VU University Amsterdam

Master Thesis Business Analytics

Investigating opportunities to increase the response of a web questionnaire

Research to find out whether it is possible to increase response rates by creating customized questionnaires for different devices

By

Dieuwe van Bergen en Henegouwen

Fall 2015

Supervisors:

Arjen Neefkes (ISIZ)

Koen Haverkamp (ISIZ)

Fetsje Bijma (VU)

Evert Haasdijk (VU)

VU University Amsterdam

Master Thesis Business Analytics

Investigating opportunities to increase the response of a web questionnaire

Research to find out whether it is possible to increase response rates by creating customized questionnaires for different devices

Dieuwe van Bergen en Henegouwen

2513239

VU University Amsterdam
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam

ISIZ B.V.
Haarlemmerstraat 121-1
1013 EN Amsterdam

Preface

This thesis is written in conclusion of the Master's program Business Analytics at the VU University Amsterdam. In this thesis, I describe the research carried out during my internship at ISIZ. During the internship I had to perform a research about a practical problem of the internship granting company.

ISIZ is a small company that started in 1996 and is located in the centre of Amsterdam. ISIZ is specialized in data collection using web questionnaires. It provides the whole technical part of developing questionnaires. Furthermore, ISIZ advises their clients for creating efficient questionnaires. This research is done to achieve more knowledge for advising clients to develop better questionnaires.

There are several people I would like to thank for helping me during this thesis. First, I would like to thank my supervisors at ISIZ, Arjen Neefkes and Koen Haverkamp, for their continuous support, feedback and discussions. Thanks are for Merel Smit for all the help with my English writing skills. Furthermore, I would thank Fetsje Bijma for all the help with the statistical models and machine learning techniques. At last, I also would thank second reader Evert Haasdijk for reading this report.

Dieuwe van Bergen en Henegouwen
Fall 2015

Abstract

Nowadays, people receive more and more e-mails which invite them to participate in web questionnaires. Filling in these questionnaires can be done on all kind of devices, like a mobile phone, personal computer (PC), and, tablet. At this moment, ISIZ only creates one version of the questionnaire for all devices, and all participating respondents have to answer the same questionnaire.

In this thesis, research is done to find opportunities to improve the response of a survey. This improvement should be done by creating customized questionnaires for different devices. A model is created to predict the fractions of respondents who are likely to use a mobile phone, PC or tablet. Nevertheless, designing customized questionnaires is only useful by sufficient interest in responding (web) questionnaires via mobile devices (mobile phone and tablet). The customized questionnaires should be designed with the least possible dropout triggers. Possible dropout triggers are the time that a respondent is already spending to answer the questionnaire and the question type of the last seen question.

Because of the technological development in recent years, questionnaires can be filled in on more and more devices. In 2012 95% of all respondents answers a questionnaire on a PC, while, in 2015 this fraction dropped to 69%. At the same time, the fractions of respondents that use a mobile phone or tablet has increased. This increase may be due to the technological development, but also to the update of the old survey3 into the new surfey4 software, for which answering questionnaires on mobile devices are much easier.

Multiple trend analysis models are used to predict the fractions of devices that will be used in completing questionnaires for the upcoming years. Two models are found that fit the data best. One of the models is selected because of the context of the problem and the pattern that is found in the classification model. The selected model predicts a positive trend in the fraction of respondents who answer the questionnaire via a mobile phone, a negative trend via a PC and a stabilized trend via a tablet.

Regular classification models predict the class to which a new observation belongs. However, in this thesis only class probabilities are assigned to respondents. These class probabilities are used to predict the fractions of devices that a panel is likely to use. This prediction is made on the base of the background information of a group of respondents and the questionnaire. For this prediction the following classification techniques are used: multinomial logistic regression, decision tree learning, support vector machines and naive Bayes classifiers.

From all classification techniques, the decision tree learning system has highest quality. On the other hand, naive Bayes classifiers have a slightly bigger error but a smaller variance, which prevents wrong predictions. Besides the low variance of the naive Bayes classifier, the decision tree learning system work most accurate. When predicting fractions of devices the respondents' age influences the outcome variable most, followed by type and expected completion time of the questionnaire.

A respondent who starts, but not completes the questionnaire is a big loss for researchers. Proportionally most of these dropouts occur on a mobile phone or tablet. In this study, the time a respondent is willing to answer the questionnaire and the type of last seen

question are examined as dropout trigger. The time as dropout trigger is divided into two parts; the number of minutes spending and the number of answered questions before a respondent leaves the questionnaire. 31% of all dropouts leave the questionnaire before having answered the questions on the first page. Between devices, there are no significant differences in time spent before a respondent leaves the questionnaire.

Another examined dropout trigger is the type of the question. In this study four question types are examined; matrix questions, multiple choice questions with only one possible answer, multiple choice questions with multiple possible answers and open (ended) questions. Matrix questions are hard to answer and lead to significant more dropouts on all devices. Moreover, multiple choice questions with multiple possible answers are significant harder to answer on a mobile phone and tablet and are on those particular devices dropout triggers.

The study in this thesis shows that more and more respondents answer questionnaires via mobile devices. Furthermore, the fraction of respondents who are likely to use a certain device can also be predicted. Nevertheless, designing different questionnaires for different devices, which would lead to higher response, is complicated. Further research is necessary to investigate more dropout triggers, such that customized questionnaires for different devices can be created.

Contents

1 Introduction

People receive e-mails which invite them to review services they have used, such as hotel rooms, an item they bought, or other services they have used. Most of these reviews make use of a web questionnaire, because this technique collects a lot of information very easily. During these review, a respondent has to select one or more answers on certain questions. Most of these questions are closed answer questions, although there can also be open (ended) questions. With all these data, researchers try to find patterns and meanings in the data.

This chapter will introduce the subject of this thesis. It consists of a company description of where the internship is done; a well formulated problem description, and the research questions, with main and sub questions.

1.1 Company description

This thesis is written on occasion of an internship at ISIZ B.V. ISIZ is a small company that started in 1996 and is located in the centre of Amsterdam. Since 2014, ISIZ expanded with another office in Barcelona. Both offices together, ISIZ counts around 30 employees. The company was founded by three friends who wanted to collect data for their thesis. At that time collecting data was complicated, so they decided to start a company which simplifies data collection.

ISIZ specializes in data collection using web questionnaires. These questionnaires are used in market research and employee or customer satisfaction. ISIZ mainly works for big companies, for which components such as data safety have high priority. Besides building questionnaires, ISIZ builds other tools like brainstorm sessions and real-time dashboards. They also manage the panels of their clients.

ISIZ programs the questionnaire, tests the questionnaire, sends invitations to all respondents, and deliver the dataset to the client. In the whole process of collecting data, ISIZ provides the technical part. ISIZ has developed their own software to program the questionnaires, as well as software (Survey4) which modifies the lay-out when a respondent uses a mobile device (mobile phone and tablet). The Survey4 scales the questions and buttons to the size of the screen, which simplifies it to select an answer. Because of the smaller screen size of mobile phones respondents has to scroll more by reading all answer options. Most of the questionnaires are standardized, but ISIZ also builds customized questionnaires. Nowadays, most of the surveys make use of the Survey4 software, but there are still surveys which make use of the older Survey3 software.

In consultation with the client a blueprint of the questionnaire is designed. The questions are prepared by the client, but ISIZ gives advice on how to design a good questionnaire. These advices are based on the knowledge that ISIZ has gained in the past years. ISIZ does research to understand the behavior of the respondent. Whitepapers [1], [2], [3] have been written by ISIZ about the best time to send invitation e-mails, but also to investigate whether a progress bar has influence on the completion rate.

1.2 Problem description

In this thesis the differences between answering questionnaires via different devices are examined. In this study only the following types of devices are used: mobile phone, tablet, and personal computer (PC), which can be a desktop computer or laptop. Nowadays more and more people have one of these devices with an internet connection. Therefore, a researcher can easily meet his panel, because most of the researchers use web questionnaires to collect data. All participants are invited by e-mail and are free to open this link on a mobile phone, PC or tablet. Furthermore, questionnaires can be opened at any moment and any place. They can answer the questionnaires at home, in their office or even in public transport.

In this thesis a model is designed which classifies the fractions of devices of a certain group of respondents for answering web questionnaires. This classification is done on the base of background information of the respondents and the questionnaire. These classification models are built into a computer based tool. Besides the classification model, a prediction of the fractions of devices that are used for the upcoming two years is made. Furthermore, the fractions of devices used over a day, week, and year are examined on trends or seasonality. This analysis is done to examine whether sufficient respondents are willing to use mobile devices.

To improve the response of a survey, the knowledge how to design customized questionnaires is achieved. Different questionnaires can be created for different devices. These questionnaires should be designed with as little as possible dropout triggers.

1.3 Research questions

The main question of this thesis is formulated as:

Is it possible to improve the response of a survey, when a researcher knows in advance the fractions of devices used in the survey?

To answer the main question, some sub questions have been formulated:

(i) *Are there any trends in the fractions of devices used for answering online questionnaires? And how do these fractions change in upcoming two years?*

(ii) *Is it possible to classify, based on background information of the respondents and the questionnaire, which device a certain group of respondents uses for answering (web) questionnaires?*

(iii) *Are there differences found between the dropout triggers of answering questionnaires on different devices?*

In the first sub question the data is modeled into a time series framework. Trend analysis is done to examine the fractions of devices used over a longer period of time. Furthermore, the different trend and seasonal components during a day, week and year are examined. Eventually, a prediction is made to figure out whether the fractions of devices that are used in the upcoming two years will increase, decrease or stabilize. This analysis tries to find whether there is sufficient interest in answering questionnaires via a mobile device. Otherwise, creating customized questionnaires would not be attractive for researchers.

The second sub question tries to predict the fractions of devices that a certain group of respondents will use in web questionnaires. A classification model is used that calculates the probability that a respondent will use a certain device (class probabilities). This model is different compared to common classification models which predict outcome variables. This sub question makes use of techniques such as statistical and machine learning techniques. A major problem in this question is the limited information which is known as input for the classification model.

The third sub question tries to find dropout triggers for different devices. A dropout trigger is a property of the questionnaire due to which a respondent would leave the questionnaire with a higher probability. Possible dropout triggers could be the length of the questionnaire or a question type. These dropout triggers are used in order to design customized questionnaires to improve the response of a survey.

1.4 Structure of the report

This thesis is structured as follows: the second chapter provides the method section. This method section is divided in four smaller sections; literature study, trend analysis, classification techniques, and dropout triggers. The literature study provides an overview of the studies that are already done in the field of (web) questionnaires. In the trend analysis section, the data is examined on trend and seasonal components. A prediction is made on how the fraction of devices will develop in the upcoming two years. Furthermore, the number of completes in questionnaires using the old survey3 software are compared with the new survey4 software. The section about the classification techniques, describes all techniques that are used in order to calculate fractions of devices for a certain group of respondents. Moreover, the time that a respondent is willing to take answering the questionnaire and the type of last seen question are examined as dropout trigger in the last method section.

In the third chapter the results of the analyses are presented. The quality of the different models discussed. The last chapter contains the conclusions, recommendations, and further research.

2 Methods

The methods section describes the different datasets and techniques that are used in this thesis. First, a literature study investigates what is already done in the field of (web) questionnaire research. Second, the dataset and methods for trend analysis, classification techniques, and dropout triggers are presented.

2.1 Literature study

In the past fifty years approaches for collecting data have changed a lot. In the earlier days, the researcher uses personal interviews to collect data. Later, questionnaires were introduced, which were sent by mail to a group of respondents, which sent back the answered questionnaires. Because of the technological development nowadays researchers make use of other mediums to collect data. With mediums such as the telephone, fax, e-mail, and the web, most of the researchers never see their panel anymore. A lot of research [4], [5], [6], [7], [8], [9], [10] has been done on the effects that a certain medium can have on response rate and given answers. The general conclusion is that the response time decreases for approaches which use superior technological hardware. However, the mean response rate [7] did not grow during 1986 and 2000. It shows only a small peak during 1995-1996, the time in which e-mail becomes extremely popular. Another study [11] showed that between the years 2004-2009 the response rate has dropped. They suggested that this drop is caused by many more interesting things to do on the web.

Not all researchers agree that technological improvements will lead to more and better data. In 2001 Porter and Whitecomb [12] predicted that in 2006 there would be so much spam mail that web questionnaires would become a less useful tool for survey research. This research suggests that personal contact between the researcher and the respondent will lead to data with higher value.

Some studies are done to improve the response rate. The Total Design Method [13] introduces the idea to send a follow-up e-mail to non-responders, one week after the initial mail-out. After three or seven weeks a whole new questionnaire will be sent to the non-responders. The Total Design Method guaranteed a response rate of 80% on e-mail and telephone surveys. Primarily, Dillman [14] claims that it is impossible to reach the complete population by making use of only one medium. He suggests the use of multiple media such as e-mail, telephone, fax, and web questionnaires for collecting interesting data. Another research by Snijders and Matza [1], [2] investigates the effects of the lay-out of the questionnaire. This research shows that the completion rate of a survey increases by dropping the progress bar.

Nowadays, almost all survey invitations are sent by e-mail. These e-mails contain a web link to participate in the survey. A respondent is free to open this link on a PC, mobile phone, tablet, or any other device like video game console, eBook reader, smart watch, and smart television. In this area, research has been done [15], [16] to detect the best lay-out for certain devices, mostly for mobile devices. Because of the smaller screen of a mobile device, the lay-out should differ: the text and answer buttons are bigger, to increase ease of use. During recent years the percentages of mobile devices which are used to answer questionnaires have increased. Fuchs and Busse [17] attempt to find out whether the increase of mobile devices in surveys has a bias with the increase of smart phones. This study suggests that it is too early to

use mobile web surveys as a mode of data collection, because mobile web coverage biases are already smaller than the coverage biases of the population with an internet connection.

Some study is done to research the influence of various conditions, like dropout triggers, under which questionnaire are designed with a lowest dropout rate possible. Some study [18], [19] shows that by increasing the amount of questions the dropout rate will increase. Brent [19] showed that by adding five extra questions to a questionnaire, the dropout rate will increase by another 2%. La Bruna and Rathod [11] showed that the response rate has decreased between 2004 and 2009. But they also showed that the dropout rate has decreased.

2.2 Trend analysis

This chapter will introduce different time series techniques for trend analysis. A time series is a sequence of data points which is dependent on time and has a natural and temporal ordering. Examples of time series data is the daily closing values of the Dow Jones or the weekly water level in a river. One of the interesting issues with time series is predicting future values. This technique is called forecasting and makes use of historical data. Besides forecasting, time series analysis detects trend, and seasonal components.

The goal of the trend analysis in this thesis is to examine which devices are used in responding to web questionnaires. Furthermore, in this section a prediction is made for the fractions of devices that will be used in the upcoming two years. First, the dataset is transformed into compositional time series data. Second, several models are trained to predict the fractions of used devices, and cross validation is used to evaluate forecast models. Finally, the model with highest quality is used to predict the fractions of devices used in the future. Besides forecasting, the fraction of devices that are used over a day, a week and a year are examined.

2.2.1 Data exploration

The data for time series analysis comes from some preselected surveys. The data in the dataset meets the following criteria: a representative age distribution, a balanced male-female ratio, and data are continuously collected between 2012 and 2015. The data is collected each week to ensure that there are enough data points for time series analysis. However, the dataset is sampled in such a way that all surveys have an equal amount of records.

The dataset contains 144.033 records and each record has deviceID, and a week number variable. A record with a deviceID that is not a mobile phone, PC or tablet is removed from the data. For each week the amount of devices that have been used are calculated. Because not all weeks have the same amount of records, the data is transformed into compositional data, which is explained section 2.2.2.1.

The fractions of devices that are used over the last couple years is shown in Figure 2.1, on the next page. Respondents who use a PC in web surveys have declined over the last years from 95.19% in 2012 to 69.52% in 2015. Contrary to the negative trend in PC users, a positive trend is shown in the fractions of tablet and mobile phone users. The fraction of tablet users increases from 3.63% in 2012 to 16.10% in 2015, while the increase in mobile phone users is even larger (2.42% in 2012 and 14.36% in 2015).

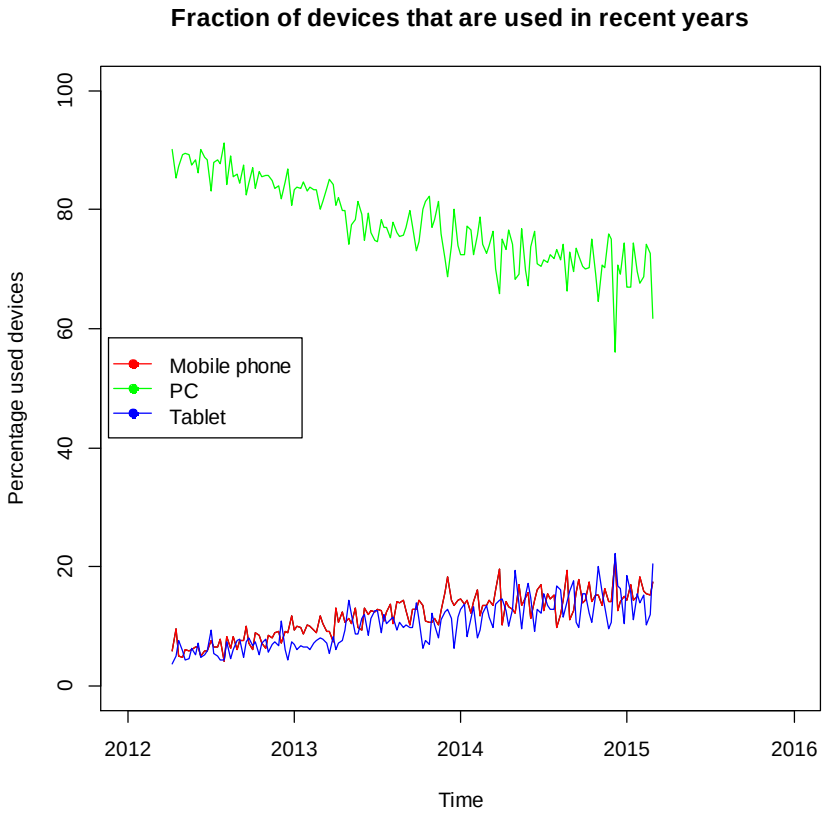

Figure 2.1: fractions of devices that are used during 2012 and 2015.

2.2.1.1 Devices during a year, a weak and a day

The left plot in Figure 2.2, on the next page, shows the fractions of devices that are used for (web) questionnaires during a year, while the right plot shows the fractions of devices during a week. The left plot of the figure shows no positive or negative trend (ANOVA; p-value = 0.996) in the fractions of devices that are used each month of the year. Only during the summer a small positive peak of personal computer users is visible. On the other hand, the right plot of the figure shows on Thursdays a small dip in PC users. Nevertheless there is no significant (ANOVA; p-value = 0.146) difference between the fractions of devices that are used during a week.

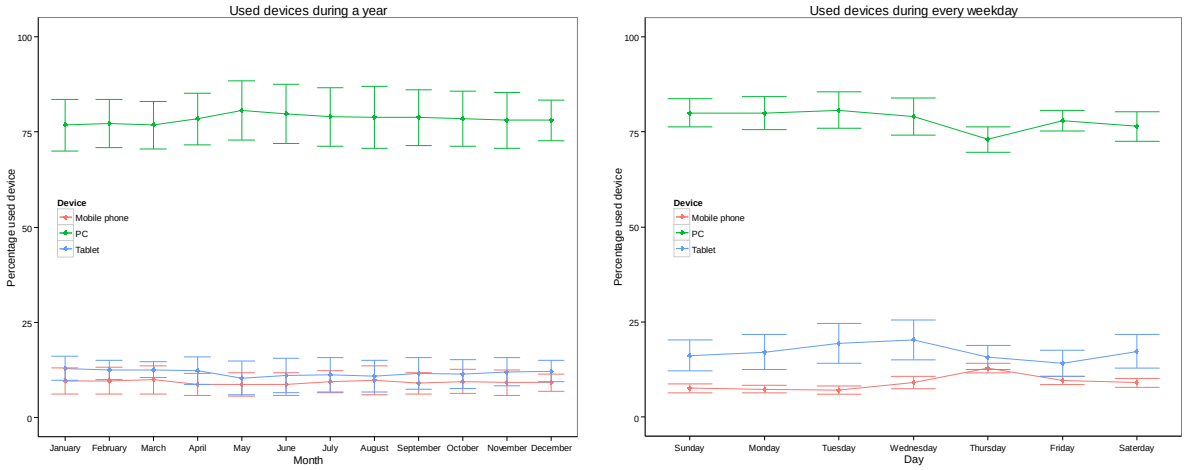

Figure 2.2: plot of fraction devices that are used during a year and a weak.

Much more interesting are the fractions of devices which are used during a day. This fraction and the amount of answered questionnaires of each hour are shown in Figure 2.3. The left plot shows the number of completed questionnaires during a day. This plot shows fewer completes during the night and early morning $\{1{:}00\ A.M., \ldots, 7{:}00\ A.M.\}$. Therefore, the error bars, in the right plot of the figure, are in that time period larger. During the night and early morning fewer PC users are active. However, the fraction of PC users is highest during traditional office hours (9:00 A.M. until 6:00 P.M.). This suggests that most PC users answer questionnaires while they are in their office. Between 12 A.M. and 1 P.M. a small dip is visible in the fraction of PC users which can likely be explained by the lunch break. After working hours the fraction of mobile phone and tablet users grows again.

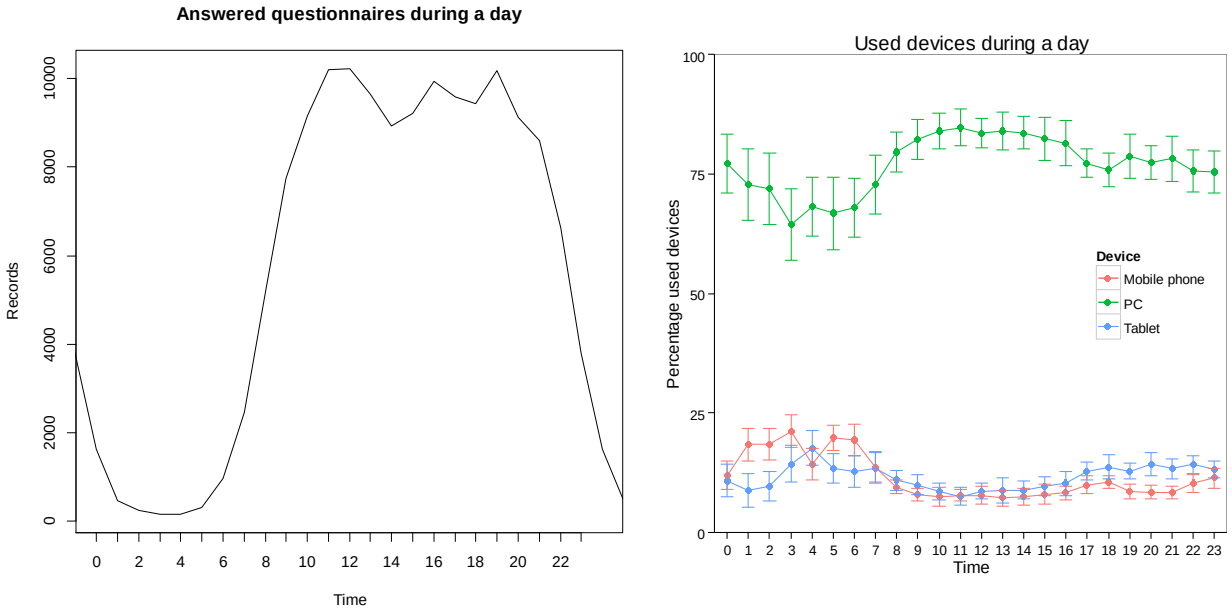

Figure 2.3: Left plot shows the number of completed questionnaires for each hour of a day. Right plot shows the fraction of devices that are used for each hour of a day.

2.2.1.2 Survey3 VS Survey4

At the beginning of 2015 ISIZ has updated their software, from Survey3 until Survey4. After this update answering questionnaires on mobile devices has become much easier, because the lay-out of the questionnaire is scaled to the size of the screen. Besides, the questions and buttons are presented in a larger size, which simplifies selecting an answer.

To investigate the differences between survey3 and survey4 software, data is used from 2013 until 2015. Each record in the dataset consists of the status of the session, which is either a complete or a dropout. In this case, a respondent who start but not complete the questionnaire is a dropout. For each week, the ratio of completes and dropouts are recorded and shown in Figure 2.4. This figure shows highest ratio of completes for PC and tablet users. For mobile phone users the ratio of completes is much lower. After updating the software, the ratio of completes are significant larger for mobile phone and PC users (t-test with dummy variable; p-value = 0.00062 and p-value = 0.0215). The fraction of completes from tablet users is not significantly (t-test with dummy variable; p-value = 0.251) different after updating the software. So, updating the software has led to more completes on a mobile phone and a PC.

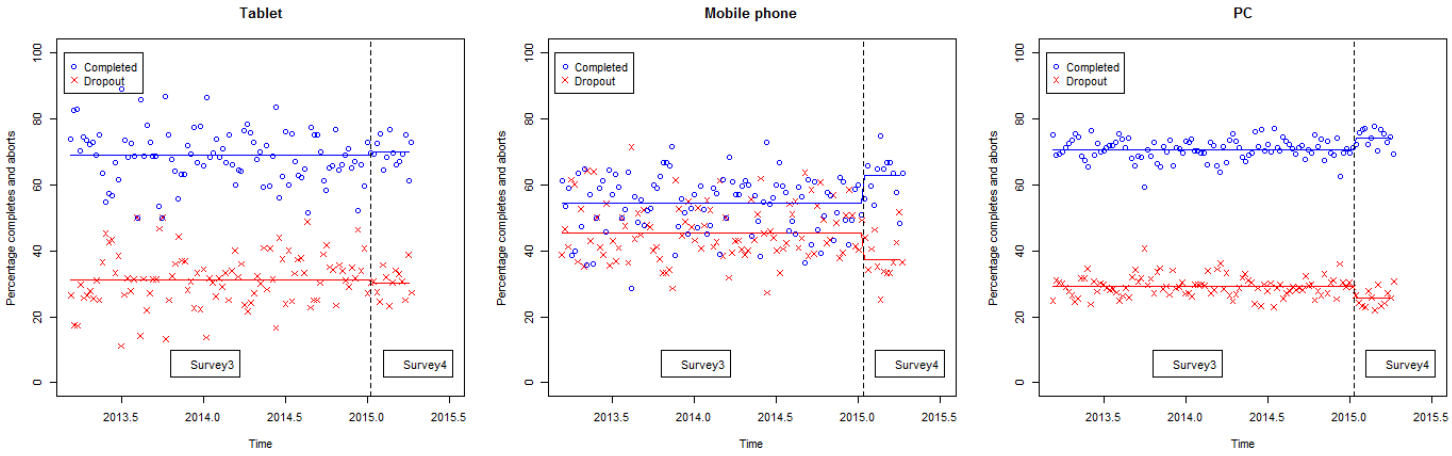

Figure 2.4: Plots of ratio completes and dropouts between 2013 and 2015 for different devices.

2.2.2 Methods

This section will introduce the methods to predict future time series values. First, the data is transformed into a compositional time series. Then, the Box-Jenkins modeling approach is used in forecasting future time series. Furthermore, different forecasting models are explained.

2.2.2.1 Compositional data

Compositional data [20] consists of vectors whose components are a proportion or percentage of a whole, and sum to one. When a time series y_t lies within (0, 1) and $1 - y_t$ is also a time series, for which the sum of both time series is one, then the time series is compositional. Therefore, a compositional time series is a time series $y_{1t}, y_{2t}, \dots, y_{mt}$ in which

$$0 < y_{it} < 1$$

And

$$\sum_{i=1}^{m} y_{it} = 1 \quad \text{for each } t = 1, \dots, T$$

The first step in forecasting compositional time series [21] is transforming the original observations. The *additive log ratio transformation* transforms a dataset from m-dimensional space into $m-1$-dimensional space, in this thesis from a 3-dimensional space into a 2-dinensional space. Before transformation, the time series has observed values $y_{i,t}$ with $i = 1,2,3$ from time period $t = 1, \dots, n$. At each time point t the values are transformed in

$$z_{i,t} = ln\left(\frac{y_{it}}{y_{mt}}\right) \quad \text{for } i = 1, \dots, m-1$$

The *generalized logistic transformation* transforms the $m-1$-dimensional forecasted data into m-dimensional space. Suppose $\hat{z}_{i,t}$ to be the point forecast of $z_{i,t}$. The transformed point forecast at time t is

$$\hat{y}_{i,t} = \begin{cases} \dfrac{\exp(\hat{z}_{i,t})}{1 + \sum_{k=1}^{m-1} \exp(\hat{z}_{k,t})} & i = 1, \dots, m-1 \\ \dfrac{1}{1 + \sum_{k=1}^{m-1} \exp(\hat{z}_{k,t})} & i = m \end{cases}$$

2.2.2.2 Box-Jenkins modeling approach

The Box-Jenkins modeling approach [22] is a step-by-step approach to design the best performing model for predicting future time series data. The Box-Jenkins modeling approach is originally designed to create ARIMA forecasting models. The traditional Box-Jenkins modeling approach is shown in the following figure.

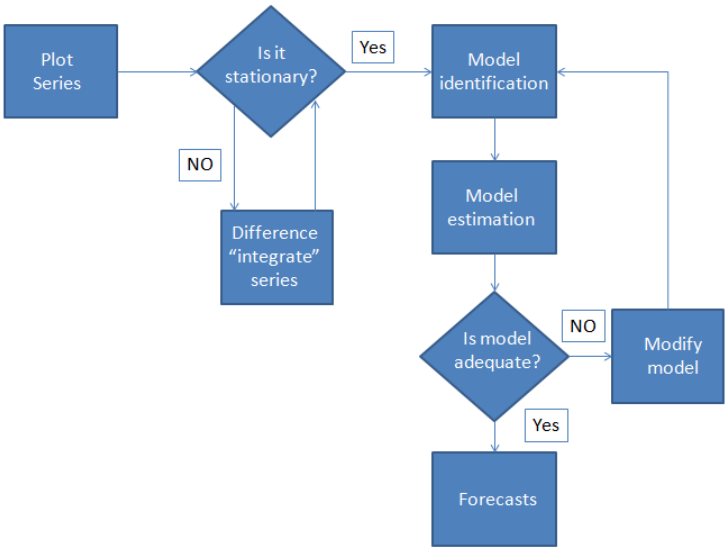

Figure 2.5: Graphical view of the Box-Jenkins modeling approach.

In this thesis the Box-Jenkins modeling approach is adapted by adding extra models (neural networks and exponential smoothing) to the model identification step. In case of neural networks and exponential smoothing the data does not have to meet the assumption of stationarity. Therefore the second step in the Box-Jenkins modeling approach can be ignored for neural networks and exponential smoothing.

The Box-Jenkins modeling approach starts with plotting the time series data, to make the data visible. This plot gives an indication as to whether the data is stationary. Apart from plotting, The Ljung-Box test, Augmented Dickey-Fuller t-statistic test, and Kwiatkowski-Phillips-Shin test can be used to test for stationary time series. In this thesis the Augmented Dickey-Fuller t-statistic test is used.

Many time series models assume stationarity. A time series is stationary when it has no trend and seasonality components. Moreover, a stationary time series has a constant variance. A strictly stationary process has the property that, given $t_1, t_2, \dots, t_m$ the joint distribution of $X_{t_1}, X_{t_2}, \dots, X_{t_m}$ is the same as the joint distribution of $X_{t_1+\tau}, X_{t_2+\tau}, \dots, X_{t_m+\tau}$ for all m and τ. This means that the joint distribution of (X_t, X_s) is the same as (X_{t+r}, X_{s+r}) and depends only on the difference between s and t, i.e. $s - t$.

There exist several methods to eliminate trend and seasonality components in time series. Such methods can transform non-stationary time series into stationary time series. The log-transformation is one of these methods and stabilizes the variance of the time series. Besides, a method such as differencing stabilizes the mean. The differenced series consists of the differences between consecutive observations in the original time series, and can be written as

$$y'_t = y_t - y_{t-1}.$$

Single order differencing does not always transform time series into a stationary time series. By increasing the order of differencing, the data (somehow) transforms into a stationary time series. Other methods to eliminate trend and seasonality components are the small trend method, linear detrending, and EMD-based detrending [23].

This study uses three types of forecasting models: ARIMA, exponential smoothing, and neural networks. The following sections will give an overview of the definitions and specifications of the three models. First, ARIMA will be handled, followed by the exponential smoothing and neural networks.

2.2.2.3 ARIMA model

The ARIMA model is a combination of an AR and MA model, and a generalization of the ARMA model. In this section the AR, MA, ARMA and ARIMA models are explained.

A common model for modeling time series is the autoregressive (AR) model. The output of an AR(p) model is linearly dependent on its previous values. The AR(p) model with a positive p can be written as

$$X_t = \delta + \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + Z_t$$

For which $X_t, X_{t-1}, \dots, X_{t-p}$ is the time series, $\varphi_1, \varphi_2, \dots, \varphi_p$ the parameters of the model, Z_t is white noise term and δ denotes the process mean.

Another common model for modeling time series is the moving average (MA) model. An MA (q) model consists of lagged white noise terms. Therefore, the output variable of the MA (q) model uses white noise errors in a regression-like model. The MA (q) model with a positive q can be written as

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}$$

For which X_t is the times series, $\theta_1, \theta_2, \ldots, \theta_q$ the parameters of the model, and $Z_t, Z_{t-1}, \ldots, Z_{t-p}$ the white noise terms.

The Autoregressive moving average (ARMA(p, q)) model is a combination of AR and MA models. This model consists of p AR terms and q MA terms. The ARMA(p, q) model can be written as

$$X_t = \delta + \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} - Z_t - \theta_1 Z_{t-1} - \ldots - \theta_q Z_{t-q}$$

The meaning of the symbols is the same as for the AR and MA models.

The Autoregressive Integrated Moving Average (ARIMA) model is a generalization of an ARMA model. The ARIMA(p, d, q) model consist of p AR terms, d integrated terms, and q MA terms. An ARIMA(p, d, q) model with d high enough is stationary by definition, due to the integrated term differencing techniques which are used to transform the time series into stationary time series. For instance the ARIMA(1,1,2) model can be written as:

$$X_t = \delta + X_{t-1} + \varphi_1 (X_{t-1} - X_{t-2}) - \theta_1 Z_{t-1} - \theta_2 Z_{t-2}$$

The meaning of the symbols is equal to those of the AR and MA models. The general ARIMA equation can be written as

$$\left(1 - \sum_{i=1}^{p} \varphi_i B^i\right) (1 - B)^d X_t = \delta + \left(1 + \sum_{i=1}^{q} \theta_i B^i\right) Z_t$$

with

$$B^k X_t = X_{t-k}$$

Where B is the backshift operator, d as order of differencing and δ as drift operator, other variables are equal to those of the AR and MA models. For ARIMA models without drift the drift operator is equal to zero.

So far only non-seasonal ARIMA models have been discussed. However, ARIMA models are capable of modeling a wide range of seasonal data. A seasonal ARIMA model is formed by including additional terms for seasonality. A seasonal ARIMA model can be written as ARIMA(p, d, q) $(P, D, Q)_m$, with m as the number of periods per season. With use of the backshift operator the seasonal ARIMA$(p, d, q)(P, D, Q)_m$ model is formulated as

$$\left(1-\varphi_1 B-\cdots-\varphi_p B^p\right)\left(1-\Phi_1 B^m-\cdots-\Phi_P B^{Pm}\right)(1-B)^d\left(1-B^m\right)^D X_t = \left(1-\theta_1 B-\cdots-\theta_q B^q\right)\left(1-\Theta_1 B^m-\cdots-\Theta_Q B^{Qm}\right)Z_t$$

With B as backshift operator and $\Phi_1, \Phi_2, \ldots, \Phi_P$ and $\Theta_1, \Theta_2, \ldots, \Theta_Q$ as model parameters for the additional seasonal terms.

Two methods can be used for identifying the order of forecast models: examine AutoCorrelation (ACF) and Partial AutoCorrelation Functions (PACF), or use an automated iterative procedure. The most common method is to examine the ACF and PACF plot. In order to select the order of an AR(p) model the lagged terms of the PACF is used. Equally, the ACF is used to select the order of an MA(q) model.

A built-in package in R with automated iterative procedure works accurately for selecting the order of ARIMA models. This automated iterative procedure returns the order of the ARIMA model with lowest Akaike Information Criterion (AIC).

2.2.2.4 Exponential smoothing models

Exponential smoothing is another method for predicting future time series values. Just like ARIMA, it uses data from the past to predict future time series values. For prediction, recent observations have more influence than older ones. The simple exponential smoothing (SES) model, sometimes called the single exponential smoothing model, ignores trend and seasonality components. Therefore, this model is often used for short range forecasting. The forecast for time $t+1$, denoted as S_t, is equal to a weighted average between the observation and forecast at time t, and can denoted X_t and S_{t-1}, respectively,

$$S_t = \alpha X_t + (1-\alpha) S_{t-1}$$

S_t is called the level at time t and parameter $0 < \alpha < 1$ can tune the influence of the current smoothed value or the current data point.

Double exponential smoothing is a generalization of simple exponential smoothing. Double exponential smoothing models contain a trend component. At each period in time, level (S_t) and trend (b_t) components are updated as follows:

$$S_t = \alpha X_t + (1-\alpha)(S_{t-1} - b_{t-1})$$

$$b_t = \beta (S_t - S_{t-1}) + (1-\beta) b_{t-1}$$

With $0 < \alpha, \beta < 1$. The β term in this formula is the weight that determines the influence of the trend component b_t. There are several methods to choose the initial value of b_t , for instance $b_1 = y_2 - y_1$, which is used throughout this study.

2.2.2.5 Artificial neural networks

An artificial neural network is a typical machine learning technique. Pattern recognition, classification, and regression problems can be solved with neural networks. A neural network exists of an interconnected group of nodes (neurons) and a set of adaptive weights. These adaptive weights are the conceptual connections between the neurons in the different layers. In a neural network there are three types of layers; an input layer, hidden layer (can be multiple) and the output layer. Training the network leads to updating the adaptive weights. Observed

data are used in order to train the neural network and the network learns an approximation of the relationship by iteratively adapting its weights. This section only considers feed-forward networks with one hidden layer. An overview of a feed-forward neural network with one hidden layer is shown in Figure 2.6. This network has four input variables, one output variable and five nodes in the hidden layer.

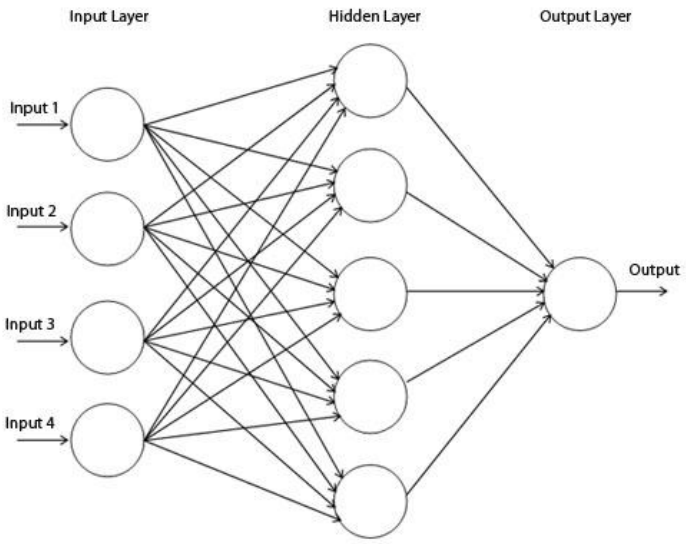

Figure 2.6: Overview of a feed-forward neural network with one hidden layer[1].

The data points $X_{t-1}, X_{t-2}, \dots, X_{t-4}$ are used as input for training a neural network. The general notation for a feed-forward neural network is NNAR(p, k). This network has p lagged inputs and k nodes in the hidden layer. The network in Figure 2.6 is an NNAR(4, 5) network. A feed-forward neural network with zero nodes in the hidden layer (NNAR(p,0)) is equivalent to an ARIMA(p,0,0), and AR(p) model.

2.2.2.6 Model accuracy

In order to compare the performance of the different forecast models, the model accuracy is computed for each model. For computing the accuracy of a model the Mean Absolute Error (MAE) is used.

$$MAE = mean(|y_t - \hat{y}_t|)$$

with y_t as the t −th observation and $\hat{y}_t$ denote a forecast y_t.

Validation of the models uses three-fold-cross validation. The training set consists of only observations that occurred prior to the observations from the test set. Thus, no future observations are used in constructing the forecast. This procedure is sometimes called “rolling forecasting origin”. The following algorithm shows this procedure

[1] http://www.codeproject.com/Articles/175777/Financial-predictor-via-neural-network

1. Select set of observations $k+i, \dots, T$ as test set, and use observations $1,2, \dots, k+i-1$ as training set to estimate the forecasting model. Calculate the error on the forecast of the test set.
2. Repeat step 1 for $i = 1,2, \dots, T-k$, where T is the total number of observations.
3. Compute the forecast accuracy measure based on the mean of the obtained errors.

In this thesis the variable k is 70%, 80% and 90% of the total dataset, and represents the amount of training data.

2.3 Classification Techniques

This section gives an introduction about the different classification techniques, which are used in this thesis. Since ISIZ is only interested in the fractions of devices, techniques that are used to predict the fractions of devices are discussed. The prediction is made based on the background information of the respondents and the questionnaire. The classification techniques that are used in this thesis are: multinomial logistic regression, decision tree learning, support vector machines, and naive Bayes classifiers.

2.3.1 Data

The dataset for the classification problem is another dataset than the one used for trend analysis. The dataset for trend analysis covers 2012 until 2015. However, for the classification problem only recent data is used in order to provide up to date results, concerning the usage of devices. ISIZ has a lot of data that can be used for classification. Despite the amount of data, one problem is that the training data is relatively old. The training data are between one and eighteen months old. Data older than a year differ from data that are only one month old.

The trend analysis section in this thesis shows a changing environment in the use of different devices by completing web questionnaires. The analysis shows a decreasing trend of PC users and an increasing trend of mobile phone and tablet users. Because of the changing environment in the use of devices, the time has to be taken into account. The best way of using the time in the dataset for classification is to use the number of months passed since a respondent has answered the questionnaire.

The respondent's age and gender information is provided by the client. Beside respondent information, questionnaires have also specifications, such as survey type and the expected completion time. The client communicates the subject and the expected completion time of the questionnaire in advance. Using this information, a respondent can choose to start the questionnaire and thus become a record. After starting the questionnaire, the system detects the type of device the respondent uses.

The dataset for training consists of 22.448 records. These records come from 21 preselected questionnaires. A record, which contains only six useful variables, is recorded after a respondent has started the questionnaire. These six variables include gender and age of a respondent, type and expected completion time of the questionnaire and the type of device that a respondent has used. The final variable is the number of months ago a respondent has answered the questionnaire. When predicting class probabilities, this variable is always set to zero.

Each record is characterized by a pair (X, y), where X stands for the input attributes and y for the labeled target attribute. An input vector is composed by the age, gender, and the months passed since a respondent has answered the questionnaire, supplemented by the type and expected completion time of the survey. The labeled attribute y is determined by the device that the respondent has used.

Like the previous dataset, this dataset has a representative age distribution, and a balanced male-female ratio. Moreover, there is a significant difference (ANOVA; p-value $<2.2\ 10^{-16}$) between the age of mobile phone users (37.78 years old on average) compared to tablet (53.82 years old on average), and personal computer (50.95 years old on average) users. Furthermore, men use personal computers more often than women. 75.4% of all male respondents use a personal computer, compared to only 68.48% of female respondents. Women are more likely to use a mobile phone (17.14% vs. 12.97%) and tablet (14.37% vs. 11.60%) compared to men.

All questionnaires are classified into one of the following questionnaire types: *Employee satisfaction survey* (type A), *Customer satisfaction survey, after buying a product or entering a contract* (type B), *Customer satisfaction survey, after a contact moment* (type C), *Customer satisfaction survey after terminating a contract* (type D), *Periodical customer satisfaction survey* (type E), *Dutch representative survey* (type F), *Product evaluation* (type G), and *Event evaluation* (type H).

The average expected completion time of the questionnaires in the dataset is 9.45 minutes, with two minutes as minimum and 30 minutes as maximum. Questionnaires with higher expected completion times are more often answered on personal computers (ANOVA; p-value $<2.2\ 10^{-16}$). The average expected completion time on different devices is highest for PC users (10.17 minutes), followed by tablet users (8.11 minutes) and mobile phone users (7.15 minutes).

The dataset contains only six background variables, which is not much for a classification problem. However other variables, such as time of the day (night, morning, afternoon, and evening) do not improve the model significantly; therefore, this variable is not included in the classification model. Other variables like education level and income level are not used in classification, because they are not known for all respondents.

Regular classification models predict the class to which a new observation belongs. However in this thesis, the training set reveals a unusual pattern. Out of all 22.448 records there are 4.514 unique combinations of input vectors. Some combinations of input vectors have the same target value (device), but there are also combinations for which different target values occur. For example, 132 respondents have the same attribute values as input vector. Nevertheless, 62 respondents have used a mobile phone, 50 respondents a PC, and a tablet was used by 20 respondents. Because of this pattern, a regular classification model can never correctly predict more than $\frac{62}{132} = 46.9\%$ of the cases. Since ISIZ is not interested in the device for each respondent individually, but rather in the fractions of the devices for all respondents together, it suffices to predict the probability for each device.

Beside the training dataset, there is also a dataset for testing. This set contains 7765 records, belonging to 8 different surveys. Each type of questionnaire is represented in this set. The questionnaires in this test set are recent surveys, so that the number of months passed since a respondent completed the questionnaire is minimized.

2.3.2 Methods

Several classification techniques are used to learn from data and classify new input data. The fractions of devices that a certain panel will use are calculated by modeling class probabilities as variables from the multinomial distribution. To assess the quality of the models, the fractions of predicted devices is compared to the actual ones.

Three types of learning systems can be used in machine learning: supervised learning, unsupervised learning, and reinforcement learning. In this thesis, the techniques use supervised learning approaches. A supervised learning approach analyzes labeled training examples and produces a function, which is used for mapping test examples. A training record is a pair consisting of an input object and a desired target value. These training records consist of a vector with information of the respondent and the questionnaire (input) and the type of device (output) that a respondent had used.

This section gives an overview of the different classification techniques. The techniques used in this thesis are: multinomial logistic regression, decision tree learning, support vector machines, and naive Bayes classifiers. At the end of this section the measurement of comparing the quality of the different techniques is discussed.

2.3.2.1 Multinomial Logistic Regression

The multinomial logistic regression model [24] is an expansion of the ordinary binomial logistic regression model. This model generalizes logistic regression to multiclass problems. The multinomial logistic regression model allows more than two classes of a dependent outcome variable. The multinomial logistic regression model is also sometimes called softmax regression or multinomial logit.

The multinomial logistic regression model is used to predict the probability $\pi_{i,K}$ that a dependent variable is a member of a certain class K based on the independent input vector X_i. The model compares multiple groups through a combination of binary logistic regressions.

The linear predictor function $f(k,i)$ constructs the probability that observation i has outcome k. This predictor $f(k,i)$ can be described by a set of weights that are linearly combined with the explanatory variables, using a dot product

$$f\ (k,i) = \alpha_k + \beta_k \cdot X_i \ \ = \ \ \alpha_k + \beta_{1,k}X_{1,i} + \ \beta_{2,k}X_{1,i} + \ldots + \beta_{M,k}\,X_{M,i}$$

where $\beta_{m,k}$ is the regression coefficient associated with the mth explanatory variable and kth outcome and with α_k as intercept constant.

In the case of K possible outcome categories, the model runs $K-1$ independent binary logistic regression models. One of the K categories is selected as the so-called reference category. Consequently, the other $K-1$ categories are separately regressed against the reference category. By default the reference category is the first category; in this thesis the

reference category is the 'mobile phone' device. The $K - 1$ categories are separately regressed against the reference category. This process leads to the following equations with category K as reference category.

$$log\ \frac{\pi_{i,j}}{\pi_{i,K}} = \alpha_j + \beta_j \cdot X_i = \alpha_j + \beta_{1,j} X_{1,i} + \beta_{2,j} X_{1,i} + \ldots + \beta_{M,j} X_{M,i} \qquad for\ j = 1, \ldots, K-1$$

Taking on both sides the exponent, and solving the probabilities yields the following equations:

$$\pi_{i,j} = \pi_{i,K}\ e^{\alpha_j + \beta_j \cdot X_i} \qquad for\ j = 1, \ldots, K-1$$

Furthermore, the class probabilities sum to one, which can be written as

$$\sum_{l=1}^{K} \pi_{i,l} = 1$$

$$1 = \sum_{l=1}^{K} \pi_{i,l} = \sum_{l=1}^{K-1} \pi_{i,K}\ e^{\alpha_l + \beta_l \cdot X_i} + \pi_{i,K-1} = \pi_{i,K}\ (1 + \sum_{l=1}^{K-1} e^{\alpha_l + \beta_l \cdot X_i})$$

such that

$$\pi_{i,K} = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\alpha_l + \beta_l \cdot X_i}}$$

With that in mind, the class probability, given an input vector X_i is given by

$$\pi_{i,j} = \frac{e^{\alpha_j + \beta_j \cdot X_i}}{1 + \sum_{l=1}^{K-1} e^{\alpha_l + \beta_l \cdot X_i}} \qquad for\ j = 1, \ldots, K-1$$

The unknown parameters the intercept and β_j can be estimated with several techniques. A widely used technique for estimating parameters is the maximum likelihood estimator. The intercept is estimated only for those categories which are not the reference category. Based on the estimate of the intercept values and regression coefficients, the estimated class probabilities can be calculated.

2.3.2.2 Decision tree learning

Decision tree learning [25] is an approach that is often used in data mining, machine learning, and statistics. The goal of decision tree learning is to create a decision tree as a predictive model that maps target values based on input variables. A decision tree is usually constructed in two phases: growing and pruning. In the growing phase, the decision tree is constructed, while the pruning phase is used to reduce the size of the tree.

A decision tree exists of nodes and interconnections between these nodes. At each node a crafted question is asked about the attributes of the input vector. At each node an answer is received, a follow-up question is asked until a leaf node is reached with concluded class probabilities. A root node has no incoming edges and two outgoing edges. In this thesis, an internal node has exactly one incoming edge and two outgoing edges. The leaf or terminal nodes have one incoming node and no outgoing nodes. In this thesis, each leaf node does not return a

single class, but the probabilities for all three classes (devices). Figure 2.7 is a simple example of a decision tree.

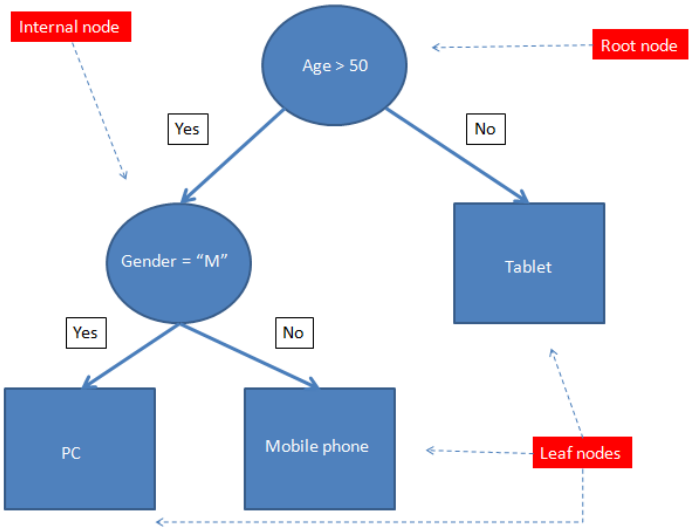

Figure 2.7: Simple decision tree structure with root, internal and leaf nodes.

When a decision tree is constructed, class probabilities can be calculated for new observations. Starting from the root node, the test condition is applied to the record and follows the outcome to a new internal node, for which a new test condition appears. This continues until the record terminates in a leaf node; for this leaf node, the class probabilities are calculated.

There are many decision trees that can be constructed for a set of attributes. Because of the exponential search space it is computationally infeasible to construct an optimal tree. Nevertheless, some efficient algorithms ID3 [26] and C4.5 [27] exist, which search through possible decision trees and detect the best splitting points for constructing decision trees. In this thesis, the best splitting points is selected by using the Gini index.

Building a tree starts at the root node and ends with the leaf nodes (top-down). At each node, the best splitting attribute is chosen to split the dataset associated with that node. Each splitting point splits the set into a left and right subtree. At each node, the splitting point has to minimize the error of a local cost function. The cost function in this greedy algorithm is the Gini index ($G(q)$). The Gini index will be computed as the weighted sum resulting partitions. The Gini index can be written as $\mathrm{G(q)} = 1 - \sum p_i^2$ with p_i the fraction of records belonging to class i. The splitting point with the lowest Gini index is the optimal splitting point and is called the final splitting point.

Now that the complete tree is built, which is possibly complex and quite large, it must be decided how much to retain. The goal of pruning in decision tree learning is to reduce the tree, without losing too much information. A large tree risks overfitting training records, while a tree that is too small might not capture the important information (underfitting). Different pruning techniques are used in order to counteract overfitting of training data. A correctly pruned tree has both low training and testing errors, instead of low training errors and high testing errors.

The reduced error pruning technique is a simple pruning technique. This technique replaces each node with its most popular class. If the prediction accuracy on the test data is not affected, the change is kept. This somewhat naive technique has the advantage of simplicity and speed.

In this thesis the cost complexity pruning technique [25] is used. Therefore, let $T_1, T_2, \dots, T_k$ be the leaf nodes of (sub) tree T and define $|T|$ to be the number of leaf nodes. $R(T)$ is the risk of T, that is, the proportion of misclassified classes in (sub)tree T multiplied by the proportion of data at sub-tree T. Now let α be a number between 0 and ∞, this α is defined to be the average increase in error per leaf of the subtree. This α represents the 'cost' of adding another variable to the tree. The cost-complexity measure $R_\alpha(T)$ can be defined as $R_\alpha(T) = R(T) + \alpha|T|$. For each sub-tree T, α_i is calculated and selects the sub-tree with lowest value of $R_{\alpha_i}(T)$ for pruning. Repeating this process until there are no sub-trees left yields a series of increasingly pruned trees. The last step in pruning a tree is to determine the final tree, which has the lowest misclassification rate on a validation set.

2.3.2.3 Support Vector Machines

In 1963 Vapnik and Lerner [28] invented the original support vector machine (SVM) algorithm. This original algorithm was only able to classify linear separable data. In 1995, the technique of soft margin classifiers was invented by Vapnik [29]; hence, the support vector machine close to their current form was introduced. The SVM became quite popular because of its good performance and theoretical foundations.

The SVM classifier presents observations as points in space in such a way that observations are divided into separate classes. The support vector machine tries to create the biggest possible margin between different classes. Records from the testset are mapped into the same space and classified by the class training records assigned. Given a set of labeled observations (in Figure 2.8 $\{dot, square\}$), there are multiple hyperplanes that might separate classes (see left plot of Figure 2.8. on the next page). A SVM identifies the hyperplane that linearly separates the dots from the squares with the largest margin. The hyperplane with the largest possible margin is called the optimal separating hyperplane, which is shown in the right plot of Figure 2.8. When classifying data into $k > 2$ classes, multiple hyperplanes are used to separate observations. Observations that lie on the boundaries are the so-called support vectors. In Figure 2.8, the support vectors are the colored squares and dots.

If data is not linearly separable in the used dimension, the kernel trick transforms the data into a higher dimensional space, so that the data becomes linearly separable.

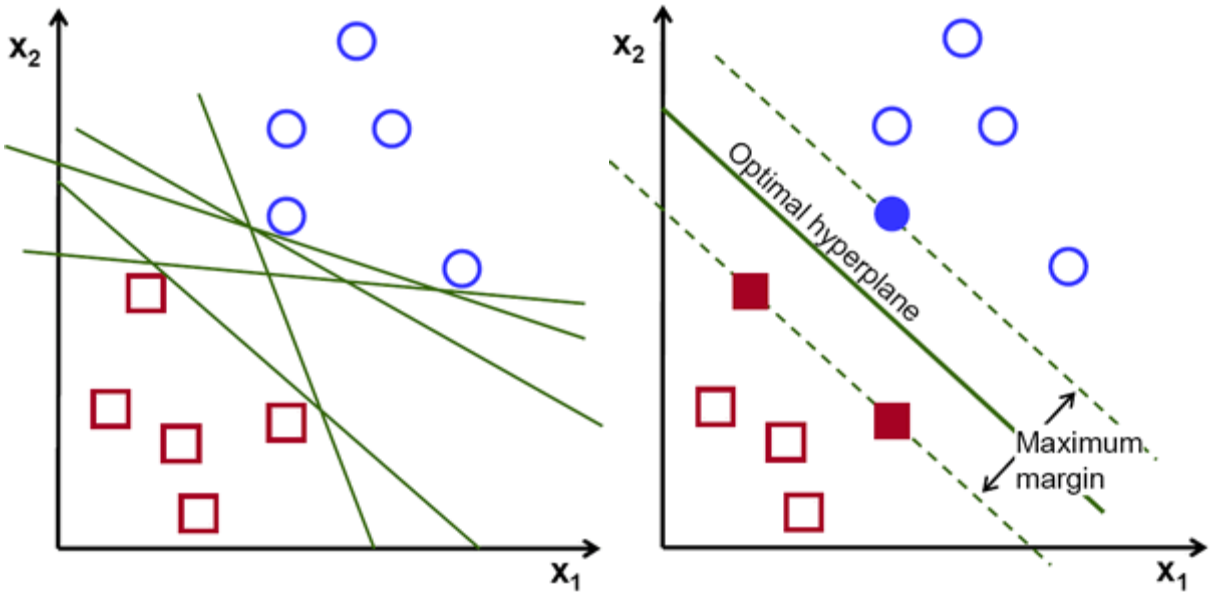

Figure 2.8: Multiple hyperplanes that separate two classes (left plot). Optimal separating hyperplane with support vectors (right plot)[2]

2.3.2.3.1 Linear support vector machines

Data that can be separated linearly into two classes uses the simple linear support vector machine [30]. A hyperplane can then be written as

$$w \cdot X + b = 0$$

with X as a set of points , w the normal vector to the hyperplane and b the offset of the hyper plane from the origin along w. With linearly separable data, two hyperplanes which are parallel are selected in such a way that there is no data between them. These hyperplanes can be described by the following equations:

$$H1: \quad w \cdot X + b = 1$$

$$H2: \quad w \cdot X + b = -1$$

The distance between the two hyperplanes $\frac{2}{\|w\|}$, has to be maximized. Therefore, the Euclidean norm of w has to be minimized. Nevertheless, the condition of no data points between $H1$ and $H2$ must still be accomplished. For each i −th data point the following constraints must be satisfied:

$$w \cdot X_i + b \geq 1 \quad y_i = +1$$

$$w \cdot X_i + b \leq -1 \quad y_i = -1.$$

This constraint can be combined as $y_i(w \cdot X_i + b) \geq 1$, for $1 \leq i \leq n$. This constrained optimization problem can only be solved with use of the Lagrangian multiplier method. Without

[2] http://docs.opencv.org/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

changing the solution of the Lagrangian multiplier the $\|w\|$ can be substituted by $\frac{1}{2}\|w\|^2$, that has to be minimized.

In general, the Lagrangian multiplier uses the formula: $L(\alpha, X) = f(X) + \alpha g(X)$ where $f(X)$ is the optimization condition, and $g(X)$ the constraint condition. In our case, this can be written as the following equation:

$$L(\alpha, X) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i\, y_i(X_i \cdot w + b) \qquad \alpha_i \geq 0\,, \forall i$$

To solve this equation the dual form is used. In the dual form is used that $w = \sum_i \alpha_i y_i X_i$. This form attempts to fix the value of f and finds the support vectors α_i, so that $L_D(\alpha, x)$ is maximized with respect to α_i.

$$L_D(\alpha, x) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j\, K(X_i, Y_j) \qquad \alpha_i \geq 0\,, \forall i$$

With $K(X_i, Y_j)$ is defined as the kernel.

2.3.2.3.2 Non-linear classification

Finding a hyperplane for linearly separable data is relative simple, but finding hyperplanes for non-linear separable data is harder. Vapnik [29] only proposed the soft margin technique that leads to the current form of the SVM. Boser, Guyon and Vapnik [31] introduce another method, the so-called kernel trick method. The kernel trick transforms non-linear observations of a finite-dimension space into a higher-dimensional feature space. In this higher-dimensional space a separating hyperplane can be constructed.

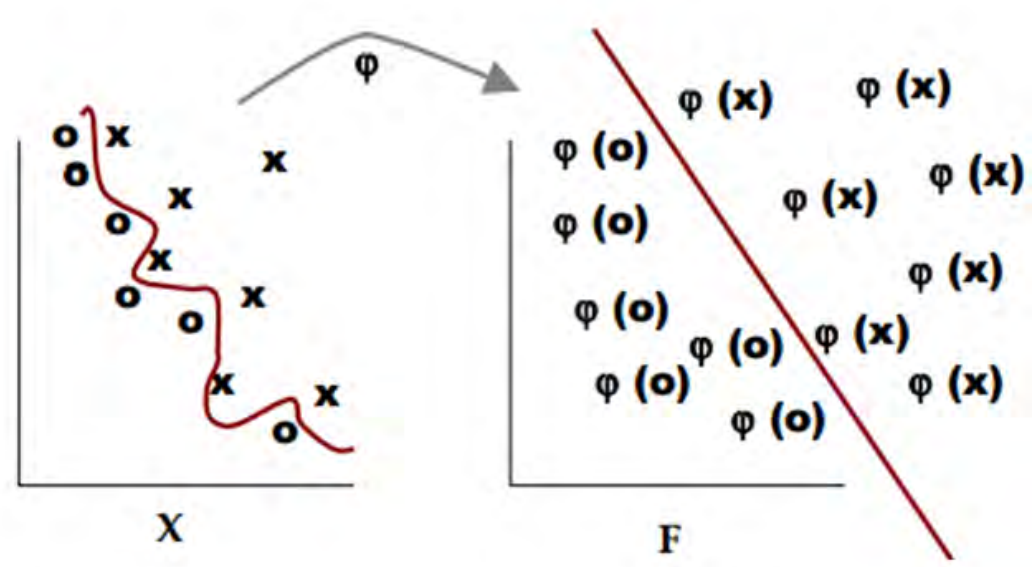

Figure 2.9: Transforming non-linear separable observations into linear separable observations into a higher-dimensional space.[3]

[3] http://www.svms.org/tutorials/Berwick2003.pdf

Figure 2.9, on the previous page, shows the transformation of non-linear observations into linear separable observations in high-dimensional space. In the figure, the kernel function is shown by the symbol phi (φ).

The kernel trick is a powerful tool because it expects that every dot product is replaced by a non-linear kernel function. This kernel function denotes an inner product in feature space and is denoted by

$$K(X,Y) = < \varphi(X), \varphi(Y) >$$

Choosing the most appropriate kernel function highly depends on the problem. In this thesis the linear, polynomial, Gaussian radial basis, and hyperbolic tangent sigmoid kernel functions are used.

The simplest kernel function is the linear kernel function. The function is given by an optional constant c and the inner product $< X, Y >$. The linear kernel function can be described as

$$K(X,Y) = X^T Y + c$$

The polynomial kernel allows model feature conjunctions up to the order of the polynomial. Polynomial kernels are well suited for problems with normalized data. The polynomial kernel, with slope γ, constant c, and polynomial degree d, is written as

$$K(X,Y) = (\gamma X^T Y + c)^d$$

The Gaussian radial basis kernel allows separating classes in circles. The radial basis kernel function is written as

$$k(X,Y) = e^{(-\gamma \|X-Y\|^2)}$$

The last kernel is the so-called hyperbolic tangent sigmoid kernel. This sigmoid kernel is sometimes called the multilayer perceptron kernel and comes from neural networks. The sigmoid kernel is written as

$$k(X,Y) = \tanh(\gamma XY + c)$$

In most kernels the parameters cost (c) and gamma (γ) are involved. Selecting the optimal combination of parameters uses a grid search approach. The SVM with kernel and best performing combination of parameters, with $c, \gamma \in \{2^{-4}, 2^{-2}, \dots, 2^1, 2^2\}$, will be compared with the quality of the other classification techniques.

2.3.2.4 Naive Bayes classifier

The naive Bayes classifier is a technique based on Bayes' theorem. In 1763, the Bayesian theorem was invented by Thomas Bayes, but formulated by Pierre-Simon Laplace in his *Théorie analytique des probabilities* in 1812 [32]. The naive Bayes classifier works usual under the assumption of independent variables. The advantage of naive Bayes classification is the simplicity of constructing classifiers, which makes it particularly useful for very large datasets.

An assumption in the naive Bayes classifier is the mutually independence between variables. Two variables $X_1$and $X_2$are independent if and only if $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)\, P(X_2 = x_2)$ The Chi-squared test of independence examines the independence between variables. The test statistic of the Chi-squared test is

$$\chi^2 = \sum_{i,j} \frac{(f_{i,j} - e_{i,j})^2}{e_{i,j}}$$

where $f_{i,j}$ is the observed frequency of events that belong to the i −th category of X and j −th category of Y. Furthermore, $e_{i,j}$ correspond to the expected count if X and Y are independent.

Let $\vec{X} = (x_1, x_2, \dots, x_n)$ be a vector of background variable to classify and C_k be one of the K potential classes. The goal of the naive Bayes classification model is to calculate the probability that a vector $\vec{X}$ belongs to a certain class C_k. For calculating these class probabilities Bayes' theorem is used, which can be written as

$$p(C_k|\vec{X}) = \frac{p(C_k)\, P(\vec{X}|C_k)}{p(\vec{X})}$$

In words, this equation is written as

$$posterior\ probability = \frac{Class\ prior\ probability * Likelihood}{predictor\ prior\ probability}$$

The class prior and predictor prior probabilities can be calculated based on training examples. The prior probability represents the relative proportion of respondents belonging to a certain group. However, direct estimation of likelihood probabilities is impossible in most cases. Therefore, $P(\vec{X}|c_k)$ is decomposed under the assumption of conditional independence between all elements of vector $\vec{X}$

$$p(\vec{X}|c_k) = \prod_{j=1}^{d} p(X_j|\, C_k)$$

After this transformation the posterior probability becomes

$$p(C_k|\vec{X}) = \frac{p(C_k) * \prod_{j=1}^{d} p(X_j|\, C_k)}{\prod_{i=1}^{d} p(X_j)}.$$

The naive Bayes classifier is originally designed for discrete attribute values. However, a simple expansion allows continuous attributes. There are two options for dealing with continuous attributes: using discretized numeric variables, or using a probability density function. In this thesis, a probability density function is used to calculate prior probabilities. The prior probabilities are calculated with the use of a normal distribution. Although, not all continuous variables come from a normal distribution, the probability density function works fine and is therefore used for dealing with continuous variables.

The standard z −table is used to assign the prior probabilities for continuous attributes. First the standard error is calculated: $SE = \frac{\sigma}{\sqrt{n}}$ for which σ is the population standard deviation and n the population size. With the standard error, the z −score can be calculated using $z = \frac{(M-\mu_i)}{SE}$, with M as the numeric attribute of the respondent and μ_i the mean of that attribute for class i. With this z −score the prior probability is assigned.

2.3.2.5 Model accuracy

The training dataset is used for comparison of the performance of the different models. 80% of the training set is used for training and the remaining 20% for validating. For testing the quality of the models the testset is used. For both validating and testing approaches, the accuracy of the class probabilities is measured with the Mean Absolute Error (MAE) measure. This error measurement has been discussed in section 2.2.2.6.

To improve the quality of model accuracy, k-fold cross validation is used. In this thesis, 5-fold cross validation is used. The 5-fold cross validation is explained with the following algorithm.

1. Divide the training examples randomly into k folds.
2. **For** $i = 1, \dots, 5$
3. Train the classifier using all records that do not belong to fold i.
4. Test the classifier on all records in fold i.
5. Compute e_i, the Mean Absolute error of fold i
6. **End for**
7. return $E = \frac{\sum_{i=1}^{5} e_i}{5}$

The error E is finally used for validating and testing the different techniques.

For testing the quality of the techniques the training set is used for training the models. Using these trained models, a prediction is made for the fractions of devices that are probably used in each survey out of the testset. These predicted fractions are compared, with the actual fractions.

2.4 Dropout triggers

When evaluating questionnaires, researchers are interested in response and dropout rates. It is obvious that lower dropout rates come from better designed questionnaires. However, creating questionnaires, that lead to fewer dropouts but still gather enough data, is not easy. This section discusses different possible dropout triggers, as well as the dataset that is used in order to find dropout triggers.

When a respondent starts the questionnaire, there are many reasons for them to leave the questionnaire; the questionnaire is too long, response to all questions is required, or the expected completion time does not match with the actual completion time. Moreover, a question type could lead to a dropout. These dropouts are a big loss for researchers, because, these respondents are willing to answer the questionnaire but because of a messy questionnaire they drop out.

2.4.1 Data

The dataset for exploring dropout triggers is comparable with the dataset for classification, but is expanded with two more variables: last seen page and the number of answered questions. The question on the last seen page could probably be the reason for leaving the questionnaire

Out of all the 9919 records, belonging to 9 surveys, 70.47% of the respondents have used a PC, 17.99% a mobile phone and 11.54% a tablet. Just like the previous datasets, this set has a representative age distribution and a balanced male-female ratio. 80.9% of all respondents who started the questionnaire, completes the whole questionnaire, 13.5% of the respondents answer at least one question but do not complete the whole questionnaire (dropout), and 5.6% start the questionnaire but do not answer any question (no start). The last category (no start) agreed to click on the invitation link, but leaves the questionnaire after reading the introduction text.

Table 2.1 shows the Proportion of completes, dropouts and no starts for each type of device. The table shows proportionally highest completion rates for PC users. Mobile phone and tablet users have the highest proportion of dropouts (18%). A lot of mobile phone users does not even answer one question and become a no start (13%). From the table can be concluded that higher response can be achieved by preventing dropouts and no starters

Status	Mobile phone	PC	Tablet
Complete	69.23%	82.75%	75.15%
Dropout	17.54%	11.77%	17.94%
No start	13.23%	5.48%	6.9%

Table 2.1: Proportion of completes, dropouts and no starts of using different devices.

The questionnaires in the dataset have 44 pages on average, with 18 pages as minimum and 94 as maximum. However, respondents do not have to answer questions on all 94 pages, because of the routing techniques. Out of all pages, 36.0% of the pages contains a matrix question, 43.4% a simple multiple choice question with only one possible answer, 12.6% contains a multiple choice question with multiple possible answers and 8.0% of the pages contains an open (ended) question.

The questionnaires in the dataset have the following expected completion times; 2, 4, 5, 7, 10, 15, and 27.5 minutes. The number of questions that a respondent has to answer grows linearly with the expected completion time of the questionnaire.

2.4.2 Dropout triggers

A dropout trigger is a property of the questionnaire due to which a respondent would leave the questionnaire with a higher probability. In this thesis two types of dropout triggers are examined: the time spent to the questionnaire and question type as dropout trigger. The time spent is discussed in two ways: time in minutes and in amount of already answered questions. When the time spent could be a dropout trigger, the questionnaire, for different devices, could be designed with different expected completion times, such that it improves the response.

The second dropout trigger examined is question type. A respondent could leave the questionnaire due to the question type as trigger. In this thesis, the last page that a respondent has visited with the unanswered question is recorded. This last seen question is labeled as one of the following question types: matrix questions (a matrix of several multiple choice questions) (A), multiple choice questions with only one possible answer (B), multiple choice questions with multiple possible answers (C), and open (ended) questions (D). Since, there are too little dropouts on other question types, other question types are not used in the analysis. Dropouts because of technical problems are also not taken into account.

Another interesting thing in the context of dropouts is the respondents who leave the questionnaire before answering any question. In the dataset 31.9% of all dropouts occur at the first page. These respondents start the questionnaire but before answering on question they decide to leave.

The records in the dataset only contain the moment that a respondent leaves the questionnaire. All the questions that the respondents have answered before leaving the questionnaire are not recorded. So, it is not possible to take the answered questions into account in this analysis. Therefore, assume that the probability of leaving the questionnaire only depends on the current state (the last seen and unanswered question on a certain device). Furthermore, routing techniques allow researchers to ask respondents only valuable questions. In order to use routing techniques, some questions are asked more than others. However, the number of times that a question is asked is not recorded in the dataset. Therefore, assume all questions are equally asked.

2.4.3 Dropout rates

To explore dropout triggers, the expected dropout rates are compared with the observed ones. The expected dropout rate is computed as the probability that a random respondent would leave the questionnaire. To explore dropout triggers, the scaled expected and observed dropout rates must be significantly different. To investigate significant differences, the binomial test is used. The binomial test is a test to compare the observed distribution to the expected distribution of only two categories.

The expected dropout rate of leaving the questionnaire before answering a question of type i, and using a device of type j, is computed as

$$P\big(dropout, QuestionType = Q_{i,} Device = D_j\big) = P(dropout) * P(QuestionType = Q_i) * P\big(Device = D_j\big)$$

With $Q_i \in \{A, B, C, D\}$ and $D_j \in \{mobile\ phone, PC, table\}$. For example, the probably that a random respondent becomes a dropout, at one of the preselected question types, is 7.03%, the probability of answering a question of type A is 36.0%, and the probability that a respondent uses a tablet is 11.54%. Then, the theoretical probability that a random respondent would leave the questionnaire, with a question of type A as last seen question, and using a tablet, is $7.03\% * 36.0\% * 11.54\% = 0.30\%$.

The observed dropout rates are calculated by the number of dropouts, belonging to a certain class, divided by the total number of respondents. For example, there are 51 dropouts who left on a tablet with a last seen question of type A and 9919 respondents in the whole dataset then the observed dropout rate is $\frac{51}{9920} = 0.51\%$.

3 Results

In this section the results of the analyses are presented. First, the results of the trend analysis are presented and later the results of the classification techniques and dropout triggers.

3.1 Trend analysis

Trend analysis is done to explore the fractions of devices that are used over a long period of time (2012-2015). Moreover, a prediction is made to explore the fractions of devices that probably will be used in the upcoming two years. In this section, each phase of the Box-Jenkins modeling approach will be discussed, such that eventually a prediction for the fractions of devices for the upcoming years can be made.

3.1.1 Stationary

As mentioned in section 2.2.2.1, the additive log ratio transformation maps compositional data from a 3-dimensional space into a 2-dimenstional space. After transforming observations into a 2-dimensional space, the dataset consist of two time series which both have to fit forecasting models. The first time series is from the transformed tablet users, while the second time series is from the transformed mobile phone users. These transformed observations are shown in the left plot of Figure 3.1. The transformed observations are not yet stationary (ADF-test: p-value = 0.077 and p-value = 0.071). Using the one order differencing method the observations become stationary (ADF-test: p-value < 0.01 and p-value < 0.01).

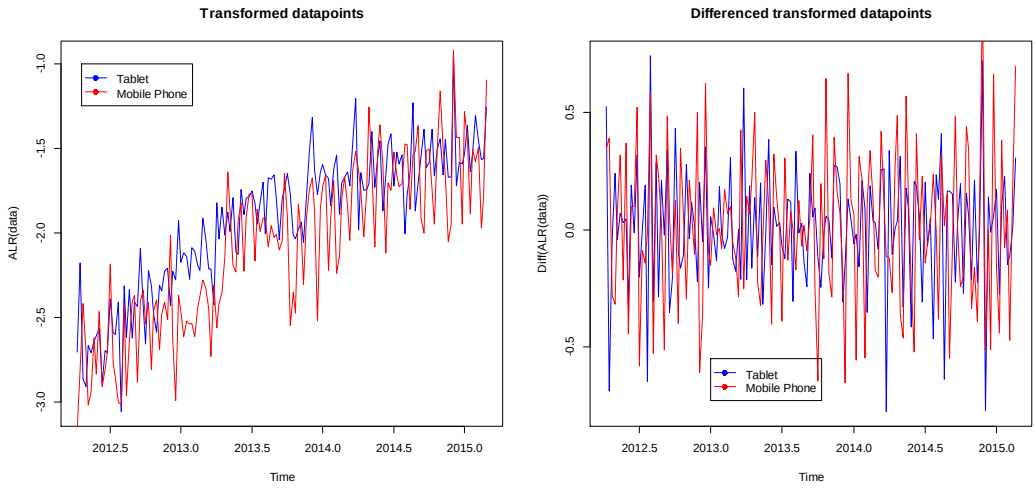

Figure 3.1: The left plot shows the transformed observations, while the right plot shows the differenced transformed observations which are stationary.

3.1.2 Model identification

Two approaches are used in order to select the orders of the ARIMA model. First, the (partial) autocorrelation function is used to examine the order of the ARIMA model. Using the (partial) autocorrelation functions only 75 lags are taken into account, because lags higher than $n/4$ are unreliable [33]. Later an iterative procedure is used for selecting the orders of the ARIMA model. The goal of the iterative procedure is to minimize the goodness of fit statistic for

certain model parameters. In case of exponential smoothing and neural networks, the default parameters are used.

3.1.2.1 ACF and PACF plot

The left plot in Figure 3.2 shows the ACF for tablet. This figure shows that lag 1 (-0.493) exceeds the significant bounds, and all other lags 1, … ,70 do not exceed the bounds. The PACF for tablet shows that lag 1, 2, 3 and 7 exceed the significant bounds, are negative and are slowly decreasing. The absolute values of the ACF are smaller than 0.15 after lag 1 and the absolute values of the PACF are smaller than 0.15 after lag 3. Therefore, the model to use for the time series of tablet users is the ARIMA(3,1,0) model.

The ACF of mobile phone users shows a more cyclical process. In the regular part there is no decay in the AR structure, whereas lags 5, 9, and 13 exceed the significant bounds. This seasonality structure indicates that there is a significant correlation between observations X_t and X_{t+4}. The PACF have same structure as the time series of tablet users, therefore the same findings are used. An appropriate model for the second time series of mobile phone users is the seasonal ARIMA(3,1,2)(1,0,0) model.

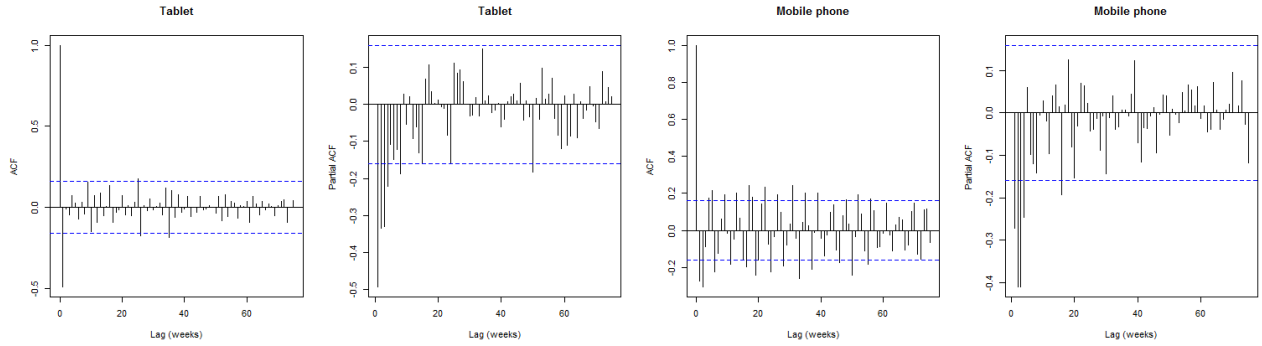

Figure 3.2: Autocorrelation and partial correlation functions for tablet and mobile phone users.

3.1.2.2 Automated iterative procedure

The Automated iterative procedure fits different possible orders to the ARIMA model and uses the goodness of fit statistic (AIC) to select the best orders. A built-in package in R simulates the process of fitting different orders in an ARIMA model (Auto.Arima). Moreover, in this thesis the goodness of fit statistic (AIC) is calculated for all the following possible 147 variations with $\{1, \dots, 7\}$ AR (p) parts, $\{1, 2, 3\}$ integrated (d) parts, and $\{1, \dots, 7\}$ parts of MA(q). The goodness of fit statistic of the models is calculated with and without drift.

Method	Tablet	Mobile phone
Auto.Arima	ARIMA(2,1,1) with drift	ARIMA(3,1,1) with drift
Iterative procedure (with drift)	ARIMA(2,1,1) with drift	ARIMA(3,1,2) with drift
Iterative procedure (without drift)	ARIMA(6,1,5) without drift	ARIMA(5,1,0) without drift

Table 3.1: Model parameters found with the iterative procedures for which the goodness of fit statistic is minimized.

The parameters of the model without drift have very high AR parts, compared to the Auto.Arima model. On the other hand, the found parameters for the model with drift are almost equal to the Auto.Arima model.

3.1.3 Quality of the models

Three different techniques (ARIMA, exponential smoothing and neural networks) are used for predicting future values. The different parameters for the ARIMA model are found in the previous section. In Table 3.2 are the errors shown for all the different models, using the rolling forecasting origin method.

Methods	MAE
ARIMA(2,1,1), ARIMA(3,1,1), With drift	0,045671
ARIMA(2,1,1), ARIMA(3,1,2), With drift	0,045631
ARIMA(6,1,5), ARIMA(5,1,0), Without drift	0,030111
ARIMA(6,2,2), ARIMA(6,0,6), Without drift	0,032732
ARIMA(3,1,1), ARIMA(3,1,2)(1,0,0)	0,03199
HoltWinters(A = 0,224, B = FALSE, G =FALSE), HoltWinters(A=0,253,B=FALSE,G=FALSE)	0,032144
HoltWinters(A = 0,602, B = 0,309, G =FALSE), HoltWinters(A=0,451,B=236,G=FALSE)	0,170238
HoltWinters(A = 0,181, B = 0, G = 0,079), HoltWinters(A=0,303,B=0,G=0,403)	0,049047
NNAR(4,1), NNAR(3,1)	0,033821

Table 3.2: Errors of different forecasting models.

The MAE of the ARIMA(6,1,5), ARIMA(5,1,0) model without drift is lowest (0.30111). Besides, The ARIMA(3,1,1), ARIMA(3,1,2)(1,0,0) model has an almost similar error (0.03199). These two models are used to predict future values, due to the comparable errors.

Forecasting models are adequate when the residuals of a model behave like white noise and are normally distributed with a mean of zero and constant variance. Testing for models with random residuals can be done with the Ljung-Box test. The Ljung-Box test rejects models with non-random residuals.

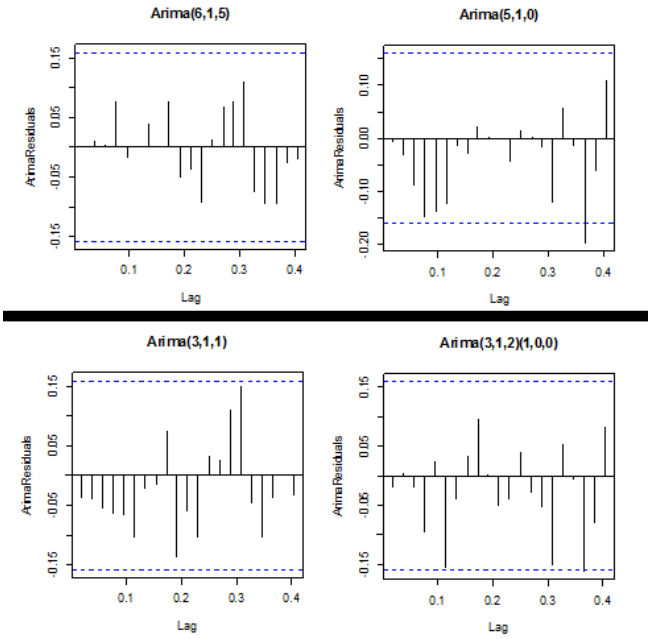

Figure 3.3: The autocorrelation functions of the residuals of different time series models.

Figure 3.3, shows the plots of the ACF of the residuals for different models. The ACF of the residuals for both models (ARIMA(6,1,5) ARIMA(5,1,0) and ARIMA (3,1,1) ARIMA (3,1,2)(1,0,0)) exceeds the significant bounds. Nevertheless, the models are appropriate for forecasting time series (Ljung-box test: p-value = 0.9771, p-value = 0.6077 (tablet) and p-value = 0.6154 and p-value = 0.5789 (mobile phone)).

3.1.4 Forecasts

Predicting future time series values is the last phase in the Box-Jenkins modeling approach. The ARIMA(6,1,5), ARIMA(5,1,0) and ARIMA(3,1,1),ARIMA(3,1,2)(1,0,0) have the highest quality of all models. Furthermore, the residuals are normally distributed with mean zero and a constant variance. ARIMA(6,1,5), ARIMA(5,1,0) have the lowest goodness of fit statistic (AIC) of all compared parameters, while the ARIMA(3,1,1), ARIMA((3,1,2)(1,0,0) is found by examining the (partial) autocorrelation function. The prediction is made only for the upcoming two years, because after these years, the field of consumer technology can be completely changed.

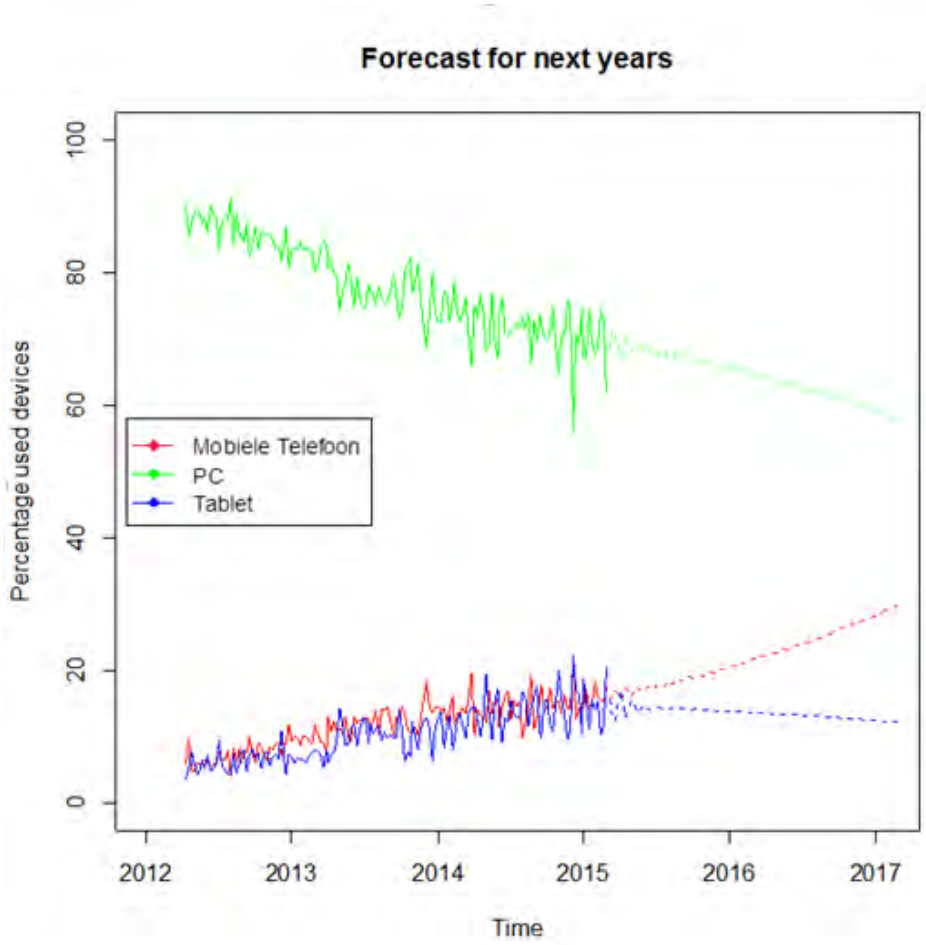

Figure 3.4: Predicted fractions of devices for the upcoming years, using the ARIMA(6,1,5), ARIMA(5,1,0) model.

The predicted future values using the ARIMA(6,1,5), ARIMA(5,1,0) model is shown in Figure 3.4. The fractions of devices that probably are used in the upcoming two years shows a negative trend in the fraction PC users, a positive trend in mobile phone and stabilized trend in tablet users. In 2017, only approximately 57% of the respondents will use a PC to fill out web questionnaires. Contrary to the negative trend of PC use, mobile phone use will increase towards 29% in 2017. In recent years the fraction of tablet users has grown, while the model predicts a stable trend (around 12%). This stabilized trend had started already last year.

The forecast is made by using an ARIMA model with a relative high p order. This high order of p can be the explanation for the positive and negative trend for PC and mobile phone user. A high order of p uses more previous values for predicting future values.

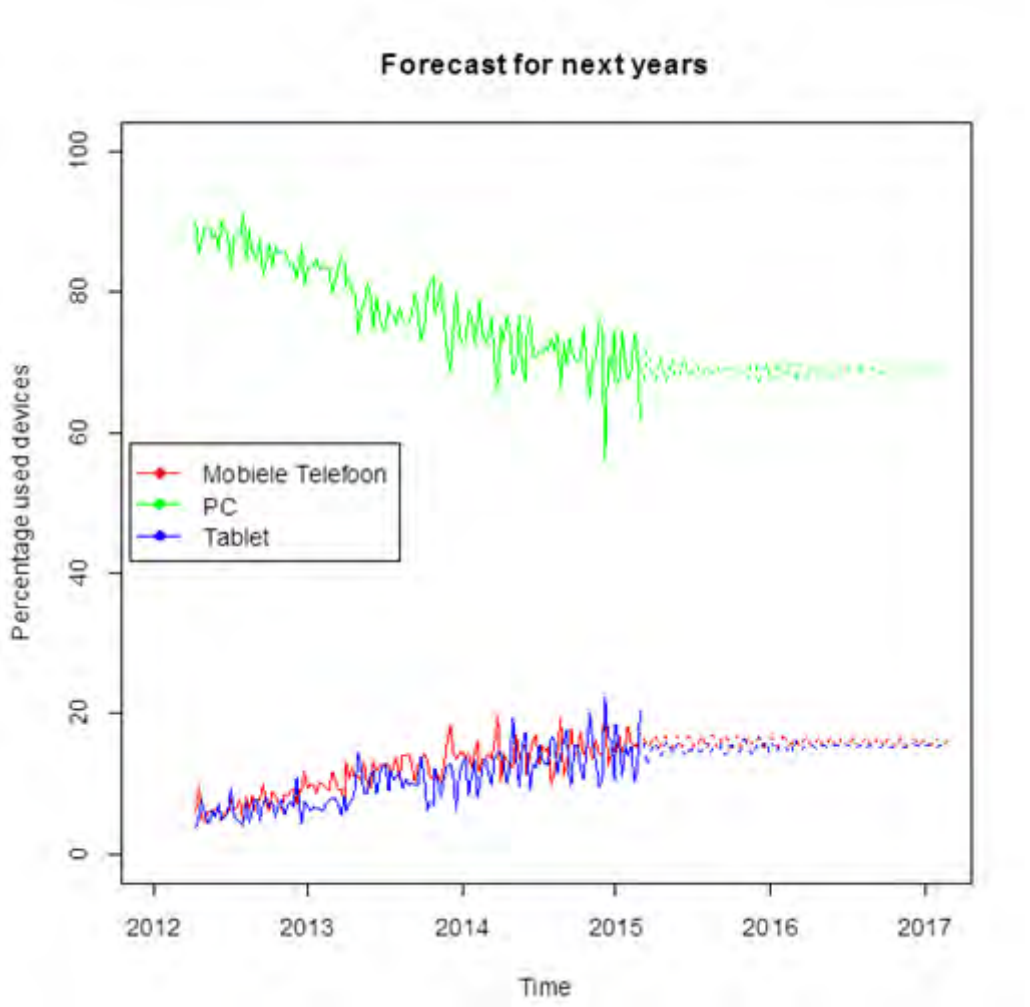

Figure 3.5: Predicted fractions of devices for the upcoming years, using the ARIMA(3,1,2), ARIMA(3,1,1)(1,0,0) model.

The predicted fractions of devices that will be used in web questionnaires for the upcoming two years are shown in Figure 3.5. This forecast uses the ARIMA(3,1,1), ARIMA(3,1,2)(1,0,0) model. This model predicts that the fraction of all devices stays approximately equal between 2015 and 2017.

The order p in the second forecast is smaller compared to the first forecast. Therefore fewer observations are used to describe the time series. The fraction of devices that were used last year (2014) looks like a convex function; the trend for all devices stabilizes after a strong trend. Because of the lower order p and convex function the forecast is stabilized for all different devices.

The difference between the first (ARIMA(6,1,5), ARIMA(5,1,0)) and second (ARIMA(3,1,1), ARIMA(3,1,2)(1,0,0)) model is best shown in the prediction for PC and mobile phone users. The fraction of PC users has decreased between 2012 and 2015; the first model predicts that this negative trend will continue, while the second model predicts a more stabilizing trend. Both of the forecast models predict a stabilized trend for tablet users between 2015 and 2017. This is somewhat unexpected, because of the positive trend between 2012 and

2015. Nevertheless, the first model predicts a positive trend for the fraction of mobile phone users, while the second model predicts a stabilized trend.

Both models have low errors and work accurately on the testset, but I prefer the first forecast (ARIMA (6, 1, 5), (5, 1, 0)). This forecast probably better predicts future values because of the context of the problem. Last year (2014), the amount of smart phones with an internet connection has grown [34]. Also, web questionnaires have become more user-friendly for mobile phone users. On the other hand, the amount of PC users has decreased in recent years, just like the number of sold tablets [34]. Because of these developments, more and more users are likely to use a mobile phone in answering web questionnaires.

3.2 Classification algorithms

Different classification techniques are used for predicting fractions of devices that a corresponding panel will use. In this section, the results from four classification techniques are described. At the end of this section, the quality of the different techniques is examined and the best classifying technique is selected.

3.2.1 Multinomial logistic regression

The multinomial logistic regression (MLR) model acts like a series of logistic regression models. Estimating intercept and regression coefficients are the expected amount of change in the logit for each one unit change in the predictor. So, the closer a coefficient is to zero, the more influence a predictor has on the prediction. The t-test examines whether coefficients are significantly different from zero. In this thesis, the mobile phone is the reference category. However, changing the reference category does not change the quality of the model.

Table 3.3, on the next page, shows the intercept and regression coefficients of the variables. The table shows that the coefficients of some variables are significantly different from zero (***). The predictors are a meaningful addition to the model, because changes in the predictors value are related to changes in the outcome variable. Conversely, predictors with large p-values do not lead to significant changes in the outcome variable. Furthermore, Table 3.3 shows that, for example, age and expected completion time highly influence the outcome variable. However, p-values for variables type and gender are more spread out.

Coefficients :	Estimate	Std, Error	t-value	Pr(>|t|)	
PC:(intercept)	-2,3061800	0,1368103	-16,8568	< 2,2e-16	***
Tablet:(intercept)	-3,8574590	0,188319	-20,4836	< 2,2e-16	***
PC: Gender	-0,2098374	0,04228615	-4,9623	6,97E-07	***
Tablet: Gender	0,0927073	0,05478113	1,6923	9,06E-02	*
PC: Age	0,0487184	0,001444315	33,7311	< 2,2e-16	***
Tablet: Age	0,0526577	0,001847864	28,4965	< 2,2e-16	***
PC: Completion time	0,1014422	0,005699185	17,7994	< 2,2e-16	***
Tablet: Completion time	0,0549990	0,007329562	7,5037	6,20E-14	***
PC: type B	0,4100055	0,1067325	3,8414	1,22E-04	**
Tablet: type B	0,6584181	0,1457763	4,5166	6,28E-06	***
PC: type C	0,5465827	0,1320208	4,1401	3,47E-05	***
Tablet: type C	0,8162119	0,1766642	4,6201	3,83E-06	***
PC: type D	-0,6687632	0,09747977	-6,8605	6,86E-12	***
Tablet: type D	-0,0801035	0,13568628	-0,5904	5,55E-01	
PC: type E	0,0466300	0,09058754	0,5148	6,07E-01	
Tablet: type E	0,4737181	0,12559015	3,7719	1,62E-04	***
PC: type F	1,0212811	0,1039138	9,8282	< 2,2e-16	***
Tablet: type F	0,7552802	0,144311	5,2337	1,66E-07	***
PC: type G	0,2322654	0,09980049	2,3273	1,99E-02	**
Tablet: type G	0,6236384	0,12768861	4,8841	1,04E-06	***
PC: type H	0,9910188	0,1190195	8,3265	< 2,2e-16	***
Tablet: type H	0,2744046	0,191304	1,4344	1,51E-01	*
PC: Months	0,0803251	0,005711929	14,0627	< 2,2e-16	***
Tablet: Months	0,0446408	0,007139152	6,2530	4,03E-10	***

Table 3.3: Estimated intercept and regression coefficients.

The estimated intercept and regression coefficients, whether they are significant or not, are used in order to calculate class probabilities. All the class probabilities taken together yield the predicted fractions of devices. Table 3.4 shows the actual and predicted fractions of devices for the MLR model, which has an MAE of 0.00606. This table shows that the classification model predicts more PC users than the actual ones. Besides, fewer tablet users are predicted.

	Mobile phone	PC	Tablet
Actual fraction of devices	16.559%	69.780%	13.659%
Predicted fraction of devices	16.826%	70.422%	12.751%

Table 3.4: The actual and predicted fraction of devices, based on 20% of the training set.

3.2.2 Decision trees learning

Decision tree learning system uses decision trees as classifier. The complexity parameter, cp, is used for constructing the tree, very small complexity parameters lead to very large trees. The first step is constructing a tree with the smallest possible training error. After that, pruning is used to retain the tree until an optimal tree.

In this thesis, decision trees are constructed with many different complexity parameters, $cp = (1 * 10^{-3}, 7.5 * 10^{-4}, 5 * 10^{-4}, 2.5 * 10^{-4}, 1 * 10^{-4}, \dots, 7.5 * 10^{-6})$. For all these trees, the training error is calculated and shown in Figure 3.6. These errors are spread out over a range from 0.0066 to 0.012. The tree with lowest training error is calculated with $cp = 7.5 * 10^{-4}$.

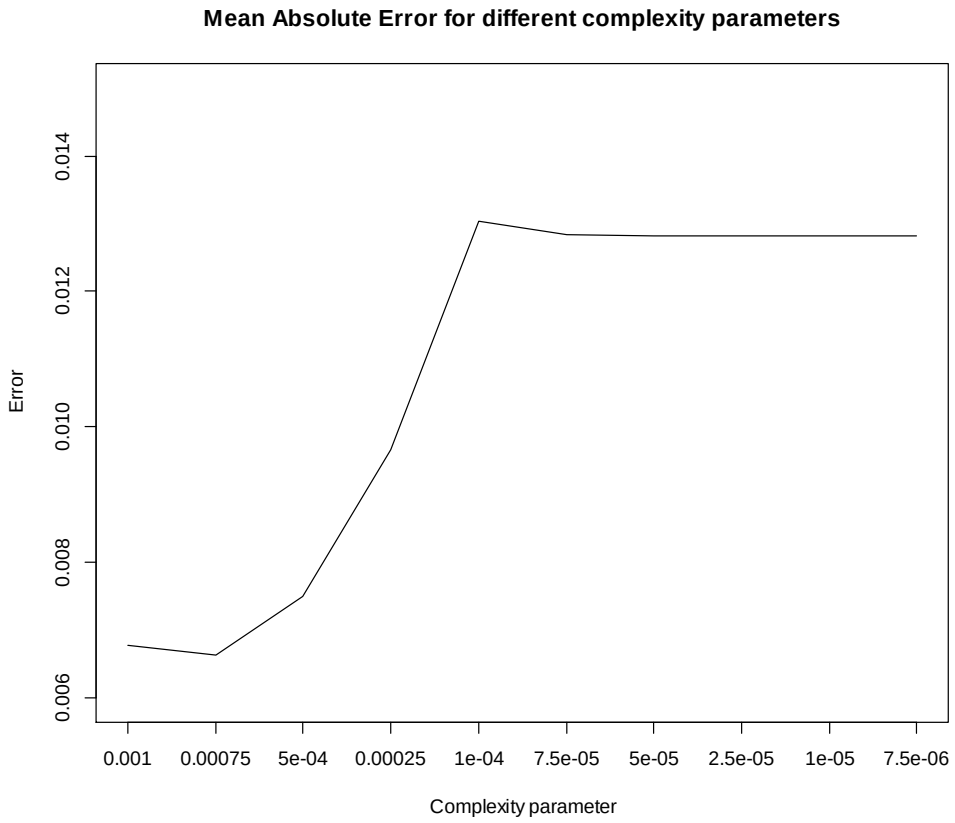

Figure 3.6: The error of constructed trees with different complexity parameters.

Figure 3.7, on the next page, shows the constructed tree with $cp = 7.5 * 10^{-4}$. At the leaf nodes of the tree are the number of devices shown for the respondents who end in that leave. Also the class is shown; this class can be used in regular classification problems. In this tree, an input vector can be split into 10 steps. The most important variable for splitting the attributes is the age of a respondent. 35% of all the 29 splitting points are dependent on the respondents' age. Other remarkable splitting points are type of survey (30%), expected completion time (16%), months passed since a respondent completed the questionnaire (16%), and respondents gender (4%).

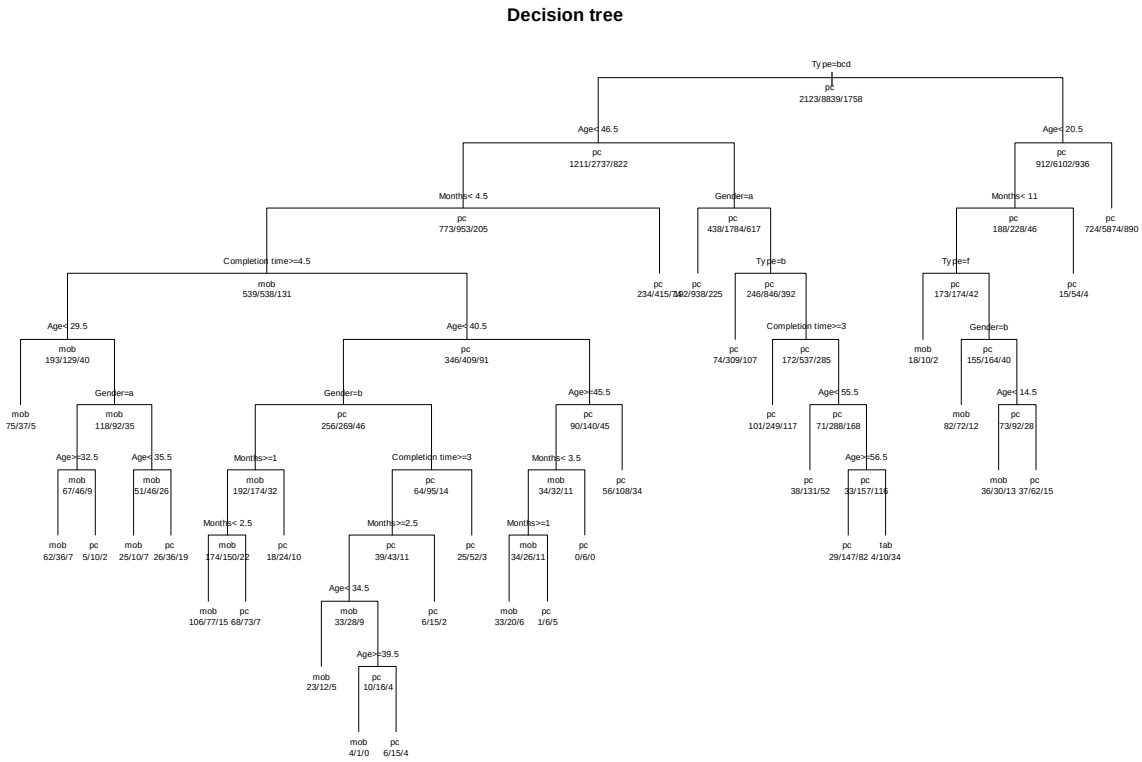

Figure 3.7: Decision tree with $cp = 7.5 * 10^{-4}$

The second phase in constructing an optimal tree is the pruning phase. This pruned tree is shown in Figure 3.8 on the next page. The goal of pruning is to prevent overfitting, and construct smaller trees with similar errors. The pruned tree has only 6 splitting points, significantly fewer than the 29 splitting points of the unpruned tree. The importance of variables has not changed much. Age is the most important variable (41%), followed by type of survey (25%), expected completion time (24%), and months passed since a respondent completed the questionnaire (11%).

The MAE of the pruned tree is slightly larger with 0.0076 than the unpruned tree (0.0066). Table 3.5 shows the actual and predicted fractions of devices. The pruned decision tree predicts higher fractions of PC users, and fewer mobile phone and tablet users.

	Mobile phone	PC	Tablet
Actual fraction of devices	16.559%	69.780%	13.659%
Predicted fraction of devices	16.398%	70.959%	12.642%

Table 3.5: The actual and predicted fraction of devices, based on 20% of the training set.

Although the error margin of the pruned tree is larger, it prevents overfitting. Therefore, the quality of the pruned version of the tree with $cp = 7.5 * 10^{-4}$ will be used and compared with the other techniques.

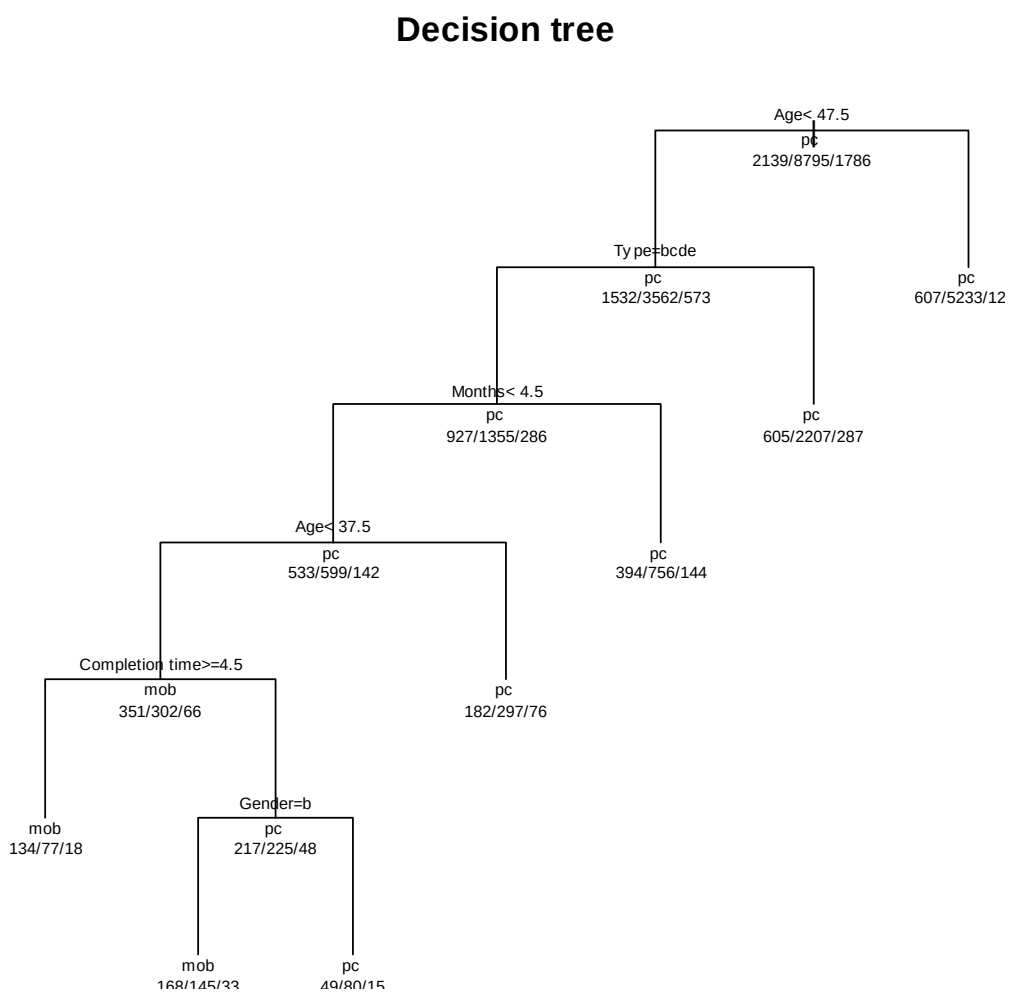

Figure 3.8: Pruned version of the decision tree with $cp = 7.5 * 10^{-4}$

3.2.3 Support vector machines

In support vector machines, four different kernels are used in order to map data into the right dimension. This transformation is necessary because the data is probably not linearly separable. In this thesis, the linear, polynomial, Gaussian radial basis, and hyperbolic tangent sigmoid kernels are used. For selecting the best fitting support vector machines with corresponding kernels, first, the kernel with lowest error is selected and later the optimal combination of parameters is found by using a grid search.

Before selecting the best combination of parameters, the kernel with the lowest error is selected. These kernels use the default parameters $\gamma = 0.333, c = 1$. Table 3.6, on the next page, shows the MAE for support vector machines with different kernels. The hyperbolic tangent sigmoid kernel has the lowest error (0.00899).

Kernel	Linear	Polynomial	Radial Basis	Sigmoid
MAE	0,012793	0,012060	0,0124760	0,008991

Table 3.6: the Mean Absolute Error of different models with default parameters $(\gamma = 0.333, c = 1)$.

Grid search is used in order to find the combination of optimal parameters. The quality of the support vector machines with different parameter combinations is shown in Figure 3.9. This figure shows lowest error rates for a γbetween 1 and 2, combined with a c between 0.125 and 0.25. An error of 0.00800 appears for the optimal parameters, which are $\gamma = 2$ and $c = 0.25$.

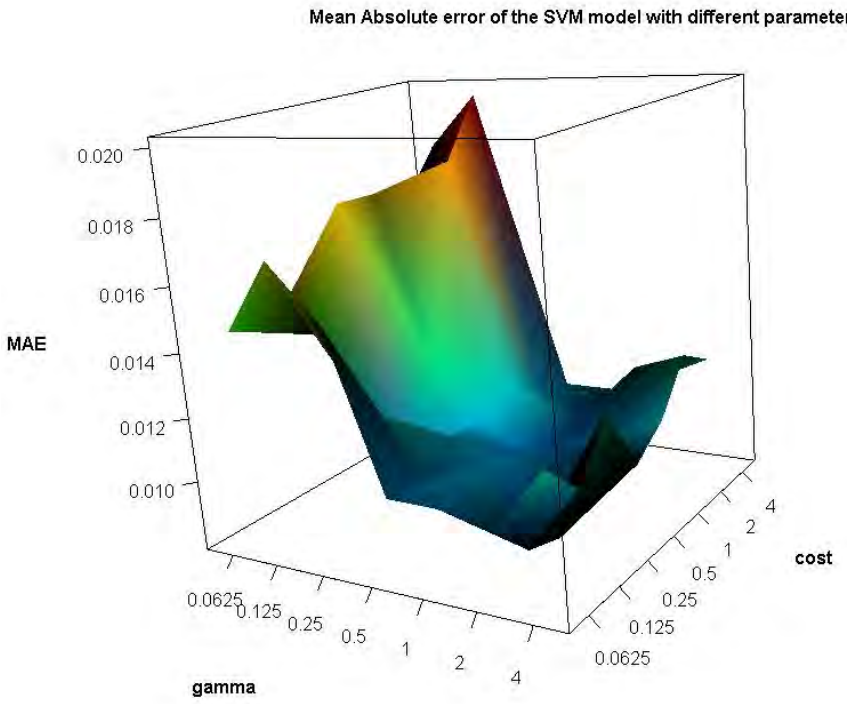

Figure 3.9: The error rate for different combinations of model parameters for a support vector machine with hyperbolic tangent sigmoid kernel.

With the previous results the kernel function that transforms the data is written as

$$\mathrm{K(X, Y)} = \tanh\left(0.25\mathrm{X}^T Y + 2\right)$$

After training the model with the optimal kernel and corresponding parameters, the class probabilities can be calculated. Table 3.7 shows the actual and predicted fractions of devices. This SVM predicts higher fractions of PC users and fewer tablet and mobile phone users.

	Mobile phone	PC	Tablet
Actual probabilities	16.559%	69.780%	13.659%
Predicted probabilities	15.851%	71.0621	13.125%

Table 3.7: The actual and predicted fraction of devices, based on 20% of the training set.

3.2.4 Naive Bayes classifier

In naive Bayes classification, conditional probabilities are used in order to calculate class probabilities. In this section, the assumption of independence between variables is checked first. Secondly, the conditional probabilities are calculated. With these probabilities, the class probabilities can be calculated, which are used for calculating fractions of devices.

The Chi-Squared Test of independence shows no independence (p-value < $2.2\ 10^{-16}$) between variables. Fortunately, Harry Zhang [35] proved that naive Bayes classifiers do not solely work for independent variables. Zhang mentioned that in a dataset two attributes depend on each other, but dependence is distributed evenly in each class. This means that the conditional independence assumption is violated, but naive Bayes is still a proper classifier. Therefore, the naive Bayes classifier is still used in this thesis.

First prior and conditional probabilities are calculated. Table 3.8 shows the predictor prior probability $p(\vec{X})$, which is the probability that a specific device is used. The probability that a random respondent would use a personal computer is 72,03%, while for using a mobile phone it is 15,01%, and for a tablet 12,96%.

Mobile phone	PC	Tablet
0,15012	0,720286	0,129588

Table 3.8: Predictor prior probabilities.

In the following tables the conditional probabilities $P(\vec{X}|C_k)$ are shown for the device classes K. Afterwards, these probabilities are used to calculate class probabilities.

Gender		
	Male	Female
Mobile phone	0,442136	0,557863
PC	0,534691	0,464530
Tablet	0,457889	0,542111

Table 3.9: Conditional probabilities of gender, given the use of device.

The probability that a woman will use a mobile phone or tablet is higher than the probability for men (see Table 3.9). Nevertheless, the probability of using a personal computer is lower for women than for men. Although the difference between these probabilities is small, it is still remarkable that gender influences the use of device.

Age		
	Mean	Standard deviation
Mobile phone	38,20097	15,17067
PC	51,17716	16,56091
Tablet	53,63208	14,88945

Table 3.10: Mean and standard deviation of age, given the use of device.

The age variable is determined as a continuous attribute. Therefore, the normal distribution and standard z-table are used in order to calculate conditional probabilities. As told before, the continuous variables do not fit a normal distribution. Nevertheless, this distribution is used for calculating conditional probabilities. Table 3.10, on the previous page, shows the mean and standard deviation of the respondents' age, given the use of device. The table shows a higher average respondent age for PC and tablet users. The standard deviation is comparable for all devices. It seems logical that respondents who use a mobile phone are much younger than respondents who use a PC or tablet.

Type of Survey								
	A	B	C	D	E	F	G	H
Mobile phone	0,125609	0,127555	0,200584	0,168451	0,112950	0,081791	0,057448	0,125608
PC	0,138041	0,114578	0,098405	0,107517	0,124601	0,142824	0,136446	0,137585
Tablet	0,054678	0,177399	0,172539	0,164034	0,142162	0,083839	0,148238	0,057108

Table 3.11: Conditional probabilities of type of survey, given the choice of device.

Table 3.11 shows the conditional probabilities of answering a questionnaire of a certain type, given the use of device. This table shows low probabilities of tablet users filling out questionnaires of type A or H. Furthermore, the probability of mobile phone users filling out questionnaires of type C and D is relatively high.

Expected completion time		
	Mean	Standard deviation
Mobile phone	7,152671	5,769856
PC	10,174222	7,110858
Tablet	8,110519	5,996710

Table 3.12: Mean and standard deviation of expected completion time, given the choice of device.

Just like age, the expected completion time is a continuous attribute. Table 3.12 shows the mean and standard deviation of the expected completion time (in minutes) of a questionnaire completed on a specific device. The table shows a higher probability for opening questionnaires on a PC, given that the questionnaire has a high expected completion time. Conversely, mobile phone users are more likely to open shorter questionnaires.

Months passed since a respondent has completed the questionnaire		
	Mean	Standard deviation
Mobile phone	8,319288	5,491418
PC	9,031356	4,780467
Tablet	8,440358	5,075730

Table 3.13: Mean and standard deviation of the months passed since a respondent has completed the questionnaire, given the choice of device.

The final attribute in the training set is the time that has passed since the respondents completed the questionnaire. For this attribute is also assumed that it come from a normal distribution, which not realistic. This attribute, which is shown in Figure 3.13, on the previous page, is also a continuous attribute and uses the z-table to become a probability.

At this point the class prior, predictor prior, and conditional probabilities are calculated. Consequently, the class probabilities can be calculated. For each respondent, the probability of using a device is calculated based on their background information. Table 3.14 shows the predicted and actual fractions of devices. The model predicts larger fractions of PC users, and fewer mobile phone or tablet users. The MAE of the naive Bayes classifier is 0.01698.

	Mobile phone	PC	Tablet
Actual ratio of devices	17.506%	68.639%	13.854%
Predicted ratio of devices	16.559%	69.781%	13.659%

Table 3.14: The actual and predicted fractions of devices, based on 20% of the training set.

3.2.5 Testing the quality of the models

In the previous sections the quality of the techniques are validated and the optimal performing parameters are found. Furthermore, the predictions in the previous sections, predicts proportionally more PC users compared to the actual fraction. The reason why it predicts, using each technique, a higher fraction of PC users is not known.

This section compares the quality of the different classification techniques. The training set is used for training the models, while the test set is used to measure the quality of the techniques. The test set consists of records belonging to one survey of each survey type. The quality of the classification technique is measured by comparing the predicted fractions to the actual fractions of devices. Table 3.15 shows the errors that are found by comparing the actual to the predicted fractions, of the whole test set.

type	Multinomial logistic regression	Decision trees	Support Vector machines	Naive Bayes	Mean
A	0,061	0,051	0,049	0,027	0,047
B	0,018	0,026	0,119	0,030	0,048
C	0,021	0,028	0,066	0,042	0,039
D	0,081	0,025	0,161	0,077	0,086
E	0,020	0,024	0,056	0,064	0,041
F	0,407	0,317	0,172	0,252	0,287
G	0,052	0,026	0,045	0,052	0,044
H	0,105	0,085	0,049	0,073	0,078
Mean	0,096	0,073	0,089	0,077	0,084
St.Dev.	0,130	0,101	0,053	0,073	0,084

Table 3.15: The MAE for the predictions of different models on different survey types.

The first things to mention are the high errors for models that attempt to predict the fractions of devices for type F. The error of type F is very high because no pattern is visible in the data. The training set contains three surveys of type F, which show the following fractions of

devices [13% : 70% : 17%], [27% : 57% : 15%], [3% : 94% : 3%] for [Mobile phone : PC : Tablet]. The actual fractions from the survey of the test set have the following fractions [0% : 97% : 3%]. Consequently, all the models predict more mobile phone and tablet users and fewer PC users. Adding more surveys of type F to the training set will probably lead to better results.

With the exception of type F, the quality of the predictions is quite well. Table 3.16 shows the mean and standard deviation of all prediction errors, except the prediction errors of type F. The average prediction error is relatively low, at 0.0547. With an average error of 0.0379, the decision tree classification models work best for this test set. Furthermore, the naive Bayes classifier does not show lowest error margin, but it does have a small standard deviation, which ensures that the predictions will not be very bad. The worst predicting technique is the support vector machine, which is not recommended to use.

	Multinomial logistic regression	Decision trees	Support Vector machines	Naive Bayes	Mean
Mean	0,051	0,038	0,078	0,052	0,055
St. Dev.	0,034	0,023	0,045	0,020	0,019

Table 3.16: The mean and standard deviation of the predictions of all types except type F.

Even when type F is not omitted, the decision tree classification model proves the most accurate. The naive Bayes classification model has an error that is quite similar: 0.0728 versus. 0.0771. However, by omitting predictions of type F, the decision tree classification model works best. Therefore, decision tree classification is recommended to predict the fractions of devices. This recommendation is made based on this dataset. For other training and test datasets, other techniques might work better.

3.3 Dropout rates

This section is divided into two parts. First, time spent as a dropout trigger is examined and later, the question types are examined as dropout trigger. In this study, the time in minutes and the number of answered questions are evaluated as dropout trigger. In case of different dropout triggers for different devices, customized questionnaires could be designed to retain dropouts.

3.3.1 Time as dropout trigger

In Figure 3.10 the proportion of dropouts that occur after an amount of minutes are shown. In the figure, the long questionnaires ($expected\ completion\ time\ \geq 10$) are separated from the shorter ones. Many dropouts (39% for short questionnaires and 25% for long questionnaires) leave the questionnaire before being one minute present in the survey.

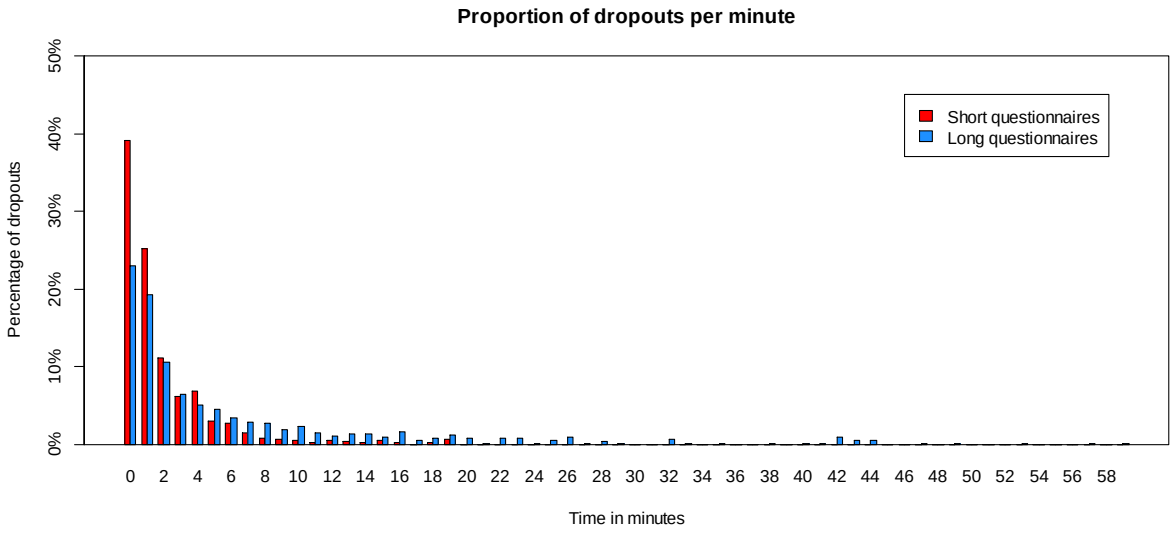

Figure 3.10: The proportion of dropouts that occur per minute.

For short questionnaires, 75% of all dropouts have occurred within 3 minutes, while, 95% of all dropouts leaves the questionnaire within 7 minutes. However, for long questionnaires, 75% of all dropouts occur within 7 minutes and 95% of all dropouts have occurred within 25 minutes. So, the first minutes are the most important to keep respondents in touch. Furthermore, the number of dropouts gradually decreases as the time passes.

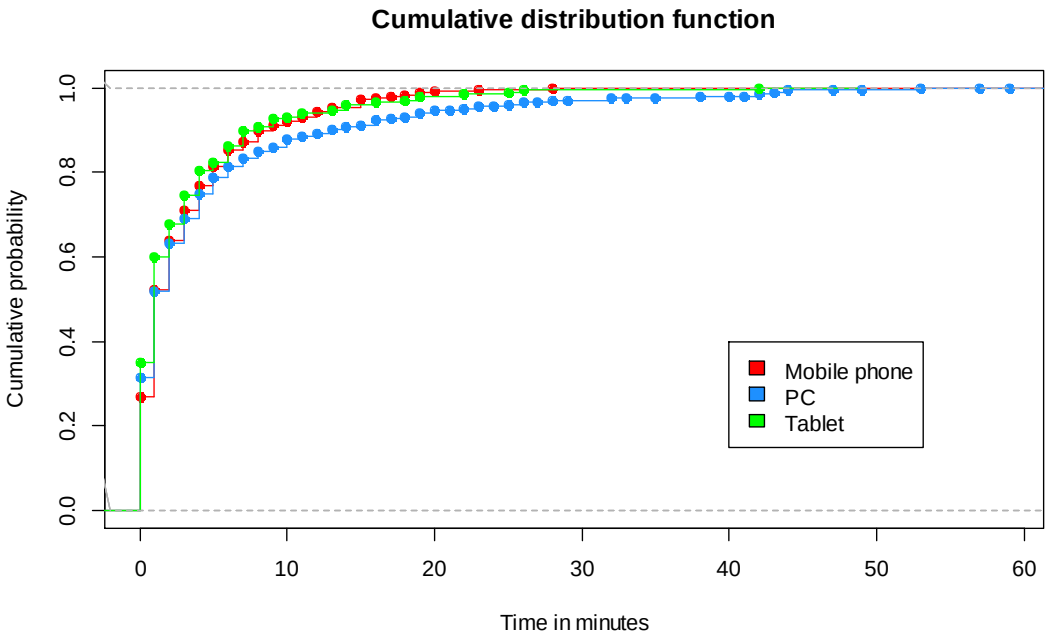

Figure 3.11: Cumulative probability distribution functions for different devices.

To investigate differences between devices, the cumulative distribution functions are used. In Figure 3.11 the cumulative distribution functions for all three types of devices are shown. The cumulative distribution function describes the probability that a random variable is less than or equal to a specified value. In this case, the function describes the probability that a dropout occurs within a specified amount of minutes.

The cumulative distribution function for a mobile phone is almost equal to that for a tablet. However, the cumulative distribution function of the dropouts on a PC is at some points in time lower than on a mobile phone or tablet. This means that a respondent, who uses a PC, will stay longer in touch before leaving the questionnaire. Between 6 and 21 minutes this difference is biggest.

The Kolmogorov-Smirnov test is used to explore significant differences between distributions. This nonparametric Kolmogorov-Smirnov test quantifies the distance between the cumulative distribution function of two samples. The distribution of the test statistic is calculated under the assumption that samples are drawn from the same distribution (the null hypothesis). A low p-value (p-value <0.05) indicates an observation that is very exceptional under this assumption. Hence, a low p-value is reason for rejecting the null hypothesis, and assuming that the underlying distributions of the two samples are different.

The Kolmogorov-Smirnov test cannot distinguish between the dropout time distributions for a mobile phone and tablet (p-value = 0.363). Furthermore, the distribution of dropouts on a PC is also not significantly different from the distribution of dropouts on a mobile phone (p-value = 0.414) or tablet (p-value = 0.245). These p-values, which are all larger than 0.05, show no evidence that the samples come from different distributions.

The second stage is to explore dropouts that occur after a certain number of answered questions. Figure 3.12 shows the proportion of dropouts that occur after a passed number of answered questions. Just like in the previous analysis, dropouts are separated in long and short questionnaires.

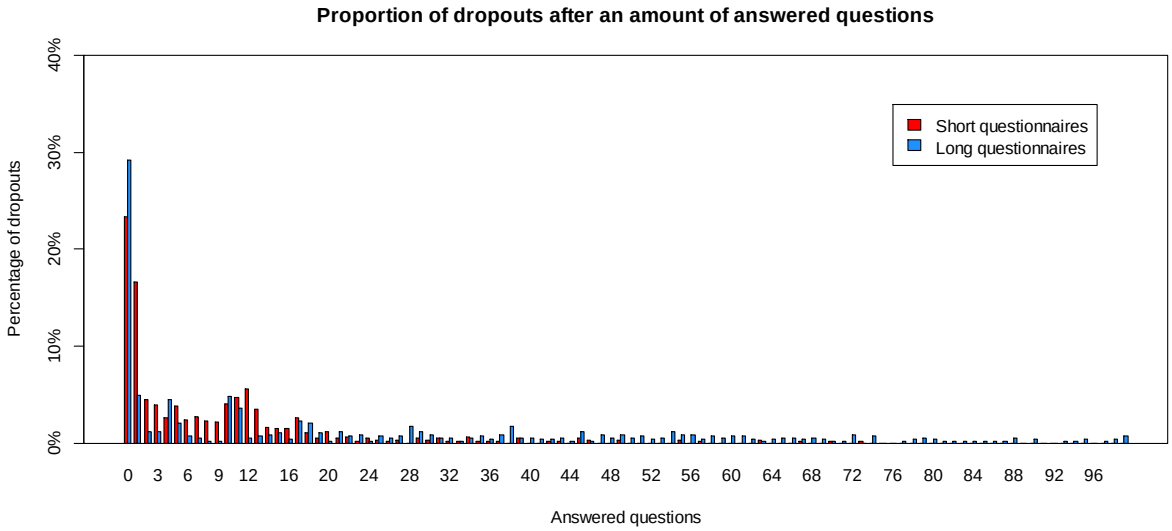

Figure 3.12: The proportion of dropouts that occur after an amount of answered questions.

Figure 3.12 shows that most dropouts occur before answering one question, these respondents who leave before answering one question are called the first question dropouts. Something remarkable are the 30% of dropouts that occur, on long questionnaires, before answering only one question. This pattern is strange, because the proportion of dropouts on short questionnaires is usually higher within the first 15 questions. After answering 15 questions, most dropouts occur on long questionnaires. Besides, some peaks are visible after answering 5, 10 and 18 questions.

In the previous section, no significant differences are found between the time respondents are willing to spend answering questionnaires via different devices. In Figure 3.13, on the next page, the cumulative distribution functions are shown for the probability that a respondent would leave the questionnaire after answering a number of questions.

.

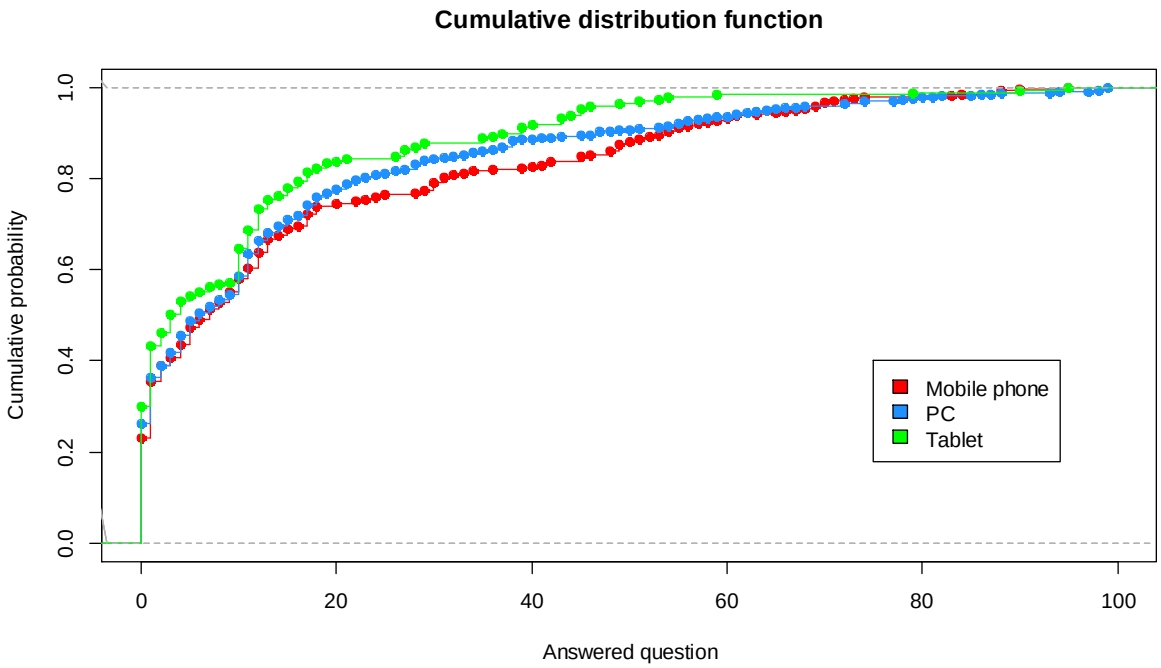

Figure 3.13: Cumulative probability distribution functions of dropouts that occur after a number of answered questions.

The cumulative distribution functions show that mobile phone users answer more questions before leaving the questionnaire. Just like the previous analysis, there are no significant differences found between the cumulative distributions for different devices.

3.3.2 Question type as dropout trigger

Four types of questions are used in this analysis; matrix questions (a matrix of several multiple choice questions) (A), multiple choice questions with only one answer possible (B), multiple choice questions with multiple answers possible (C), and open ended questions (D). Questions that do not belong to these question types are not taken into account.

The amount of dropouts that occur at each question type is shown in Table 3.17, on the next page. The expected and observed dropout rates are also shown in Table 3.17.In this table, dropouts are taken into account that occur on all types of devices. According to this table, most dropouts occur at questions of type B (simple multiple choice questions), followed by type A questions (matrix questions). The expected dropout rate for leaving the questionnaire before answering a question of type A is 3.47%. This expected dropout rate is significantly smaller (binomial test: p-value = 0.007) than the expected dropout rate from 3.97%. As a result, questions of type A lead to more dropouts than predicted, and can be selected as a dropout trigger.

Question type	Dropouts	Asked questions	Expected dropout rate	Observed dropout rate	p-value	
A	394 (41.3%)	134 (36.2%)	3.47%	3.97%	0.0071	**
B	422 (44.2%)	161 (43.5%)	4.18%	4.25%	0.7067	
C	100 (10.5%)	46 (12.4%)	1,21%	1,01%	0.0461	*
D	39 (4.1%)	29 (7.8%)	0.77%	0,39%	2.2E-16	***

Table 3.17: Dropouts, number of asked questions, dropout rates and p-values of binomial test for each question type.

The observed dropout rate is for questions of type A and B larger than the expected dropout rate. Nevertheless, only for type A questions this difference is significant. On the other hand, for questions of type C and D the expected dropout rates are larger than the observed ones. That means that fewer dropouts occur at questions of those types.

In section 2.4.2 it is found that 31.9% of all respondents leave the questionnaire before answering one question. Assume that these dropouts do not occur because of the question type as dropout trigger. These dropouts occur likely because of no interest in the survey, long loading times or other interesting things to do on the web. Therefore, in the following analysis, the dataset is reduced by omitting the first question dropouts.

Table 3.18, is the same table as 3.17, but in 3.18 the data is reduced by omitting the first question dropouts. Eight out of nine first questions appear to be a type B question; the other question is a type A question. After reducing the dataset, the number of type A dropouts dropped from 394 to 342 and type B dropouts from 422 to 186. This huge difference confirms the importance of first question dropouts.

Question type	Dropouts	Asked questions	Expected dropout rate	Observed dropout rate	p-value	
A	342 (51.5%)	133 (36.8%)	2,59%	3,61%	3.71E-07	***
B	186 (28.0%)	153 (42.4%)	2,98%	1,96%	6.71E-10	***
C	100 (15.0%)	46 (12.7%)	0,89%	1,05%	0.09	
D	39 (5.4%)	29 (8.0%)	0,56%	0,41%	0.0436	*

Table 3.18: Dropouts, asked questions and dropout rates and p-values using the binomial test for each question type, by omitting first question dropouts.

Different results are found in Table 3.18 (reduced dataset) than in 3.17 (total dataset). The total dropout rate has dropped from 9.63% (total dataset) until 7.37% (reduced dataset). Furthermore, the positive difference of type B questions in Table 3.17 is changed in a negative significant difference. The positive differences in dropouts of type A en D questions remain equal.

Previously, dropouts are examined that occur on all devices together. Now, dropouts are separated from devices and question types. In Table 3.19 the expected and observed dropout rates are shown, as well as P-values using the binomial test.

Question type	Expected dropout rate	Observed dropout rate	p-value	
Mobile phone				
A	0,47%	0,83%	2.68E-06	***
B	0,54%	0,66%	0.1064	
C	0,16%	0,28%	0.00628	**
D	0,10%	0,07%	0.5164	
PC				
A	1,82%	2,26%	0.0016	***
B	2,10%	1,07%	2.46E-14	***
C	0,63%	0,57%	0.5159	
D	0,40%	0,24%	0.01158	*
Tablet				
A	0,30%	0,51%	0.00067	***
B	0,34%	0,22%	0.04839	*
C	0,10%	0,20%	0.00498	**
D	0,06%	0,09%	0.1997	

Table 3.19: Expected, observed dropout rates and p-values using the binomial test of question types on different devices, by omitting first question dropouts.

Table 3.19 shows, that type A and C questions have significantly more observed dropouts than predicted ones, on a mobile phone and tablet. This means that on mobile devices, answering questions of type A and C lead to significant more dropouts and is selected as dropout triggers. Moreover, Questions of type A are hard to answer on all devices, and are on al devices a dropout trigger.

The analysis shows that dropouts occur most frequently at the first page of the questionnaire. Furthermore, there is a difference (not significant) found in the time that respondents are willing to take answering questionnaires. This difference (not significant) shows that respondents, who use a PC, tend to take more time to answer questionnaires. Moreover, questions of type A lead to more dropouts than expected on all devices, so questions of that question type is a dropout trigger. Besides, type C questions are only dropout triggers on mobile devices (mobile phone and tablet).

4 Conclusion and recommendations

In this section, all the conclusions and recommendations are discussed. First the problem is introduced and the conclusions and recommendations are described. The opportunities to improve the study are described in the section 'further research'.

4.1 Conclusions

In this thesis, a study is done to find opportunities to improve the response of web questionnaires. Investigating differences between answering questionnaires via different devices should probably lead to improvement. The purpose of this thesis can be summarized in the following question:

Is it possible to improve the response of a survey, when a researcher knows in advance the fraction of devices used in the survey?

But before solving the main question, some other problems have to be solved.

First, trend analysis is done to examine which devices are used in responding to web questionnaires. Nevertheless, a prediction is made to figure out in what fraction devices are used during the upcoming two years. This analysis shows the interest of answering web questionnaires via mobile devices. Without sufficient interest, improving the response via customized questionnaires for different devices would not be attractive for researchers.

A computer based tool is created such that it predicts, using a classification model, the fractions of devices that a group of respondents are likely to use. Besides predicting, the computer based tool is also capable of training new input data in such a way that its parameters stay up to date. A major problem is the limited information which is known as input for the classification algorithm.

In the last problem, dropout triggers are explored. There will always be respondents who are not completing the whole questionnaire. However, a researcher tries to minimize these respondents who become a dropout. A dropout trigger is a property of the questionnaire due to which a respondent has a higher probability of leaving the questionnaire. Possible dropout triggers are the time that a respondent has spent and the question type of the last seen question before leaving the questionnaire. Customized questionnaires are only worth to be designed when there are different dropout triggers for different devices.

During the technological development in the past years, more different devices are used in order to respond to web questionnaires. The mobile phone, personal computer (PC) and tablet are the most used devices. During recent years, the use of a type of device in web questionnaires has changed. At the beginning of 2012 95% of all respondents use a PC, while, in 2015 only 65% of all respondents did. In the meantime, more and more respondents are using mobile phones and tablets. The fraction of mobile phone users have increased from 2% to 14% while, the fraction of respondents who uses a tablet has grown from 4% to 16%.

Multiple trend analysis models are used in order to predict the fractions of devices that are likely used in the upcoming years. From all the models with different parameters (ARIMA, Exponential smoothing en neural networks), the two models with highest quality are used in order to predict future time series values. Out of the two best fitting models, the ARIMA(6,1,5), ARIMA(5,1,0) model shows a forecast that fits best the context of the problem. The forecast predicts a positive trend in the fraction of respondents answering questionnaires via a mobile phone, a negative trend via a PC and a stabilized trend via a tablet. These predicted trends are equal with the trend of sold items that belong to the corresponding devices.

Moreover, the choice for the first model can be supported by the results from the classification problem. Predicting fractions of devices for the same panel, but on different time periods, lead to different results. The predicted fraction of PC users is lower for a panel that answers the questionnaire over 5 months compared with the same panel that answers the questionnaire today. This decreasing trend is also shown in the forecast, based on the first model (ARIMA(6,1,5) ARIMA(5,1,0)). Furthermore, the classification techniques provide an increasing trend in mobile phone users by decreasing the time variable. However, a very small upward trend is visible for tablet users, in the classification problem. Besides the small upward trend for tablet users, the classification techniques predict the same pattern as the first forecast model.

Customized questionnaires are only valuable if a large proportion of respondents are likely to use a mobile device. Therefore, the fraction of respondent who are likely to use a certain device is predicted, using a classification model. Since there are only 4.514 unique combinations of background variables within the 22.448 records, in this classification, the classifier assigns class probabilities for all devices to input vectors. Later, these class probabilities are used to predict the fraction of devices that a panel probably would use.

Classifying input variables is done using four classification techniques; multinomial logistic regression, decision tree learning, support vector machines and naive Bayes classifiers. The quality of all models is compared based on the best fitting parameters for each model. Except for one questionnaire type, the techniques predict fractions of devices with relative low errors. The average error of all techniques is 0.055; this error is the average absolute difference between the predicted and actual fractions. All models predict higher fractions of PC users compared to the actual one. Respondents' age influences most the outcome variable. The expected completion time and type of the questionnaire are the second most influencing variables. The gender of a respondent has least influence on the predictions.

Out of all techniques, the decision tree learning system performs best. The naive Bayes classifier has a slightly larger error (0.052 versus 0.038), but a smaller variance (0.0003 versus 0.0005). This naive Bayes classifier prevents wrong predictions, because of the low variance. The decision tree learning system works most accurate, because of the lower error. Out of all techniques the support vector machines has lowest quality, therefore, the support vector machines technique is not recommended to use.

Customized questionnaires, which lead to higher response, should be designed with the least possible dropout triggers. For this purpose, two types of dropout triggers are examined: the time spent and question type. Most dropouts occur in the first few minutes of a questionnaire; already 53% of all dropouts have left the questionnaire in the first two minutes.

Moreover, 31% of all dropouts occur before answering the questions on the first page. Only a small and not significant difference is found in the time that a respondent is willing to take answering a questionnaire via different devices. This shows that respondents, who use a PC, tend to take more time to answer questionnaires.

In this thesis, the following four question types are examined as dropout trigger; matrix questions (a matrix of several multiple choice questions) (A), multiple choice questions with only one possible answer (B), multiple choice questions with multiple answers possible (C), and open ended questions (D). There are too little dropouts on other question types; therefore other question types are not taken into account. It is found that a lot of dropouts occur at the first question. These dropouts do not occur because of the type of the question. Therefore, the first question dropouts are omitted in this analysis.

For exploring dropout triggers, the difference between theoretical and empirical dropout rates is examined. When the empirical dropout rate is significant larger than the theoretical one, then the question type is harder to answer and is classified as a dropout trigger. The analyses shows that on all devices, type A questions are harder to answer, and selected as a dropout trigger. On the other hand, type B and type D questions are easy to answer and lead to fewer dropouts.

By separating dropouts for different devices comparable results are shown as in the overall case. Answering questions of type A and C is significantly harder to answer on a mobile phone, and is classified as a dropout trigger. Furthermore, question type A is a dropout trigger on a PC. Responding to questionnaires via a tablet, shows that question types A and C are dropout triggers. Between devices, only differences are found in the question of type C, which is only a dropout trigger on a mobile phone and tablet.

The analysis of the dropout triggers shows that creating customized questionnaires for different devices is complicated. The time as dropout trigger shows no significant differences between devices. So, creating questionnaires with different lengths, for different devices, would probably not lead to higher response. More interesting is the question type as dropout trigger. That analysis shows that a question of type C is only a dropout trigger on a mobile device and not a PC. Other question types, as dropout trigger, are equal for all devices. So customized questionnaires could only be created by omitting type C questions on mobile phone or tablet. By means of an experiment it can be tested whether these customized questionnaires indeed lead to higher responses.

To conclude, there is enough interest in answering web questionnaires via mobile devices. Also, it is possible to predict the fractions of respondents who are likely to use a mobile device. However, designing customized questionnaires for different devices, which lead to higher response, is complicated. Further research has to be done to investigate more dropout triggers, which makes it possible to create customized questionnaires.

4.2 Further research

The analyses in this thesis have some shortcomings, which probably could be solved with some additional research. This additional research could make it possible to improve the response of a survey by designing customized questionnaires.

Right now, the classification model works quite well with the limited information. Improving this model could be done by adding extra background variables. For example variables like education and income level of a respondent, but also the type of panel that answers the questionnaire. At this moment, these variables are not known and are therefore not used. With adding the variables in the dataset, the classification error would probably decrease.

Further research should definitely be done, to investigate more dropout triggers for different devices. The analysis to explore dropout triggers is a very small part of this thesis, and has therefore limited results. However, a bigger research can lead to more differences between dropout triggers on different devices. For example the whole process can be modeled in a network to investigate more dropout triggers. The amount of times a question is asked can then be taken into account by calculating dropout rates. With more different dropout triggers it could be possible to design multiple questionnaires for different devices, which could lead to higher response. Then, by meaning of an experiment can be tested whether customized questionnaires improve the response of a questionnaire.

Another possible method to investigate dropout triggers is to create a score function which will be updated after every answered question. When a questionnaire have a high score function, the questionnaire is designed such that more respondents drop out. The score function takes the already answered question as well the sequence of asked questions into account. A researcher could then create a questionnaire such that this score functions is minimized.

5 References

[1] Snijders, D., Matzat, U., „Gebruiksvriendelijkheid van webenquetes," ISIZ, Amsterdam, 2005.

[2] Snijders, D., Matzat, U., „Het effect van voortgangsbalken op de completion rate van online vragenlijsten," ISIZ, Amsterdam, 2005.

[3] Haak, D., „De invloed van invultijd op de uitval bij online enquêtes," ISIZ, Amsterdam, 2010.

[4] Sheehan, B. K., McMillan, J. S., „Response variation in e-mail survey: an exploration," *Journal of Advertising,* vol. 39, nr. 4, pp. 45-54, 1999.

[5] Yun, G. W., Trumbo, C. W., „Compartive response to a survey executed by post e-mail and web form," *Journal of computer mediated communication,* vol. 6, nr. 1, pp. 1-25, 2000.

[6] Cobanoglu, C., Warde, B., Moreo, P., „A comparison of mail, fax and web-based survye methods," *International Journal of Market research,* vol. 43, nr. 4, pp. 441-452, 2001.

[7] Sheehan, B.K., „E-mail Survey response rate: A Review," *Journal of cumputer mediated communication,* vol. 6, nr. 2, p. 0, 2006.

[8] Markovitch, D.G., „Comparing online and mail survey methods in a sample of chief marketing officers," *Inovative marketing,* vol. 5, nr. 4, pp. 55-62, 2009.

[9] Millar, M.M., Dillman, D.A., „Encouraging survey response via smartphones: Effect on respondents' use of mobile devices an survey response rates," *Survey practice,* vol. 5, nr. 3, pp. 1-6, 2012.

[10] Mavletova, A., „data Quality in PC and Mobile Web Surveys," *Social Science Computer Review,* vol. 31, nr. 6, pp. 725-743, 2013.

[11] La Bruna, A., Rhathod, S., „Questionnairre length and fatigue effects," Survey Sampling International (SSI), Rotterdam, 2010.

[12] Porter, S.R., Whitcomb, M.E., „The impact of contact type on Web Survey Response Rates," *Public Opinion Quarterly,* vol. 67, nr. 4, pp. 579-589, 2001.

[13] Dillman, D.A., Mail and Telephone, New York: John Wiley & Sons, 1978.

[14] Dillman, D.A., Mail and internet Surveys: The Tailered Design Method, 2nd red., New York: John Wiley & Sons, 1999.

[15] Tarkus, A., „Usability of Mobile Surveys," *Mobile Market Research,* vol. 7, p. 154, 2009.

[16] Couper, M.P., „Visual design in online surveys:Learning for the mobile world," in *The mobile research conference*, London, 2010.

[17] Fuchs, M., Busse, B., „The coverage bias of mobile web surveys across European countries," *International journal internet science,* vol. 4, nr. 1, pp. 21-33, 2009.

[18] MacElroy, B., „Variables influencing dropout rates in Web-based surveys," *Quirk's Marketing Research Review,* 2000.

[19] Brent, C., „Does adding one more question impact survey completion rate?," 12 August 2010. [Online]. Available: https://www.surveymonkey.com/blog/2010/12/08/survey_questions_and_completion_rates/.

[20] Aitchison, J., „The statistical analysis of compositional data.," *Journal of the Royal Statistical Society,* vol. 44, nr. 2, pp. 137-177, 1986.

[21] Koehler, A.B., Snyder, R.D., Ord, J.K., Beaumont, A., „Forecasting compositional time series with exponential smoothing methods," Monash University, Melbourne, 2010.

[22] Box, G.E.P., Jenkins, G.M., Time series analysis: forecasting and control, 2nd red., San Fransisco: Holden Day, 1976.

[23] Pouzols, F.M., „Effect of different detrending approaches on computational intelligence modles of time series," in *WCCI 2010 IEEE World Congres on computational intelligence*, Barcelona, 2010.

[24] Rodriguez, G.,, Lecture notes on Generalized Linear Models, Princeton University, 2007.

[25] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., Classification and regression trees, Belmont: Wadsworth statistics, 1984.

[26] Quinlan, J.R., „Introduction of decision trees," *Machine learning,* vol. 1, nr. 1, pp. 81-106, 1986.

[27] Quinlan, J.R.,, Programms for machine learning, San Fransisco: Morgan Kaufmann, 1993.

[28] Vapnik, V., Lerner, A.,, „Pattern recognition using generalized portrait method," *Automation and Remote Control,* vol. 24, pp. 774-780, 1963.

[29] Vapnik, V., The nature of statistical learning theory, New York: Springer, 1995.

[30] Burges, C.J.C.,, „A tutorial on Support Vector Machines in Pattern Recognition," *Data Mining and Knowlegde discovery 2,* pp. 121-167, 1998.

[31] Boser, B.E., Guyon, I.M., Vapnik, V.N., „A training algorithm for optimal margin classifiers," in *Proc. 5th Annual ACM Workshop on Comput. Learning Theory*, New York, ACM Press, 1992, pp. 144-152.

[32] Laplace, P.S.„ Theorie de analytique des probabilities, Paris, 1812.

[33] Box, G.E.P., Jenkins, G.M., Time series analysis: forecasting and control, 1st red., San Fransisco: Holden Day, 1970.

[34] G. group, „Evenveel nederlanders met tablet als vaste computer," 12 December 2014. [Online]. Available: http://www.gfk.com/nl/news-and-events/press-room/press-releases/Paginas/Evenveel-Nederlanders-met-tablet-als-vaste-computer.aspx.

[35] Zhang, H., „The optimality of Naive Bayes," in *FLAIRS Conference*, Miami, 2004.

[36] Koehler, A.B., Hyndman, R.J., „Another look at measures of forecast accuracy," *International journal of forecasting,* vol. 22, nr. 4, pp. 678-688, 2005.