Selection models for information channels

Applications in multichannel digital marketing

Nils Eckerdal and Victor Winberg
Supervisor: Shaobo Jin
Uppsala University
The department of statistics
Spring 2015
Abstract

In all digital marketing efforts, different information channels must be selected and used to reach customers. In this thesis, data describing the interactions that members of the loyalty program of a Nordic airline company have had with three information channels is used to estimate four selection models. These multinomial logistic regression models have the purpose of selecting which channel(s) best suits a given member. The models are evaluated and the one that best fits the given situation is examined in detail. It is found that the estimated models have satisfactory predictive power in estimating which channel best suits a given customer, and that this method can have useful practical applications.

**Key words:** multinomial logistic regression; predictions; model estimation; customer engagement.
Table of contents

Abstract ........................................................................................................................................... 1

1. Introduction .................................................................................................................................. 4
   1.1 Background .............................................................................................................................. 4
   1.2 Structure of the thesis .............................................................................................................. 5
   1.3 Research question and purpose .............................................................................................. 6

2. The data ......................................................................................................................................... 7
   2.1 Variable descriptions ................................................................................................................ 7
      2.1.1 Channel ............................................................................................................................. 7
      2.1.2 Sex ..................................................................................................................................... 8
      2.1.3 Nationality ......................................................................................................................... 8
      2.1.4 Age ..................................................................................................................................... 8
      2.1.5 Source ................................................................................................................................ 9
      2.1.6 Time ................................................................................................................................... 10

3. Method .......................................................................................................................................... 11
   3.1 Binary logistic regression ......................................................................................................... 11
   3.2 Multinomial logistic regression ............................................................................................... 12
   3.3 Interpretation of the parameter coefficients ............................................................................. 12
   3.4 Assumptions ............................................................................................................................ 13
   3.5 Goodness-of-fit statistics ......................................................................................................... 13
   3.6 Overdispersion ........................................................................................................................ 13
   3.7 Multiple split evaluation ......................................................................................................... 14

4. Results .......................................................................................................................................... 15
   4.1 Model specification and selection ........................................................................................... 15
      4.1.1 Predictions ......................................................................................................................... 15
      4.1.2 Statistical properties of the models .................................................................................... 16
   4.2 Test of assumptions and sample considerations ....................................................................... 17
      4.2.1 Sample size considerations and overdispersion ............................................................... 17
   4.3 Interpretations of the multinomial logistic model ................................................................. 18
      4.3.1 Test of the global null ......................................................................................................... 18
      4.3.2 Maximum likelihood estimates ......................................................................................... 18
      4.3.3 Interpretations of maximum likelihood estimates ............................................................ 20
      4.3.4 Odds ratios ....................................................................................................................... 22
4.4 Alternative rule for channel selection ..............................................................24

5. Discussion ..........................................................................................................26

5.1 Applications of the model ...............................................................................26
5.2 Potential limitations of the model ....................................................................26
5.3 Suggestions for further research .....................................................................27

6. Final conclusions ...............................................................................................28

References ...............................................................................................................29
1. Introduction
1.1 Background
In order to increase sales and profits, companies all over the world are trying to reach customers with information and offers that might interest said customers in purchasing new goods and services. In the context of online customer relationship management, there is a rising attention to the concept of *customer engagement* (Singh and Kumar, 2010). With the developments of recent years in internet trends, digital platforms are changing from static places where information can be found to webs of interactive sites where the customers can actively engage in a large set of activities. Given the fact that customers may at any time leave companies and their brands, using the customers’ engagement is an important key in retaining customers and recruiting new ones (Singh and Kumar, 2010).

Three levels of customer engagement has been found (Oriesek, 2010), based on the level of energy the customers spend in their engagement with a brand. The first level, *viewers*, are estimated to account for about 75 % of the internet users and have a relatively low level of engagement and weak emotional ties to brands. They focus their activities on finding good deals rather than participating in customer discussions. Most viewers find insistent attempts of contact from the company annoying. The second level of customer engagement, *contributors*, account for about 20 % of internet users and take a more active role in web activities than viewers. Their activities are characterized by commenting the quality of products and services through ratings and comments and through sharing of information and deals. The last level of engagement, *creators*, account for only about 5 % of users. They create content that benefit both the company and other customers and are willing spend considerable time investment in choice brands.

These customer groups are of gradually greater importance for a firm. Creators have the ability to create auxiliary services that can greatly promote the web presence of a company, especially with the attention of collaborators who share and spread knowledge of products and services. Catching the attention of viewers is important in its own right, since they compose the majority of internet users and thus a majority of potential customers (Oriesek, 2010).

This thesis is concerned with the first level of customer engagement, viewership. Companies all over the world are trying to reach customers at this level to present offers and information in order to increase sales and revenues. Sales cannot come at the price of reduced profit, however, and as such it is important to keep the marketing efforts used to create customer engagement cost efficient. One way to do this is to limit the number of channels used to contact customers, using only the select channels where specific customers are most likely to engage with and obtain the information sent. To an increasing extent, this effort to reduce costs by using channels selectively is also true for digital information channels (Informant 1, 2015). Highly competitive companies are now looking for ways to be more selective in the information they send to cut costs and increase efficiency while retaining maximum effectiveness from their digital marketing efforts.
In this context, channel selection is an important topic for practitioners. Well planned channel selection can decrease marketing costs without strong adverse effects on sales. The question of whether to contact a customer through a specific direct marketing channel is older than digital information channels, however. There is a large body of research on selection models for methods such as direct mail, where marginal costs have always been an issue (see for example Baesens et al. 2002, Bijmolt et al. 2010 and Bult 1993). Selection models are used to select which customers the company should select to contact via direct mail.

In the current study, three digital information channels used by the company are examined: a mobile app, a webpage and e-mail communication. In order to work on the issue of channel selection that the airline company is facing, new models are created and evaluated. This is done by drawing on the body of knowledge available regarding direct mail selection models. Baesen et al. (2002) used Bayesian neural networking and direct mail marketing data in order to try to classify if customers would repurchase or not and which kind of techniques could be used to estimate this. They provided new insights into the usefulness of said techniques in that particular context. However, there are some differences between direct mail selection models and the models proposed in this thesis. The models proposed here are not binary (send mail or do not send mail) and they are not designed to select which customers that are to be contacted and which customers that will not be. Instead, the models created in this thesis has more than two response categories (since more than two digital channels are examined) and their purpose is to select which channel fits a given customer, rather than the other way around. Based on the dataset provided by the airline company, logistic regression models are estimated and evaluated.

1.2 Structure of the thesis
The structure of the thesis will be as follows: In Section 2 the data set is presented. Univariate and multivariate descriptive statistics as well as comprehensive explanations of any independent variables that is used for any or all models are presented. Then, Section 3 aims to explain the statistical methodology that is used in order to try to answer the research question. Here the multinomial logistic regression approach is explained. Also, potential problems with this particular technique is discussed. In Section 4, the results of the multinomial logistic regression is presented. Specifically, the different models that are deemed useful are presented and commented on and the best model is chosen for further investigation. Furthermore, in Section 5, the results of the study are discussed and analyzed from a point of view relevant to the purpose of the thesis, namely customer engagement and statistical methodology in a business context. Lastly, in Section 6, the conclusions drawn from the results are summarized and commented on.
1.3 Research question and purpose

The purpose of this study is to propose a model that can be used to find the most probable digital information channel to use given a certain customer. Such a model would enable the airline company to reduce costs by sending information via the digital information channel that is most probable that a given customer uses. If this model could help distinguish demographical patterns in channel use, this could also prove helpful in a business context.

Furthermore, from a scientific point of view it is interesting to estimate this particular kind of model using established statistical methods and therefore applying said models to a business context.

Rephrasing this purpose into a pair of research questions can be done as follows:

A. *Which digital information channel is most probable to be preferred by a given customer?*

In other words, given information on a particular customer, which channel is best used to reach him or her?

B. *Which demographic groups in the studied dataset are best reached through which type of digital information channel?*

It is also possible to investigate whether particular demographic groups among the customers of the herein studied airline company are more likely to use certain channels than other such groups.
2. The data
The purpose of the study has been developed in collaboration with the airline company in the case study. In order to be able to answer the research question, a sample of 28,922 customers for the Nordic countries, Sweden Denmark, and Norway, forms the basis of the examined data. The dataset contains variables describing the interactions these customers have had with each of the three digital information channels between 2014-12-18 and 2015-03-31. Further, background variables of these customers were provided by the airline company. All customers in the dataset are anonymous. The data was extracted early in April, 2015.

The customers in the dataset are all part of the membership loyalty program of the airline company. Further, only members who do not have a specifically negotiated company contract with the airline company is included. These restrictions, along with the restriction of customers being Norwegian, Swedish, or Danish, are applied to the population and inference outside of the population of the study should be made with caution.

2.1 Variable descriptions
The following variables are included in the dataset and will be briefly described in the Data section: Channel, sex, age, nationality and source (of registration).

2.1.1 Channel
The dependent variable used in the logistic regression is which channel(s) a customer uses. This variable has four categories: one for each channel and one denoting customers using more than one channel. Table 2.1 is a frequency table over this variable.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A mobile app</td>
<td>762</td>
</tr>
<tr>
<td>The webpage</td>
<td>4218</td>
</tr>
<tr>
<td>E-mails</td>
<td>776</td>
</tr>
<tr>
<td>More than one channel</td>
<td>320</td>
</tr>
<tr>
<td>Total</td>
<td>6076</td>
</tr>
</tbody>
</table>

As can be seen in table 2.1, only 6076 of the 28,922 customers in the sample used any of the three channels during the recorded time period. Since logistic regression is used, and the response variable need to be observed for estimation and evaluation of the models, only the subset of the sample that used a digital information channel are used in to that end. This limits the generalizability of the model and has been taken into account when in the inference and when conclusions have been drawn. In Section 2.1.2 to 2.1.6, only this smaller subset of 6076 customers is described. As can be seen in table 2.2, channel use is rather unevenly distributed over nationality, Swedes seem to like web and app whereas Danes like do use the e-mail and the Norwegians multiple use of the channels.
Table 2.2 Percentage of channel use for each nationality and in total

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Channel</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>App</td>
<td>Mail</td>
<td>Web</td>
<td>Multiple</td>
</tr>
<tr>
<td>Swedish</td>
<td>20.06 %</td>
<td>11.15 %</td>
<td>62.93 %</td>
<td>5.86 %</td>
</tr>
<tr>
<td>Norwegian</td>
<td>7.18 %</td>
<td>4.53 %</td>
<td>83.73 %</td>
<td>4.57 %</td>
</tr>
<tr>
<td>Danish</td>
<td>14.08 %</td>
<td>26.70 %</td>
<td>53.41 %</td>
<td>5.81 %</td>
</tr>
<tr>
<td>Total</td>
<td>12.54 %</td>
<td>12.77 %</td>
<td>59.42 %</td>
<td>5.27 %</td>
</tr>
</tbody>
</table>

2.1.2 Sex
The sex of the customers was one of the background variables provided. 50.39 % of the customers were male, the rest female. The channel use was fairly evenly distributed among the customers, as can be seen in table 2.3.

Table 2.3 Channel use among men and women

<table>
<thead>
<tr>
<th>Channel</th>
<th>Sex</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td></td>
<td>Female</td>
<td></td>
</tr>
<tr>
<td>App</td>
<td>492</td>
<td>(16.07 %)</td>
<td>270</td>
<td>(8.96 %)</td>
</tr>
<tr>
<td>Mail</td>
<td>404</td>
<td>(13.19 %)</td>
<td>372</td>
<td>(12.34 %)</td>
</tr>
<tr>
<td>Web</td>
<td>1986</td>
<td>(64.86 %)</td>
<td>2232</td>
<td>(74.05 %)</td>
</tr>
<tr>
<td>Multiple</td>
<td>180</td>
<td>(5.88 %)</td>
<td>140</td>
<td>(4.64 %)</td>
</tr>
<tr>
<td>Total</td>
<td>3062</td>
<td>(100 %)</td>
<td>3014</td>
<td>(100 %)</td>
</tr>
</tbody>
</table>

2.1.3 Nationality
This study has been restricted to include only Swedish, Norwegian and Danish customers. Each nationality is equally represented in the sample and make up almost exactly one third of the sample. For the sample subset of customers who have used any of the channels during the recorded period, the distribution is shown in table 2.4.

Table 2.4 Frequencies of nationalities

<table>
<thead>
<tr>
<th>Nationality</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swedish</td>
<td>1570</td>
</tr>
<tr>
<td>Norwegian</td>
<td>2716</td>
</tr>
<tr>
<td>Danish</td>
<td>1790</td>
</tr>
<tr>
<td>Total</td>
<td>6076</td>
</tr>
</tbody>
</table>

Worth noticing is that a lot more Norwegians than Danes or Swedes have used any or multiple channels (Table 2.2).

2.1.4 Age
The age of the customers was also included in the dataset. The eldest member of the dataset was 105 years old by the time the dataset was collected. The youngest was one year old (it is possible
to register children to the membership loyalty program). A histogram of the age of the customers can be seen below.

![Distribution of age](image)

**Figure 2.1. Histogram over age**

### 2.1.5 Source

Source is the name of the variable that describes the channel that each customer used to register to the airline’s loyalty program. This variable has five categories: online, the iOS app, the android app, through a partner of the airline or through a call center. Table 2.5 shows the frequencies of each source.

<table>
<thead>
<tr>
<th>Source</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online</td>
<td>3161</td>
</tr>
<tr>
<td>iOS app</td>
<td>1435</td>
</tr>
<tr>
<td>Android app</td>
<td>226</td>
</tr>
<tr>
<td>Partner</td>
<td>77</td>
</tr>
<tr>
<td>Call center</td>
<td>1177</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6076</strong></td>
</tr>
</tbody>
</table>
Table 2.6 Channel use among members registered through an app and other members

<table>
<thead>
<tr>
<th>Channel</th>
<th>Source: an app</th>
<th>Source: other</th>
</tr>
</thead>
<tbody>
<tr>
<td>App</td>
<td>40.94 %</td>
<td>1.86 %</td>
</tr>
<tr>
<td>Mail</td>
<td>14.33 %</td>
<td>12.19 %</td>
</tr>
<tr>
<td>Web</td>
<td>33.53 %</td>
<td>3.04 %</td>
</tr>
<tr>
<td>Multiple</td>
<td>11.20 %</td>
<td>82.92 %</td>
</tr>
<tr>
<td>Total</td>
<td>100 %</td>
<td>100 %</td>
</tr>
</tbody>
</table>

One variable which is derived from this variable is the dummy variable App_dummy. It takes on the value of 1 if the customer registered through the iOS or the Android app and 0 otherwise. The total number of people who registered through an app is 1661. Table 2.6 describes the multivariate distribution of the App_dummy variable.

2.1.6 Time

In order to assess whether the time since a customer enrolled in the loyalty program has an impact on the choice of channel, the number of days from enrollment to the last day of customer information included in the dataset was used as an independent variable, referred to as time. Figure 2.2 is a histogram showing the distribution of the time variable among the 6076 customers that used at least one digital information channel.

![Histogram over time](image)
3. Method

Given the characteristics of the data it is necessary to find a model that has the following properties; the response variable is categorical and the explanatory variables are either metric or nonmetric. Another necessary purpose of the model is predictions of category classification of the dependent variable.

For classification modeling, two different statistical methods are commonly used, the discriminant function analysis and the multinomial regression analysis (Hair et al., 2014). The discriminant function analysis requires the independent variables to be normally distributed and for the covariances to be equal. It does also have problems with heteroskedasticity and non-linearity. If these assumptions were met discriminant analysis would be preferable to multinomial logistic regression due it being the more powerful of the two options (Starkweather and Moske, 2011). However, the data set in this study does indeed violate said assumptions making multinomial logistic regression the best approach for the purposes of this thesis. There are several categorical independent variables, and these are not normally distributed.

In order to better understand how multinomial logistic regression works it is necessary to account for how it works in the case where the independent variable is binary. Binary logistic regression is a widely used method because of its specific properties. Binary logistic regression does not require that any assumptions are fulfilled. Non-linear, heteroskedastic or non-normal data does not affect the usefulness of the analysis (Simonoff, 2014). However, problems with overdispersion is common, resulting in a larger variance than what is normally accounted for in the logistic regression model. Unless this is mitigated, the risk of type one errors will increase (Carruthers et al., 2008). Logistic regression may also require very large sample sizes, if there are categorical variables among the independent variables (Schwab, 2002).

3.1 Binary logistic regression

Binary logistic regression is a generalized linear model in which the dependent variable has a binary outcome, 1 or 0. Usually, the 1 and the 0 represents success and failure respectively (Dobson and Barnett 2008, pp 123) making it a Bernoulli trial that is distributed (B,π) where Pr(Z=1) = π and Pr(Z=0) = (1-π) (Dobson and Barnett 2008, pp 123).

Because of the characteristics of the binary variable, its predicted value (which is a probability) must fall within the range 0 to 1. To capture this relationship, the logistic regression uses a logistic curve. This logistic curve is a transformation of the dependent variable. In the logistic regression the slope of the logistic curve decreases as it approaches 0 and 1 respectively until it flattens out, making it impossible for the predicted value to reach or exceed 0 or 1. This gives it a shape that resembles the letter s (Hair et al., 2014, pp 317).

The logistic curve is derived from the natural logarithm of the odds of the dependent variable. These log-odds come from the odds of success, that is, the odds of the dependent variable taking on the value 1. This logit transformation is referred to as the link function and makes the dependent variable metric upon which the data of a given sample can be fitted. The metric values
(the log-odds, or sometimes just plain odds) can then be reverted to a predicted value or probability that lies within range of 0 to 1, thus not violating the aforementioned binary constraint.

The logit function in its general form is as follows:

\[
\text{Logit } \pi_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^T \mathbf{\beta}
\]  

(1)

Where \( \mathbf{\beta} \) is the vector for the parameters and \( \mathbf{x}_i \) is a vector for all continuous and dummy variables who in turn corresponds to covariates and factor levels respectively (Dobson and Barnett 2008, pp 131).

Given a certain cutoff-value, it is now possible to predict success or failure. A cut-off value of 0.5 will classify all observations with a probability of success exceeding 0.5 as 1 (success) and all observations with probability of success that is lower than 0.5 will be classified as 0 (failure).

3.2 Multinomial logistic regression

Multinomial logistic regression is an extension of the binary logistic regression. It allows for a dependent variable with more than two unordered outcomes (a category variable). Unlike binary logistic regression, multinomial logistic regression requires two assumptions to be fulfilled, namely independence among dependent variable choices and non-perfect separation. (Simonoff, 2014).

The log-odds are in this case calculated differently as they use a reference category rather than the possibility of an event not happening \((1 - \pi)\). The reference category is either arbitrarily chosen, or chosen for interpretation purposes, and the odds for a certain event happening is calculated through the probability of an event happening divided by the reference category. A mathematical representation for category \( j \) is as follows:

\[
\text{Logit } \pi_j = \log\left(\frac{\pi_j}{\pi_1}\right) = \mathbf{x}_j^T \mathbf{\beta}_j
\]  

(2)

where \( \mathbf{x}_j^T \) is the vector for continuous and/or dummy variables and corresponding factor levels and \( \mathbf{\beta}_j \) is the parameter vector (Dobson and Barnett 2008, 151).

Unlike the cutoff classification of the binary logistic regression the multinomial case uses a cut-off where the most probable value for a given category is chosen as a prediction.

3.3 Interpretation of the parameter coefficients

The parameter coefficients of any logistic model are interpreted as the change in log-odds given a one-unit change in \( X \), all else being held equal. The log-odds can easily be converted into regular odds making the intuitive interpretation easier. This interpretation however does only apply for the linear case. Non-linear variables can be included to explore whether non-linear relationships exist. When this is done, the interpretation differ because the rate of change depends on the value of that particular variable.
3.4 Assumptions
As mentioned earlier in the thesis, a great advantage of using multinomial logistic regression is its relative lack of assumptions that are necessary to fulfill. The only two assumptions, however, that are necessary to fulfill for the analysis to be valid are the independence of the choices of the dependent variable (in our case the choice of digital information channel) and the assumption of non-perfect separation. A perfect separation means that the independent variables without fail classifies the variable outcomes in prediction analysis (Starkweather and Moske 2011). When this happens, the maximum likelihood estimates cannot be computed and the model will have unrealistic coefficients with inflated estimated effects.

In this case, independence among information channels means that the choice of whether a customer uses a specific channel is independent of whether he or she uses another channel. Since nothing prohibits a customer from selecting more than one channel, this variable can be viewed as independent. Moreover, no perfect separation exists in the dataset used in this study.

3.5 Goodness-of-fit statistics
Two chi-square tests are commonly used to assess goodness-of-fit for multinomial logistic regression. The first is the Pearson chi-square test and the other is the deviance chi-square test. For detailed information on how these tests are computed, see Dobson and Barnett (2008).

Hosmer and Lemeshow (1980) proposes a goodness-of-fit test for binominal logistic regression, called the Hosmer-Lemeshow test. This test is a common way to assess goodness-of-fit for a logistic model (Fagerland and Hosmer, 2012) and lack of goodness-of-fit may indicate overdispersion (Carruthers et al., 2008). The test can be generalized to apply to multinomial logistic regression as well (Fagerland and Hosmer, 2008), and this generalized Hosmer-Lemeshow test is also used to assess overdispersion in this thesis. The Hosmer-Lemeshow test statistic is computed using the following formula

\[ C_g = \sum_{k=1}^{g} \sum_{j=0}^{c-1} \frac{(O_{kj} - E_{kj})^2}{E_{kj}} \]  

where \( C_g \) is the test statistic. \( j \) is the specific response category (c), ranging from 0 (the reference category) to c-1. \( O_{kj} \) is the observed frequency of a given response category in a given group. \( E_{kj} \) is the estimated frequency of a given response category in a given group. This test statistic is \( \chi^2 \) -distributed with \((g-2)*(c-1)\) degrees of freedom, under the null hypothesis that the sample is sufficiently large (larger than 400) and that the fitted model is correct (Fagerland and Hosmer, 2012).

3.6 Overdispersion
While multinomial logistic regression does not assume that overdispersion is not present, the presence of overdispersion will affect this method in the same way it will all generalized linear models. For example, type one errors will become more common (Carruthers et al., 2008). The most common way to assess overdispersion is to use the Pearson and the deviance chi-square tests and compare the chi-square value with its degrees of freedom. If the chi-square value is
greater than the number of degrees of freedom, overdispersion may be present. If this quotient is very large (larger than 5) the model is probably of little use (Carruthers et al., 2008).

One should be cautious not to rely too heavily on these quotient values though, since some researchers, like Bedrick (2009), point out that high values of deviance statistics could indicate a bad model fit rather than presence of overdispersion. Czado (2004) adds further caution when drawing conclusions about results from deviance statistics as she identifies a number of issues that can be a possible cause of a potential high value for said results. Those issues are missing covariates/interaction terms, not taking a non-linear relationship into account, using the wrong link-function and having large outliers in the data.

If overdispersion is present because of the failure to include important variables, rather than a misspecified error distribution, it cannot easily be removed from the model. However, it can be corrected for by multiplying the covariance matrix with the chi-square quotient used to assess its presence. This correction is viable as long as the chi-square quotient is not too large (Carruthers et al., 2008).

3.7 Multiple split evaluation

One purpose of this thesis is to create a selection model with strong predictive power in order to be able to accurately select which channel that can best be used to reach a specific customer. Evaluation of predictive power is notoriously difficult, however. Prediction models are supposed to capture trends within customer behavior, but perfect separation between the lasting trends and the sample noise is impossible. Thus, such models will never perform perfectly, since sample noise will affect the model estimation. This problem is called over fitting (Stein, 2014).

With this problem in mind, it is still preferable to somehow evaluate the predictive power of the models that will be created in this thesis, since they will primarily be used to this end. The method chosen to perform this evaluation is called multiple split evaluation. It is an extension of the single split method, in which the sample is randomly divided in two equally large groups. One used to estimate the model (called training set) and another to test how well the model performed (the test set). The advantage of making such a split is that the model does not have to be evaluated within the same sample in which it has been evaluated, thus the problem of over fitting can be somewhat mitigated (Malthouse, 2001).

One known problem of the single split approach, however, is that model evaluation variance between different splits is unknown. It has been showed that significant differences can be found between how good a model is evaluated to be between different random splits, even in large sample sizes (Malthouse, 2001). To counteract this problem, a multiple split approach has been used in this study. Instead of splitting the sample once, the sample is randomly split in two equal parts several times and the model performance is recorded each time. The average performance of all splits are then used for the evaluation. Using this method, the between-split variance of evaluation can also be examined.
4. Results
All estimations and outputs have been computed in SAS software version 9.

4.1 Model specification and selection
In order to approach the research question, different models have been estimated in order to detect a good model fit for the data. In the end, four models were chosen and estimated. These models were chosen because they give fairly good predictions and suit the purpose of the study. One of the initial models that was estimated was one including all independent variables available in the sample. However this model had too small sub groups in order for the model to be useful (more on this in Section 4.2.1)

All models have choice of channel as dependent variable. The reference channel is app, it was chosen as a reference because it is the most contrasted information channel. The category variable “Multiple channels” is hereon abbreviated as “Multi”.

The models themselves look as follows:

<table>
<thead>
<tr>
<th>Model #</th>
<th>Independent variable estimations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>$\beta_0 + \beta_1 Age + \beta_2 Time + \beta_3 Nationality + \beta_4 Sex$</td>
</tr>
<tr>
<td>Model 2</td>
<td>$\beta_0 + \beta_1 Age + \beta_2 Time + \beta_3 Nationality + \beta_4 Sex + \beta_5 Age^2$</td>
</tr>
<tr>
<td>Model 3</td>
<td>$\beta_0 + \beta_1 Age + \beta_2 Time + \beta_3 Nationality + \beta_4 App_dummy$</td>
</tr>
<tr>
<td>Model 4</td>
<td>$\beta_0 + \beta_1 Age + \beta_2 Time + \beta_3 Nationality + \beta_4 App_dummy + \beta_5 Age^2$</td>
</tr>
</tbody>
</table>

The difference between the models is the different use of independent variables. Models 3 and model 4 use the App\_dummy variable as an independent variable whereas models 1 and model 2 use Sex instead. Models 2 and model 4 are extensions of models 1 and model 3 respectively, where Age^2 has been added to account for possible non-linear relationships.

Since every single one of these models require a substantial amount of outputs to be thoroughly analyzed, we have decided to focus on the model we consider to be the best one. To determine which model this is, the predictive power of the models is first considered, since it is important for their business application. Second, statistical properties are taken into account.

4.1.1 Predictions
The purpose of the estimated models is to predict which channel is best suited to reach a specific customer. Thus, one of the most important characteristics to evaluate of each model is how well it performs in these predictions.

As explained in Section 3.7, a multiple split approach has been selected to evaluate the prediction performance of the estimated models. 30 splits have been run, and the average prediction performance of each model can be seen in figure 4.2.
Figure 4.1. Prediction performance of model 1-4.

Note that the x-axis scale in figure 4.1 ranges from 69% to 80%, rather than 1 to 100. This is because 69% of the sample uses web as their chosen information channel and to be useful, models must perform better than one that would always predict the web channel. 80 has been chosen as the upper bound to further zoom in on the range of correct prediction rates among the four models and make the between model difference easier to see.

It can be noted that among the 30 splits used to evaluate the models, model 4 showed the best performance 18 times and model 3 did so 10 times, model 3 and 4 tied twice. The difference between the model performances, \( \Delta \)performance\(_{M4-M3} \), defined as

\[
\text{(share correct guesses, model 4)} - \text{(share correct guesses, model 3)},
\]

had the average value of 0.18% with a standard deviation of 0.56%.

Thus, with regards to the predictive power of the models, model 4 is the best model, but the difference between model 3 and model 4 is very small.

4.1.2 Statistical properties of the models

Notice in table 4.1 that model 3 is nested in model 4. If \( \beta_5 \) in model 4 was restricted to 0, model 3 and model 4 would be identical. To differentiate whether a nested model is as good as the unrestricted mode, a likelihood test can be used (Dobson 1990, pp 57). This test can distinguish whether an unrestricted model is significantly better than an unrestricted one. The hypothesis of the test can be formulated as follows:
\( H_0: \) The restricted model is as good as the unrestricted model

\( H_1: \) The restricted model is not as good as the unrestricted model

The test statistic (\( T \)) looks as follows:

\[
T = 2 \times \left( \log \text{likelihood of restricted model} - \log \text{likelihood of unrestricted model} \right) \tag{5}
\]

And is chi-square distributed with a number of degrees of freedom equal to the restrictions imposed on the restricted model. In this case, this is three, since three channels are compared to the reference channel (see e.g. table 4.4 below). The log likelihood of model 3 (the restricted model) is -2722.27 and the log likelihood of model (the unrestricted model) is -2963.71. Thus,

\[
T = 2 \times \left( (-2722.27) - (-2963.71) \right) = 482.88 \tag{6}
\]

And

\[
\Pr(T > 482.82) = \Pr\left( \chi^2_{(3)} > 482.88 \right) \approx 0.01 \tag{7}
\]

Given this result, the null hypothesis stating that the unrestricted model is not better than the restricted model is rejected. Thus, model 4 is selected as the best model and the one to be analyzed in detail.

4.2 Test of assumptions and sample considerations
4.2.1 Sample size considerations and overdispersion

According to Schwab (2002) a guideline of a minimum of 10 observations per sub-group is necessary, where a sub-group is a group created by the inclusion of category variables (e.g. Danish women). This guideline is not violated by any of the four models. However, if all available independent variables were to be included a problem of tiny sub-groups would arise. In our case, this means one has to choose between using the variable Sex or App dummy. To some, this feels counter intuitive given the relatively large sample size. However, even with big sample sizes, the sub groups created from the use of categorical variables can become very small very rapidly as categorical variables are added to the model.

Two different chi-square tests of residual deviance and the generalized Hosmer-Lemeshow test statistic have been used to assess overdispersion. The first chi-square test is the Pearson chi-square, which returned a chi-square value of 24422 with 16522 degrees of freedom, with a quotient of 1.48. The other is the deviance chi-square, which returned a chi-square value of 7739 with 16522 degrees of freedom, this a quotient of 0.47. When calculating the generalized Hosmer-Lemeshow test statistic, a squared difference is divided by the estimated frequency of each category (see Section 3.6). None of the specified models (model 1 to model 4) estimated that any observation belonged to the response category of “multiple channels”. Thus, the Hosmer-Lemeshow test statistic included a division by zero, returned an undefined value and could not be used to indicate a good model fit.
All in all, one of the two chi-square tests indicated overdispersion, the other does not and the generalized Hosmer-Lemeshow test did not indicate a good model fit. According to Carruthers et al. (2008), there are no adverse effects of accounting for overdispersion, and as such we have elected to account for it by multiplying the covariance matrix with the Pearson chi-square quotient (SAS Institute Inc., 2010).

4.3 Interpretations of the multinomial logistic model
A 5 % significance level is chosen for the forthcoming analysis.

4.3.1 Test of the global null

<table>
<thead>
<tr>
<th>Table 4.3. Test of global null hypothesis: Beta=0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test</strong></td>
</tr>
<tr>
<td>Likelihood ratio</td>
</tr>
<tr>
<td>Score</td>
</tr>
<tr>
<td>Wald</td>
</tr>
</tbody>
</table>

In table 4.1 the output of three different tests of a global null is presented. The null hypothesis of the mentioned tests is formulated as:

$H_0$: *No independent variables have non-zero coefficients*

$H_1$: *At least one independent variable has a non-zero coefficient*

It is interesting to test this since not rejecting the global null hypothesis indicates a model which serves no purpose in explaining the relationship between the dependent and independent variables. As can be seen in table 4.1 all three different hypothesis tests reject the global null on a 5% significance level, thus making the dependent variables potentially relevant, which warrants further analysis.

The degrees of freedom are based upon the number of parameters included in the model. Because all the hypothesis tests test the same null hypothesis, the degrees of freedom are also equal. The probabilities are clearly low. The *Chi-square* column itself just tells the value of the test statistic for each test.

4.3.2 Maximum likelihood estimates

In table 4.4 the maximum likelihood estimates are shown. The coefficient estimates that are not statistically significant on the chosen 5 % significance level are marked in grey. The null hypothesis these p-values stem from is as follows:

$H_0$: *The parameter coefficient is zero, given the inclusion of all other independent variables in the model*

$H_1$: *The parameter coefficient is non-zero, given the inclusion of all other independent variables in the model*
The test statistic of that particular null hypothesis follows the chi-square distribution and its corresponding p-values for each coefficient estimate can be found in column $Pr>Chisq$.

Table 4.4. Analysis of maximum likelihood estimates

<table>
<thead>
<tr>
<th>Channel</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
<th>Wald Chi-sq</th>
<th>Pr &gt; Chisq</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mail</td>
<td></td>
<td>2.345</td>
<td>0.666</td>
<td>12.409</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Web</td>
<td></td>
<td>8.892</td>
<td>0.592</td>
<td>225.937</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Multiple</td>
<td></td>
<td>2.673</td>
<td>0.739</td>
<td>13.066</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Mail</td>
<td>Time</td>
<td>-0.004</td>
<td>&lt;.001</td>
<td>42.170</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-0.015</td>
<td>0.033</td>
<td>0.209</td>
<td>0.648</td>
</tr>
<tr>
<td></td>
<td>Age²</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>1.665</td>
<td>0.029</td>
</tr>
<tr>
<td>Nationality (DK)</td>
<td></td>
<td>1.337</td>
<td>0.176</td>
<td>58.011</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Nationality (NO)</td>
<td></td>
<td>-0.058</td>
<td>0.204</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>App_dummy</td>
<td>-2.850</td>
<td>0.184</td>
<td>241.365</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Web</td>
<td>Time</td>
<td>-0.009</td>
<td>0.001</td>
<td>228.082</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-0.161</td>
<td>0.029</td>
<td>30.128</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Age²</td>
<td>0.002</td>
<td>&lt;.001</td>
<td>28.220</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Nationality (DK)</td>
<td></td>
<td>-0.091</td>
<td>0.160</td>
<td>0.324</td>
<td>0.570</td>
</tr>
<tr>
<td></td>
<td>Nationality (NO)</td>
<td></td>
<td>0.612</td>
<td>0.161</td>
<td>14.517</td>
</tr>
<tr>
<td></td>
<td>App_dummy</td>
<td>-3.369</td>
<td>0.162</td>
<td>431.233</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Multiple</td>
<td>Time</td>
<td>-0.003</td>
<td>&lt;.001</td>
<td>18.324</td>
<td>&lt;.001</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>-0.082</td>
<td>0.037</td>
<td>4.793</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>Age²</td>
<td>0.001</td>
<td>&lt;.001</td>
<td>4.493</td>
<td>0.034</td>
</tr>
<tr>
<td>Nationality (DK)</td>
<td></td>
<td>0.373</td>
<td>0.213</td>
<td>3.080</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>Nationality (NO)</td>
<td></td>
<td>0.663</td>
<td>0.209</td>
<td>9.7289</td>
</tr>
<tr>
<td></td>
<td>App_dummy</td>
<td>-1.536</td>
<td>0.207</td>
<td>54.843</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Since multinomial logistic regression is constructed in a way where a category outcome is evaluated relative to a given reference category, it is made up of multiple partial models. Since the reference category is "App" and the number of category outcomes are four, the partial models are as follows:
1. **Partial model (Mail):** Mail relative to App
2. **Partial model (Web):** Web relative to App
3. **Partial model (Multi):** Multi relative to App

4.3.3 Interpretations of maximum likelihood estimates
The multinomial logit estimates seen in the estimate column in table 4.4 are interpreted as the change in log-odds given a one unit increase in its respective variable value (or a category variable taking on a certain outcome, e.g. nationality). In turn, a change in log-odds means the change in the log-odds of an observation preferring one channel over app. The nationality reference is Swedish, which means that the nationality coefficients (NO and DK) in each partial model should be interpreted in relation to the nationality Swedish.

**Partial model 1 - Coefficient interpretations**
In partial model 1, holding all other constant, the multinomial logit estimate for a one unit increase in time is -0.004. A negative sign before the coefficient estimate in this case indicates that loyalty program members have a small tendency to prefer app over mail as time goes by. If the observation is Danish, it is on average 1.337, meaning that Danes are more likely to use mail compared to Swedes. If the observation is Norwegian, it is on average -0.058, meaning that Norwegians are less likely to use mail than Swedes. However, this result is not statistically significant. If the source of registration is app, it is on average -2.850, indicating that if the source of registration is app then app is preferred to mail.

Age and age\(^2\) are not interpreted the same way as above since the relationship that is estimated is non-linear. Thus, the change in log-odds in dependent upon the value of age. For a given value of age, this is calculated as follows:

\[
[\alpha*(age+1) + \beta*(age+1)^2] - (\alpha*age + \beta*age^2)
\]

which simplifies to:

\[
\alpha + \beta*(1+2*age),
\]

where \(\alpha\) and \(\beta\) are coefficients. For example, a value of age=21 that increases with one year, would lead to a change in log-odds that is as follows:

\[-0.015 + 0.000*(1+2*21) \approx -0.015,\]

which means that when age changes from 21 to 22, keeping all other variables constant, it is more likely to use app comparing to mail. Serving as another example, when age is 44 and is increased by one the corresponding change in log-odds is:

\[-0.015 + 0.000*89 \approx -0.015\]

meaning that for partial model 1, the effect of age is almost linear, since \(\beta \approx 0\).
Partial model 2 – Coefficient interpretations

In partial model 2, holding all other constant, the multinomial logit estimate for a one unit increase in time is -0.009. The negative sign before the coefficient estimate indicates that loyalty program members have a small tendency to prefer app over web as time goes by. If the observation is Danish, it is on average -0.091, meaning that Danes are likely to use web compared to Swedes, also true for this result is that it is not statistically significant. If the observation is Norwegian, it is on average 0.612 indicative of Norwegians being more likely to use web than Swedes are. If the source of registration is app, it is on average -3.369, indicating that if the source of registration is app then the app is preferred to using web.

Like in partial model 1, \( \text{age} \) and \( \text{age}^2 \) are not interpreted the same way as above since the relationship that is estimated is non-linear. Thus, the change in log-odds in dependent upon the value of age.

For example, a value of \( \text{age}=21 \) that increases with one year, would lead to a change in log-odds that is as follows:

\[
-0.161+0.002(1+2*21) = -0.075, \quad (12)
\]

thus when age changes from 21 to 22, keeping all else constant, it is more likely to use app comparing to mail. Serving as another example, when age is 44 and is increased by one the corresponding change in log-odds is:

\[
-0.161+0.002*89 = 0.017, \quad (13)
\]

meaning that a change in age from 44 to 45, keeping all else constant, it is less likely to use app comparing to mail.

Partial model 3 – Coefficient interpretations

In partial model 3, holding all other constant, the multinomial logit estimate for a one unit increase in time is -0.003, indicative of loyalty program members preferring the app to using multiple models as the time goes by. If the observation is Danish, it is on average 0.373, meaning that Danes are more likely to use multiple channels to a greater extent than Swedes, this result is, like the previous one, not statistically significant. If the observation is Norwegian it is on average 0.663, meaning that Norwegians are also more likely to use multiple channels than Swedes. This results contradicts a previous result presented for partial model 1 but since that result is not statistically significant it will be discarded from further analysis, along with all other non-significant estimates. If the source of registration is app, it is on average -1.536, indicating that if the source of registration is app then the sole use of app is preferred to using multiple channels.

Like in partial model 1, \( \text{age} \) and \( \text{age}^2 \) are not interpreted the same way as above since the relationship that is estimated is non-linear. Thus, the change in log-odds in dependent upon the value of \( \text{age} \).
For example, a value of \( age = 21 \) that increases with one year, would lead to a change in log-odds that is as follows:

\[-0.082 + 0.001 \times (1 + 2 \times 21) = -0.039 \quad (14)\]

Serving as another example, when \( age \) is 44 and is increased by one the corresponding change in log-odds is:

\[-0.082 + 0.001 \times 89 = 0.007 \quad (15)\]

These two examples show that when \( age \) changes from 21 to 22 it is more likely to solely use \textit{app} comparing to multiple channels. When \( age \) changes from 44 to 45 it less likely to solely use the \textit{app} comparing to using multiple channels.

4.3.4 Odds ratios

As previously mentioned, the interpretations of log-odds can be somewhat non-intuitive. Therefore, adding odds ratios to the analysis can be useful.

<table>
<thead>
<tr>
<th>Channel</th>
<th>Covariate</th>
<th>Point estimate</th>
<th>95% confidence intervals (Wald)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mail</td>
<td>Time</td>
<td>0.996</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>0.985</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>( Age^2 )</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Nationality (DK)</td>
<td>3.807</td>
<td>2.699</td>
</tr>
<tr>
<td></td>
<td>Nationality (NO)</td>
<td>0.943</td>
<td>0.633</td>
</tr>
<tr>
<td></td>
<td>App_dummy</td>
<td>0.058</td>
<td>0.040</td>
</tr>
<tr>
<td>Web</td>
<td>Time</td>
<td>0.991</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>0.851</td>
<td>0.803</td>
</tr>
<tr>
<td></td>
<td>( Age^2 )</td>
<td>1.002</td>
<td>1.001</td>
</tr>
<tr>
<td></td>
<td>Nationality (DK)</td>
<td>0.913</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>Nationality (NO)</td>
<td>1.843</td>
<td>1.346</td>
</tr>
<tr>
<td></td>
<td>App_dummy</td>
<td>0.034</td>
<td>0.025</td>
</tr>
<tr>
<td>Multiple</td>
<td>Time</td>
<td>0.997</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>Age</td>
<td>0.922</td>
<td>0.857</td>
</tr>
<tr>
<td></td>
<td>( Age^2 )</td>
<td>1.001</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Nationality (DK)</td>
<td>1.452</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td>Nationality (NO)</td>
<td>1.921</td>
<td>1.274</td>
</tr>
<tr>
<td></td>
<td>App_dummy</td>
<td>0.215</td>
<td>0.143</td>
</tr>
</tbody>
</table>
In table 4.5, like in table 4.4, the point estimates with non-significant p-values are marked in grey. The odds ratios can be for each partial model separately interpreted as follows:

**Partial model 1 – Odds ratio interpretations**

In partial model 1, a point estimate of 0.996 in row 2 indicates that a one unit increase in time translates to a 0.4 % decrease in the odds of an observation preferring *mail* relative to *app*, all other held constant. If the observation is Danish, the odds change is 280.7%. If the observation is Norwegian, the odds change is -5.7%. If the source of registration is *app*, then the odds change is on average -94.2 %. The negative sign before the odds percentage indicates in this case, that the odds of the loyalty program member preferring *mail* to *app* decreases with 94.2% if the member in fact registered on the *app*. Same interpretational logic holds for the nationality variable with a negative sign.

Like with the maximum likelihood estimates in Section 4.3.4, the odds ratios of *age* and *age*^2^ require a different approach than previous ones did since they estimate something non-linear.

The change in odds ratios will calculated using the following quotient, where \( O_0 \) is the odds for *age* and \( O_1 \) is for *age* + 1 :

\[
\frac{(O_1 - O_0)}{O_0} \quad (16)
\]

where \( (O_1-O_0) \) is:

\[
\exp(\alpha \cdot age + \beta \cdot age^2) \cdot \exp(\alpha + \beta (2 \cdot age + 1)) - 1 \quad (17)
\]

Simplifying this quotient yields in:

\[
\frac{(O_1 - O_0)}{O_0} = (\exp(\alpha + \beta (2 \cdot age + 1)) - 1) \quad (18)
\]

As an example, a value of *age*=21 that increases with one year, would lead to an odds ratio change that is as follows:

\[
\frac{(O_1 - O_0)}{O_0} = \exp[-0.015 + 0.000 \cdot (2 \cdot 21 + 1)] - 1 = (0.985 - 1) = -1.5 \% \quad (19)
\]

Another example for when *age* = 44:

\[
\frac{(O_1 - O_0)}{O_0} = \exp[-0.015 + 0.000 \cdot (2 \cdot 21 + 1)] - 1 = (0.985 - 1) = -1.5 \% \quad (20)
\]

**Partial model 2 – Odds ratio interpretations**

In partial model 2, a point estimate of 0.991 in row 8 indicates that a one unit increase in time translates to a 0.9 % decrease in the odds of an observation preferring multi relative to app, all other held constant. If the observation is Danish, the odds change is -8.7%. If the observation is
Norwegian, the odds change is 84.3%. If the source of registration is app, then the odds change is on average -96.6 %.

Like in partial model 1, the odds ratios of age and age² require a different approach than previous ones did since they estimate something non-linear. Recalling equation 18 from partial model 1, an example can be calculated. A value of age =21 that increases with one year, would lead to an odds ratio change that is as follows:

\[
\frac{(O_1 - O_0)}{O_0} = \exp[-0.161 + 0.002 \times (2 \times 21 + 1)] - 1 = (0.928 - 1) = -7.2 \%
\] (21)

Another example for when age = 44:

\[
\frac{(O_1 - O_0)}{O_0} = \exp[-0.161 + 0.002 \times (2 \times 21 + 1)] - 1 = (1.017 - 1) = 1.7 \%
\] (22)

**Partial model 3 – Odds ratio interpretations**

In partial model 3, a point estimate of 0.997 in row 14 indicates that a one unit increase in time translates to a 0.3 % decrease in the odds of an observation preferring web relative to app, all other held constant. If the observation is Danish, the odds change is 45.2 %. If the observation is Norwegian, the odds change is on average 92.1 %. If the source of registration is app, then the odds change is on average -78.5 %.

Like in partial model 1, the odds ratios of age and age² require a different approach than previous ones did since they estimate something non-linear. Recalling equation 18 from partial model 1, an example can be calculated. A value of age=21 that increases with one year, would lead to an odds ratio change that is as follows:

\[
\frac{(O_1 - O_0)}{O_0} = \exp [-0.082 + 0.001 \times (2 \times 21 + 1)] - 1 = (0.962 - 1) = -3.8 \%
\] (24)

Another example for when age = 44:

\[
\frac{(O_1 - O_0)}{O_0} = \exp [-0.082 + 0.001 \times (2 \times 21 + 1)] - 1 = 1.007 = 0.7 \%
\] (25)

**4.4 Alternative rule for channel selection**

When the predictive power of the model was assessed (see Section 4.1.1), only one way to use the selection model for predictions has been discussed, to select the highest estimated probability for each customer and use the corresponding channel. However, a number of other selection rules can be considered, depending on the situation that the selection model is used in.

In this section, one other selection rule is examined. Instead of choosing the single most probable channel, the two most probable channels are selected. Since more channels are being used under such a rule, the number of correct guesses will increase. However, the number of selected
channels is effectively doubled, so whether this selection rule is preferable depends on the marginal costs of using more channels, a consideration outside the scope of this thesis.

To evaluate this selection rule, multiple split evaluation is again used. The same 30 splits of the dataset has been used as the ones used to evaluate the selection rule evaluated in 4.1.1. The average results and standard deviation of these 30 splits can be seen in figure 4.2.

![Figure 4.2. Prediction performance of model 1-4 under an alternative selection rule.](image)

Note that, as with figure 4.1, the scale is altered to zoom in and highlight the differences between the models. Among the 30 splits, model 4 performed best 15 times, model 3 did so ten times and there was a tie between model 3 and 4 five times.
5. Discussion

5.1 Applications of the model

Since the model(s) are better than at predicting a given customer’s choice of digital information channel rather than just classifying the customer by the most common channel, it can be argued that this model is a useful addition to the airline company from which the data stems from. The model can indeed be used to reduce costs for the particular airline company by adding knowledge to the channel of choice for a given customer, and consequently targeting the customer through its predicted preferable digital information channel. From the model, point estimates can be made, which relatively accurately predicts which channel that is the actual preferred one. Decision rules such as the one discussed in Section 4.4.3 can also be formulated based on the model in order to customize how it is used for other contexts than the one presented here.

Furthermore, the model can be used to draw some more general conclusions about the loyalty program members of the airline company, about which demographic members group prefers which channel. The model also gives some insight in how source of and days since registration works as predictors of channel preference.

Judging from the estimated coefficients of the model, older members of the loyalty program are more likely to use mail rather than app, compared to younger members. The estimated coefficient of the nationality Danish indicates that Danish people are more likely to use mail than Swedes and that Norwegians are more likely to use web than Swedes.

The odds ratios that are statistically significant shows that for either partial model the odds of choosing the mail, web or multiple channels decreases dramatically if the member has registered on the airline company’s mobile application. Given these odds ratios it can be argued that the airline company should strive to reach members who have registered via said application through the application itself. The odds ratios for time in each partial model is significant and indicates that preference of the mobile application becomes more probable as more days since registration elapses.

5.2 Potential limitations of the model

There are some possible limitations to the model. Since it uses data from existing loyalty program members, its usage when it comes to future members is not as certain. One can argue that members a few years into the future will not be engaged in the same way as members of this dataset is being now. Since forms of digital output rapidly changes one or more of the channels used in this thesis could be considered decreasingly important for customer engagement as time passes by. Also, a significant improvement of any of the present digital channels used by the airline company could distort the current paradigm and become increasingly popular relative to the others. Another thing to consider is the relative lack of independent variables. More independent variables, preferable variables that would allow for greater distinction between the observations, would make the predictions better and the usefulness of the model greatly increased.
Lastly, the introduction of a new digital information channel could make this model irrelevant, as it does not take this new channel into account.

Although the target population for the inferential analysis in this paper is the members of the loyalty program, observations where the members actually interacted were used as basis for the inference. This can sound faulty by logic, since the target population in this case should be the customers who actually engage in any of the channels, rather than the entire population of members. The choice of target population in this motivated with the relatively small time span that the members are able to interact with any of the channels in mind. There is no clear pattern in the data that would suggest that the customers who actually engage in any of the channel during the approximately three month-long time window act in a way that would separate them from the target population of this thesis. Therefore it is assumed that the customers who engage in the time window of the case study is part of the pool of customers that make up the members list, and the inference is based upon this assumption.

5.3 Suggestions for further research

Further researches can be conducted in many different ways. For example, a new effort to estimate the preferred channel for a given customer engaging in digital information channel with a larger dataset could enhance the analysis presented in this thesis. A larger dataset would perhaps make it possible to make use of more categorical variables since the subgroups would become larger. If new kind of data would be collected by the airline company, more explanatory variables could be created which could also improve the model.

Furthermore, since the data is extracted from a three month period, potential season variation is not captured in this model. Data from a year-long period would account for this possibility and further validate the results.

There is another reason to redo the analysis with data spanning over a longer period of time. If the time window, during which the interactions were registered, was longer more observations could be used to estimate the model.

In this thesis, we have chosen to use multinomial logistic regression. Other methods have also been tried in the direct marketing context and may very well be suitable for this kind of model as well. For example, semi-parametric logistic regression (Li and Racine, 2007) is implemented in a direct mail setting in Bult (1993). Neural networking approaches (Abdi et al., 1999) are implemented in the same setting in Baesens et al. (2002). Furthermore, in order to incorporate the longitudinal nature of the time variable, a longitudinal multinomial logistic regression would a good alternative to consider, for further reading in this subject, see (SAS Institute Inc., 2004).
6. Final conclusions

This paper is based on a case study of a Nordic airline company. The purpose is to estimate a model that can be used by said company for cost reduction by sending information via the channels that are most probable that a given customer to use. By using multinomial logistic regression, a satisfactory model has been created. This indicates that this method might create satisfactory results in other settings than this particular airline company. Thus, while the results found and presented in this thesis may not be very generalizable, the method proposed here may prove useful for practitioners in other situations as well. Patterns regarding which demographic groups prefer which digital information channels have been found, but conclusions based on these patterns must be restricted to the members of the loyalty program of the studied airline company.

Given the purpose and the research questions of this thesis and the presented results it can be concluded that:

A. The model(s) predicts, to a satisfactory extent, the information channel of choice, given data provided for a given customer.

B. Different demographic groups show tendencies towards preferring some channels to others.

We believe that the conclusions from research questions A-C, as well as the models themselves, will serve as a useful addition to the airline company’s collective pool of knowledge about their customer engagement and towards cost minimization.
References


Informant 1, employee at the airline company. 2015. Personal communication, 13 April.


The data analysis for this paper was generated using SAS software, Version 9 of the SAS System for Windows. Copyright © 2012 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.