Vrije Universiteit Amsterdam

# Direct Marketing Optimization

## Optimizing the email marketing strategy of an airline using data modeling: a literature study

June 30, 2018

# Direct Marketing Optimization

**Research Paper Business Analytics**

Vrije Universiteit
Faculty of Science
Business Analytics
De Boelelaan 1081A
1081 HV Amsterdam

**Author:**
Mathijs Koopman
m.j.koopman@student.vu.nl

**Supervisor:**
Bernard Zweers MSc

# Preface

This research is conducted as a part of the curriculum of the Master Business Analytics. For this course are 6 ECTS awarded, meaning that the expected workload will be four weeks of full time research. The purpose is to develop research, writing, as well as presentation skills.

In August 2017, I started as a dual student at Air France - KLM Benelux as a Customer Intelligence Analyst in the department Direct Sales and Digital Marketing. Therefore, I wanted to conduct a research study that will be relevant for the organization. Due to the amount of time, it would not have been possible to optimize the email marketing effort on real data, because there would be a lot of time invested in the data cleaning and preparation part. That is why a literature study is conducted, which will be relevant for the organization and forms a basis for further research. I would like to thank Bernard Zweers for his advice and support.

Mathijs Koopman

# Abstract

Email marketing is widely used to communicate to (possible) customers. However, no research has been conducted about the optimization of the marketing effort for an airline. In this paper the goal is to find data mining techniques that could be used for segmenting customers and optimizing the response. This is done based on research conducted in other lines of business. K-means clustering and decision trees seem to fit a general segmentation the best. For response modeling, next to decision trees, also artificial neural networks show promising results. Most of the studies only consider one email moment instead of the long-term relationship with the customer. This makes it more difficult to conclude which models should be used in order to optimize the effort.

# Contents

# 1 Introduction

Direct marketing is used widely by companies to communicate to possible purchasers in various forms of media. According to the Direct Marketing Association, marketeers in the US spent \$153.3 billion in 2010, which led to \$1.798 trillion incremental sales. Next to this, the efforts account for approximately 9.8 million jobs. Compared to mass marketing (television, radio and newspapers), the results can be measured directly, making it for companies possible to determine the success of a campaign. According to Ling and Li (1998), the response rate of mass marketing, so people actually buying a product after seeing a promotion, is often very low, around the 1%. Using data mining techniques, the direct marketing effort can be optimized, leading to a higher pay-off of the marketing budget.

Email marketing is a form of marketing where the customer has explicitly indicated that he or she wants to receive communication from the company. However, direct marketing also includes text messages (e.g. Whatsapp), calling customers, the companies own website, banners (advertisements on other websites) and advertising on social media (e.g. Facebook, Instagram and YouTube).

For this research, the main focus will be on email marketing, since including all the forms of direct marketing will make the paper to broad. Besides, the industry of interest will be an airline. There has been done research about the optimization of email marketing within other lines of business, but not specifically for the airline industry. Therefore the paper will be a literature study in which the methods and techniques used in other businesses, will be compared with an airline. The main focus will be on why the techniques work good for other businesses and if this effect is expected to be the same for the airline industry.

In this research there will be dealt with the following subquestions:

1. Which data modeling techniques can be applied the best for customer segmentation for targeted marketing?
2. Which data modeling techniques can be applied the best to improve email marketing (response modeling)?
3. What are good indicators (features) for predicting the response?
4. How can timing, frequency and content influence the response?

# 2 Comparison Different Lines of Business

Since this research will study the techniques used in other lines of business and conclude what will work for the airline industry, there will be started with comparing the different businesses found in the literature.

## Airlines

The airline industry can be divided into two type of customers: business and leisure travelers. Typical characteristics of business travelers are that they book relatively short in advance and have a short trip duration. These travelers have to be at a destination at a certain time. Next to this there is a large group of leisure travelers, who fly less frequently and are more price-conscious. Flight tickets are not cheap and for leisure customers most times bought for a holiday, so there are no impulse purchases. Most customers first orient themselves what will be the next travel period and destination and then search for tickets or go to a tour operator.

It is not so easy to distinguish different people, because not so much information needs to be filled in when buying a flight ticket. Most of the times analyses or targeting will be done based on a frequent flyer number, but many customers do not have or use this. Distinguishing people based on name (combination of last and first name) is risky, because different people may be mapped to the same person. Next to this, the name is sometimes misspelled or only the first letter of the first name is filled in. Email addresses cannot be used either, because people may have multiple email addresses and when the booking is made via a tour operator this information is not known.

## Financial Institutions and Loyalty Companies

Financial institutions and loyalty companies differ based on the fact that they have more information available about the customer. Especially financial institutions such as banks have information about the income of the customer as well as their spendings. This is useful information which can have a huge impact on the predictive power of a predicting model. Next to this, loyalty companies such as Air Miles also know what customers are spending at different companies. These customers can be easily identified when they use their credit/debit card (banks) or loyalty card (loyalty companies). The products they sell may be comparable with flight tickets, in the sense that loans or an insurance will not be taken out overnight.

## Catalog Companies

Especially case studies for direct marketing are conducted on datasets of catalog companies. Most of these companies offer a wide range of different products, increasing the chance that a customer may find more relevant products. A customer buys utensil instead of a service. Prices of many products are lower compared than flight tickets and therefore these kind of companies can stimulate impulse purchases. Next to this, they use basket cluster analyses to recommend other products in order to boost sales. On the other hand, the sector can be comparable in the sense that there will also be customers who do not buy frequently, which may be price shoppers. Most of these companies distinguish customers based on an account (email-address including a password). Most customers will buy products for themselves, so this will be useful information for possible offers in the marketing strategy.

## Non-Profit Organizations

Next to the catalog companies, there are also case studies conducted on non-profit organizations (mainly charities). These organizations especially make an appeal to the emotions of the donor. They show what they do and achieve with the donations and then ask for more money. Most of the times this are small amounts of money and these spendings cannot be compared to spendings at a company. Just like airlines, these organizations do not have much information about who the customer is.

## Hospitality Companies

Hotel chains or car rental companies can be comparable lines of business, due to the fact that they both offer services for which delayed consumption is not possible. However, also for these sectors no research is conducted on direct marketing optimization. Therefore no useful information can be gathered from these sectors.

# 3 Segmentation for Targeted Marketing

Before 1960, most of the companies used mass marketing in order to approach possible customers. However, with the current technology, such as Customer Relationship Management (CRM), a market can be broken down into different segments. Then within a segment, there are different people, but all with the same characteristics and/or needs. For the marketing of products, a certain group of interest can be targeted, saving marketing costs compared to targeting everyone. Whereas nowadays companies are focusing on personalization of emails, because every person (within a cluster) is different and has different needs, segmentation of the market still is important in order to select people which will be interested in a certain offer in first place.

In this section different methods to segment the market will be discussed. The advantages and possible disadvantages will be exposed and the section ends with an argumentation which method will fit an airline the best.

## 3.1 Segmentation Models

Customers can be segmented based on different bases:

- Demographic: quantifiable population features, such as age, gender and education.
- Psychographic: quantifiable features, such as number of consumptions and purchases behavior.
- Geographic: location features, such as province and postcode.
- Contextual: digital features, such as website visits and email opens.

In this subsection, different models for segmentation of the market will be explained.

### 3.1.1 RFM

First mentioned by Shepherd (1990), RFM is a method to make segmentations for analyzing the customers value. A RFM model consists of three dimensions:

- Recency: how recent was the last purchase of the customer?
- Frequency: how frequent does the customer purchase?
- Monetary Value: how much does the customer spend?

This model is used frequently in many sectors, due to the easiness to make different segments. Next to this, only transactional data is needed to set up a segmentation. Every dimension can be split into a self-determined amount of categories. If three different categories are chosen, this already leads to 27 (3x3x3) different segments. Furthermore, it is possible to determine the borders by business ruling, for example splitting the Recency into 0-3, 4-12 and 13-24 months. However, splitting the borders based on equally divided categories can also be an option. Sometimes this model is extended to a LRFM model, where the L represents the Length: since when is the customer active at the company. In this way the distinction between new and existing customers can be made.

In the study of McCarty and Hastak (2007), three different widely used methods in the direct marketing field are compared. Next to the RFM model, also decision trees and logistic regression are used as segmentation methods (limited to only transactional information). The evaluation metric used in the research is the gain percentage. This percentage corresponds to the percentage of actual responders when a subset of the whole population is contacted. If 10% of the people is targeted, the percentage of actual responders in that set minus the 10% of the actual responders from the total dataset is the gain percentage. Two types of datasets are researched, a dataset from a catalog business with a low response rate and a non-profit organization with a high response rate. For the dataset with a low response rate, the RFM model seems to work significantly better than logistic regression. There is no reason given for this, but it may depend on the dataset. However, for the dataset with a high response rate, none of the three models outperforms another model. The RFM model entirely focuses on the past purchases of a customer, which is a very sales driven approach. According to Zahay et al. (2004), this may yield sales in the short run, but might be bad for the long term relationship with customers.

The study of Coussement et al. (2014) is a follow-up study of the study by McCarty and Hastak (2007) and focuses on how segmentation performance is impacted by the accuracy of the data. They conclude that RFM is not the best option when there is no data inaccuracy. Furthermore, when the accuracy of the data decreases, RFM has a tendency to be more sensitive than the other models and performs worse.

Jonker et al. (2004), propose also a RFM model for segmentation. They started by a random initial segmentation and the optimal actions (how many mails does a customer receive in one year to achieve the highest possible revenue) for each segment is determined. Afterwards local search is used to determine new segmentations. Again the profitability of this cluster will be calculated and the iterative process will go on until the optimum is found. This model leads to a significant improvement compared to decision trees.

There can be concluded that the amount of available data used (only transactional data) is brief and therefore will not make the best usable segmentation. Furthermore, the discretization into categories causes a loss of information as well. On the other hand, it is easy to implement and interpret, and the model can be extended using other models.

### 3.1.2 K-Means Clustering

The K-Means Clustering algorithm is compared to the RFM a more mathematical approach. The idea was invented by Steinhaus (1957). The K indicates the number of clusters (or segments) that will be found by the algorithm. Given a set of data observations N (the amount of customers), the algorithm tries to separate the N observations into the K clusters. The mathematical formulation (MacQueen, 1967) is given by:

$$MIN \sum_{j=1}^{K} \int_{S_i} |z - u_j|^2 dp(z) \tag{1}$$

Where $u_j$ is the geometric centroid of the data points in $S_i$. In general the algorithm minimizes the total sum of the squared error between the data observations and the closest cluster centroid.

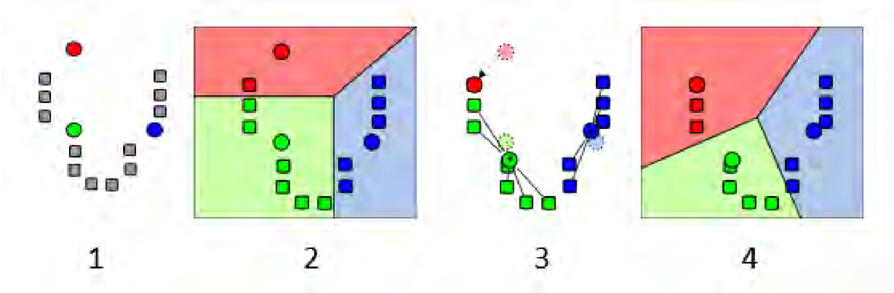

*Figure 1:* Simple illustration of the 3-means clustering algorithm[1]

In Figure 1, a simple illustration in a two dimensional plane shows how the algorithm basically works. In this case K=3. The steps are as follows:

1. Three initial means (red, green and blue) are randomly generated within the data space.
2. Then the clusters are generated (so the data points are mapped to the closest mean).
3. The centroid of each of the three clusters will become the new mean.
4. Steps two and three are repeated until convergence is reached.

For a more detailed explanation, please see Vattani (2011). All numerical features (F) can be included. In order to let every feature have an equivalent influence, the features should all be normalized. The data observations will be plotted in a F-dimensional space. Categorical data cannot be implemented in the algorithm, but there are certain alternatives as a K-Prototype algorithm, that can handle categorical data as well (Huang, 1998). By setting a dissimilarity measure, the influence of categorical compared to numerical can be set. The optimal K can be determined using the Elbow method. Two variables are plotted against each other, the percentage of the variance explained and the number of clusters. At a certain point, adding another cluster will not gain so

[1] https://en.wikipedia.org/wiki/K-means_clustering

much, but this can be challenging to see in a figure (KetchenJr & Shook, 1996). Therefore the second derivative can be used as well and the amount of clusters should be chosen where this is a strong minimum.

Liao et al. (2011), applied the K-means algorithm on the cosmetics market (comparable with a catalog company). The goal is to distinguish the purchase behavior between different customers. It states that it is important to have to most actual data on customers available, since peoples lifestyle and purchase behavior are changing rapidly. The advance of the K-means algorithm is that the means of each cluster can be compared per feature, leading to insights such as "the consumers in this cluster are more likely to have this behavior, whereas another cluster is distinctive on other features." The article is more marketing than data mining focused, but illustrates the insights that the algorithm gives. No other methods are used in the article to make segmentations.

Another example is found in an article from Reutterer et al. (2006). Here the goal is to distinguish customers based on shopping basket data. Due to the long running time when a lot of data points (and features) are included, it proposed vector quantization (VQ) algorithms. Via stochastic approximation the objective function will be minimized, leading to a speed up of the running time, making large databases able to cope with the algorithm.

The disadvantage is that the problem is NP-hard (Aloise et al., 2009), so it takes much time to solve the problem for a huge dataset. An option is to 'train' the data on a random subset of the dataset and then classify the other points based on the nearliest centroid. The global minimum will in general not be achieved due to the random starting points.

### 3.1.3 Logistic Regression

Logistic regression (LOG) is a technique for predicting a binary dependent variable invented by Cox (1958). Therefore, logistic regression is useful for marketing purposes to discriminate responders from non-responders. The mathematical formulation is given by (Menard, 2010):

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{2}$$

With $\alpha$ is the intercept and $k$ the number of independent variables. $Y$ is binary (either 0 or 1) and $\epsilon$ the error. The $\beta$'s will be estimated by the model to minimize the error. For more than two categories, multinomial logistic regression can be used. The technique is that a likelihood function will be maximized to fit to the data. Instead of only displaying a category, the probability will be calculated. Different reasons why the technique is popular in marketing can be given:

- The technique is conceptually simple (Bucklin & Gupta, 1992).
- The posterior probabilities are used rather than creating discrete groups of individuals (as in RFM).
- The technique provides relatively quick and robust results (Neslin et al., 2006).

According to Coussement et al. (2014), in which a case-study was presented (3.1.1), the logistic regression method only scored significantly better in the case of a high response rate. In the research only transactional features are used, whereas extending it with demographical and psychographical data may lead to improved results. The paper also concludes, alongside of the paper of Neslin et al. (2006), that the method is more robust, also under the circumstance of less accurate data.

In a research of Heilman et al. (2003), multinomial regression is applied and both demographical as transactional data is included. The data used is a panel dataset for Consumer Packed Goods (daily used goods by the average consumer). The goal is to segment the consumers for a better direct marketing impact and next to multinomial regression, artificial neural networks (ANN) are used. In all cases (demographical only, demographical + one purchase and demographical + all purchases), ANN outperforms the regression model.

With all the new data mining techniques available, logistic regression may seem to be out-dated. However, the model is still helpful for predicting a probability to buy, which will for a decision tree be mostly the same for a large segment (Antipov & Pokryshevskaya, 2010). This is caused due to the fact that many customers are in the same node, so will all have the same probability. Next to this, a smaller sample satisfies compared to decision trees. Finally, Antipov and Pokryshevskaya (2010) claims: "It often performs better than some state of the art techniques in terms of AUC, accuracy and other performance measures." Where AUC is the Area under the Curve, most of the times under the receiver operating characteristic curve (AUROC). However, for this statement no prove is provided.

### 3.1.4 Decision Trees

There are different types of decision trees created over time. In the basis they work the same, but have a different type of evaluating measure to construct the tree. In specific the Chi-square automatic interaction detection (CHAID) algorithm is used for segmentation. Therefore, this one will be explained extensively.

The CHAID algorithm was first published by Kass (1980). For customer segmentation, a decision tree will be constructed by splitting the entire group (in the root node), into smaller homogeneous subsets of customers. These splits can be into two or more nodes, depending on different values of an independent variable. In the CHAID algorithm, the child nodes (node below a parent node) will be distinguished based on the Chi-square test (categorical features) and F-test (continuous features). The predictor is one of the features that predicts the dependent variable.

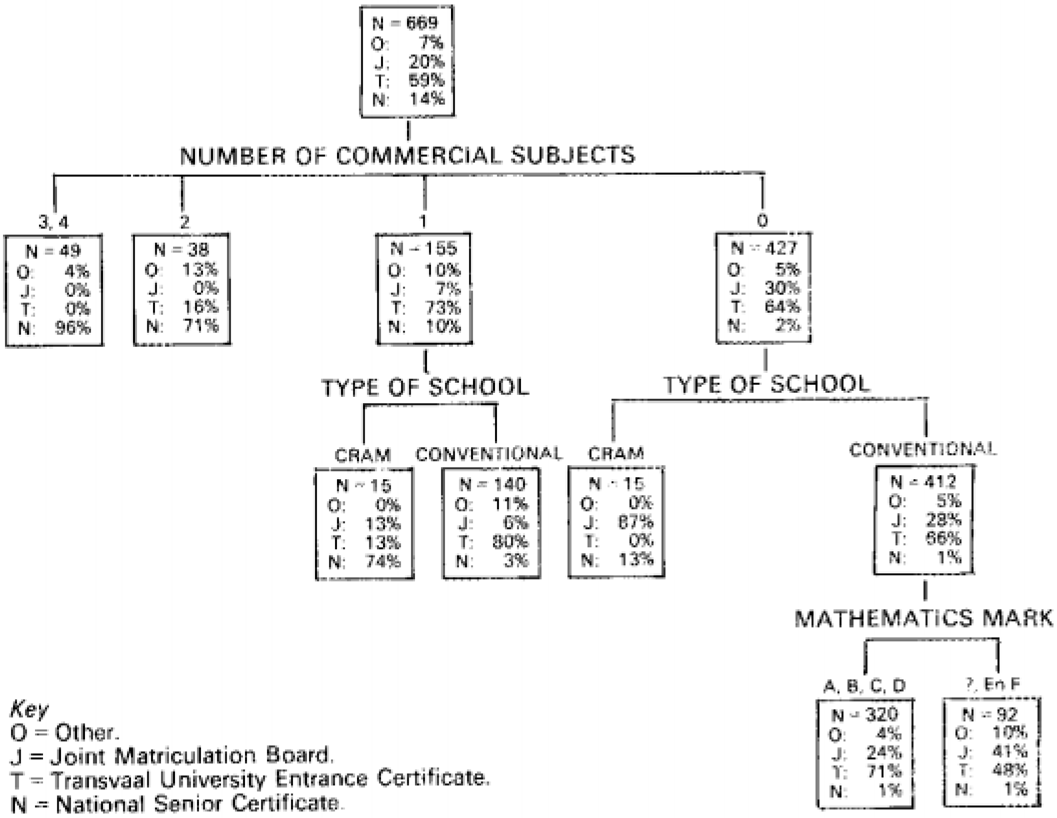
N = 669 O: 7% J: 20% T: 59% N: 14%
NUMBER OF COMMERCIAL SUBJECTS
3, 4
N = 49 O: 4% J: 0% T: 0% N: 96%
2
N = 38 O: 13% J: 0% T: 16% N: 71%
1
N = 155 O: 10% J: 7% T: 73% N: 10%
0
N = 427 O: 5% J: 30% T: 64% N: 2%
TYPE OF SCHOOL
CRAM
N = 15 O: 0% J: 13% T: 13% N: 74%
CONVENTIONAL
N = 140 O: 11% J: 6% T: 80% N: 3%
TYPE OF SCHOOL
CRAM
N = 15 O: 0% J: 87% T: 0% N: 13%
CONVENTIONAL
N = 412 O: 5% J: 28% T: 66% N: 1%
MATHEMATICS MARK
A, B, C, D
N = 320 O: 4% J: 24% T: 71% N: 1%
?, En F
N = 92 O: 10% J: 41% T: 48% N: 1%
Key
O = Other.
J = Joint Matriculation Board.
T = Transvaal University Entrance Certificate.
N = National Senior Certificate.

*Figure 2:* Simple illustration of the outcomes of a CHAID algorithm[2]

The following steps will be executed for categorical features (Kass, 1980):

1. For each predictor, cross-tabulate the categories of the predictor with the categories of the dependent variable.

2. Find the pair of categories of the predictor whose 2 × d (number of categories) subtable is least significantly different. If the critical value is not reached, this can be merged into a single compound segment. Then repeat this step for the remaining predictors.

3. For each existing compound category with three or more than the original categories, find the most significant binary split. If the value is beyond a critical value, the split should be implemented and return to step 2. This can be seen in Figure 2 where now the different sub-categories are created. For `number of commercial subject` four categories 0, 1, 2 and 3+4 are created and for `mathematics mark` the two categories A+B+C+D and E+F are created. This step may be modified in the case of a minimum number of observations that should be present in a root.

[2] from: An Exploratory Technique for Investigating Large Quantities of Categorical Data (Kass, 1980)

4. Calculate the significance of each optimally merged predictor and isolate the most significant one. If this significance is higher than the critical value, subdivide the data according to the categories of the chosen predictor. In the example (Figure 2), `number of commercial subjects` is the most significant one and therefore ends up as the first split. The data is divided in the categories determined in step 3.

5. For each partition of the data that has not yet been analyzed, go back to step 1.

The process of the algorithm will stop when either some criterion (which can be given as input) is met or when the tree is fully grown. The final child nodes (also known as terminal nodes) are the segments that will be the output of the algorithm to be used for segmentation. Figure 2 shows an example of a small decision tree. Based on seven different features, the matriculation will be predicted. The algorithm only selects three features to distinguish different segments.

According to the case study of Coussement et al. (2014) (section 3.1.1), decision trees are recommended over logistic regression and RFM when no data inaccuracy is known or suspected. Only in the case with a low response rate decision trees perform just as good as RFM. Also in the case of a lower data accuracy the CHAID decision tree seems to be the most robust.

Just like with logistic regression and K-means clustering, there can be added more features than transactional data only. Min et al. (2002) uses another type of decision trees, C5.0 which is invented by Quinlan (1993). Instead of using the Chi square statistic, information entropy is used as statistic. The goal of the research is to develop a customer retention strategy. Noteworthy, is that the researchers did not use the terminal nodes as outcome for possible clusters, but instead use the tree to create if-then rules. Only rules with a high accuracy (so a high chance of being true) are selected. Especially when there are many terminal nodes, which is difficult to interpret for the marketing purposes, these rules can be useful. Unfortunately, they only test one algorithm, so there cannot be concluded if the algorithm would be the best choice for segmentation.

From the literature where decision trees are used can be concluded that the method is promising. Difficulties may be that for using the algorithm a target feature should be set on which the distinctions will be made. For a general segmentation strategy, this will be difficult to determine, but does have a huge impact on the outcomes. Furthermore, the algorithm can end up with too many segments for which it will be impossible to create a marketing strategy on. If this amount should be smaller, only a few features will be selected on which the segments will be divided.

## 3.2 Application on Airline

The basis of this research is about which models will be applicable to an airline for segmentation. Using the observations that are made in the section before (3.1), this section will come to the conclusions.

The RFM model is easy to implement for an airline, since the transactional information is available and the interpretation is easy. However, some disadvantages can already be mentioned. Firstly, an airline has much more information available about the customer, such as the destination flown to, length of stay and the number of days booked before departure, which are all good indicators of what kind of travel a customer is interested in. The marketing strategy can be adapted to this, but all this kind of information is not available to implement in the model. Secondly, there are many customers that only fly once or even not in the last years, so many customers will be in the category 1-1-1 (not recent, not frequent and low/no spending). It is doubtful that all these customers are in need of the same communication strategy. The only way in which this approach can be useful, is to combine it with other segmentation techniques.

A K-means clustering algorithm seems to fit the marketing purposes, because all kind of features can be included. Another advantage is that the features on which a cluster is distinctive from other clusters can be easily interpret. For example, people who travel mostly with children will have different needs than business travelers and these may become different clusters (if this is the most distinctive). What may give difficulties, is that all kind of features can be included (also categorical with K-prototype), but all features will have an evenly weighted impact on the objective function. Therefore, outliers should be excluded (or adjusted), because after normalization, the other points will not be distinctive anymore. The amount of clusters (K) that needs to be determined, can be seen as an advantage. The objective function becomes lower when more clusters are added, but the optimal number of clusters can be determined using the Elbow method. However, in the end it is important for the marketing purposes that the segments are clear and workable with.

Although there are different reasons stated why logistic regression is useful, it seems from the case studies that have been conducted using logistic regression (which are not so many for general segmentation), that it is in most of the cases not the optimal technique. The probability that will be in the output for a customer can be very useful for response modeling, but the distinctive features in a segment are difficult to interpret. Therefore, there can be concluded that this modeling technique will not suite segmentation.

In the different studies conducted, decision trees perform good, are robust, categorical as well as numerical features can be used and next to transactional also other type of features can be included. No normalization is needed and no distribution of the features is assumed. Therefore it seems to be the best algorithm. However, for a general segmentation for an airline some problems are foreseen. Most important is that there are either two cases. When many features are added to a decision tree (and there are many customers in a database for an airline) the tree will create many segments. For these segments it will simply be impossible to set up a separate marketing strategy. On the other hand, when not so many segments may be built, the algorithm will only choose a few features. The example in Figure 2 illustrates that already seven segments are created based on only three features. Probably it will be the best option to give as input of the algorithm that the minimum observations in a node should be above a certain level (e.g. 10,000). Then the segments with the highest impact on the business may be selected to create a separate marketing strategy for.

# 4 Response Modeling

Also without predefined segments, a selection of customers may be targeted for a certain (new) product in order to save marketing costs. In response modeling, the probability of response (based on historical responses) as a function of features will be predicted. For email marketing, the response can be measured on different levels:

- Open Rate: number of first opens divided by the total mails delivered.
- Click-Through Rate: number of clicks divided by the number of opens.
- Conversion Rate: number of actions undertaken by the receiver divided by the total mails delivered.

In this section, the different product types wherefore data mining in email marketing may be deployed will be discussed. Afterwards, some comparative studies in the direct marketing context are shown. Then the different algorithms and there application in different case studies will be described and the chapter ends with the expectation of the application of the techniques on an airline.

## 4.1 Product Types

According to Fayyad et al. (1996), data mining can be used in order to explore implicit and useful information from a large database. Patterns can be discovered to determine likely buyers and increasing the response rate. There can be distinguished two situations (Ling & Li, 1998): an existing product or a new product that will be promoted.

In the first situation, there are already buyers of the product that have found the product due to previous marketing actions (say X%). Most of the times X is really small, because additional marketing actions are needed, but should be sufficient large in order to explore patterns. This might be around the 1 percent, but it depends on the size of the database. Data mining can be used to see the patterns and to search for likely to buy customers in the other (100-X)% of all customers. Ling and Li (1998), created a road map for this:

1. Get the data of all customers of which the X% is the amount of buyers of the product.
2. Apply data mining to the dataset:
   - Overlaying: add additional information to the customers, such as geographical and demographical information.
   - Data pre-processing: transform data to a usable format, for example change birth dates to age (ranges) and deal with the missing values.
   - Split the data into a train and test set.
   - Apply a learning algorithm (decision tree, neural network, etc.) to the train set.
3. Evaluate the patterns found on the test set. It may be the case that the results are not satisfying. This can either depend on the learning algorithm (settings) or the pre-processing of the data. Multiple iterations may take place.
4. Use the patterns found to predict likely buyers of the product.
5. Promote the product to the likely buyers.

In the second situation, a new product will be branded, so there are no purchases yet. In this case Ling and Li (1998) propose that the company can conduct a pilot study, in which customers will be randomly assigned to the pilot, for example 5% of the total customers. Then again a X% may respond to the offer and the same steps as the first situation with buyers can be applied now. The situation is quite the same, but in the second situation the amount of customers that bought the product (X% of 5% of the total customers) will be lower than in the first situation.

Sometimes the new product only satisfies the needs of certain customers. Take for example KLMs Flight Bundle, in which a customer can buy a bundle of multiple flights to a certain destination at a fixed price. This product will only serve the needs of customers that fly multiple times a year to the same destination. It will in such a case not be a good idea to assign random people to the pilot, but to only target the people who have shown the recurring behavior. Then again a X% may respond to the offer and the data mining steps as described by Ling and Li (1998) can be followed to find new potential customers.

Problems where the data mining algorithm will ran into is that the datasets will be most of the times extremely unbalanced, due to the low percentage of actual buyers of the product. The learning algorithm will classify everyone as a non-buyer and will score a high accuracy. A possible solution can be to down-sample the dataset, for example select a ratio of 1:5, meaning that there will be one positive case (buyer) against 5 non-buyers (Kubat et al., 1997). Next to this, a customer is only classified as either a buyer or a non-buyer. It may lead to cases where the business owner wants to sent the email to 20% of all customers, whereas the algorithm only predicts 5% as potential buyers. Also strategies as we want to call the first 100 of most likely buyers and sent an email to the next 5,000 of most likely buyers will be impossible. The algorithm therefore needs some kind of probability estimation.

## 4.2 Comparative Studies

Fortunately, some comparative studies have been conducted, in which multiple datasets and multiple algorithms are both tested. This can be very useful input for which algorithms may work for an airline.

Coussement et al. (2015) did a very useful research. A case-study was performed on multiple datasets of different types of companies on which many different types of models are tested (Table 1). In this way, there can be made a good comparison, with still in mind that it are only four datasets and it does not guarantee that it will work for an airline as well. The evaluation score they use is AUROC, the Area Under the ROC curve, which is according to Provost et al. (2000) good to compare classification performance. Beside, it is many times used in the evaluation of response models (Baesens et al., 2002). The performances of all models on the different datasets are stated in Table 2.

*Table 1:* Explanation datasets to which is referred in Table 2[3]

| Dataset | Direct marketing business | # Independent variables | Response rate |
|---|---|---:|---:|
| 1 | Non-profit organization | 11 | 27.42% |
| 2 | Catalogue company | 142 | 2.46% |
| 3 | Specialty catalogue company | 250 | 5.36% |
| 4 | Gift catalogue company | 87 | 9.56% |

*Table 2:* AUROC performance measured on four datasets[3]

| Dataset | LOG | LDA | QDA | ANN | NB | CHAID | CART | C4.5 | KNN-10 | KNN-100 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.6630 | 0.6685 | 0.6104 | 0.6843 | 0.6206 | 0.6839 | 0.6733 | 0.6685 | 0.6285 | 0.6693 |
| 2 | 0.6420 | 0.6411 | 0.5969 | 0.6025 | 0.5822 | 0.6366 | 0.6394 | 0.5372 | 0.5235 | 0.5665 |
| 3 | 0.8199 | 0.8150 | 0.7701 | 0.8225 | 0.7305 | 0.8622 | 0.8360 | 0.8192 | 0.6783 | 0.7503 |
| 4 | 0.7900 | 0.7959 | 0.7528 | 0.8231 | 0.6706 | 0.8157 | 0.8007 | 0.7763 | 0.6976 | 0.7635 |

Olson and Chae (2012) did a comparative research as well. They used two different datasets, one with catalog sales and one of individual donor distributions to a non-profit organization. The response rates of the datasets are respectively 9.6% and 6.2%. Only transactional information is taken into account. The models are evaluated on the prediction accuracy as well as the gain percentage. The performances of all models on the different datasets are stated in Table 3. Degenerate shows the performance in the case all customers are classified as non-responders.

*Table 3:* Accuracy measured on two datasets[4]

| Dataset | Degenerate | RFM | LOG | ANN | DT |
|---|---:|---:|---:|---:|---:|
| 1 | 0.9040 | 0.7650 | 0.9070 | 0.9110 | 0.9840 |
| 2 | 0.9371 | 0.6625 | 0.9385 | 0.9386 | 0.9386 |

[3] from: Improving direct mail targeting through customer response modeling (Coussement et al., 2015)

[4] from: Direct marketing decision support through predictive customer response modeling (Olson & Chae, 2012)

Then the last comparative study that will be presented is done by Cui et al. (2008). Instead of using multiple datasets, they investigated only one. However, they compare different models with two different evaluation methods and three different performance criteria. The dataset is originating from a catalog company based in the United States. It contains more than 100,000 customers and a response rate of approximately 5.4%. For the Bootstrap method the 0.632 bootstrap validation method is used and for Cross-fold a ten-fold cross-validation is applied. For every performance criteria, the best score is underlined. In case of a high cost of misclassification, the simple error rate should be used. On cross-validation the most models and performance criteria score higher than using bootstrap. There can be seen that none of the models scores the best on all performance criteria and validation methods.

*Table 4:* Performance of different models under different performance measures[5]

| Performance criteria | Validation method | ANN | CART | LCM | LOG |
|---|---|---|---|---|---|
| Simple error rate | Bootstrap | 0.054 | 0.071 | 0.040 | 0.054 |
| | Cross-fold | 0.054 | 0.061 | 0.042 | 0.054 |
| AUROC | Bootstrap | 0.765 | 0.669 | 0.595 | 0.742 |
| | Cross-fold | 0.771 | 0.706 | 0.611 | 0.744 |
| Top decile lift | Bootstrap | 385 | 337 | 397 | 334 |
| | Cross-fold | 401 | 389 | 394 | 342 |

## 4.3 Data Mining Algorithms

Different data mining algorithms can be used to either predict which customers are likely to buy a certain product. In this section the different algorithms found in the literature and case studies will be briefly described. Since linear discriminative analysis (LDA), quadratic discriminative analysis (QDA), naive bayes (NB) and K-nearest neighbor (KNN) are only used in one comparative study, do not perform well and are not mentioned in other articles, they will not be further explained.

### 4.3.1 Logistic Regression

For the explanation of the technique, please see section 3.1.3. Levin and Zahavi (1998) produced a comparison study in which the models were evaluated on three different levels: profitability, goodness-of-fit and prediction accuracy. In the study a mailing campaign for a home equity loan is conducted. The tested models are logistic regression, ordinal regression, linear regression, tobit regression and two-stage models. They state that logistic regression performs well in selecting people for a promotion, but that it lacks the possibility to predict the profits and returns of customers. To improve this, a combination with a second model is made, called a two-stage model (Heckman, 1979):

1. The first model is the same as the first logistic regression model to estimate the probability of response: a binary discrete choice model on all observations.

2. The second is a linear regression model which is only applied to the responders (conditional model) to estimate the expected return.

The two-stage model is concluded to be the most suitable of the tested models for continuous choice problems. Furthermore, the algorithm is efficient in computation resources and it provides purchase probabilities for all customers. Unfortunately, only one case study is present and the model has not been tested on other datasets.

In the study of Coussement et al. (2015) (Table 2), there can be seen that decision trees and neural networks outperform logistic regression. However, from the statistical algorithms (logistic regression, linear discriminative analysis, quadratic discriminative analysis and naive bayes) logistic regression performs the best. In the study of Olson and Chae (2012) (Table 3), logistic regression performs well on the second dataset. From Table 4 can be concluded that Logistic Regression performs robust under the different validation methods, but never scores the best. This is probably due to the fact that customer responses are often non-linear and non-compensatory (Coussement et al., 2015).

[5] from: Model selection for direct marketing: performance criteria and validation methods (Cui et al., 2008)

### 4.3.2 Decision Trees

For the explanation of the technique, please see section 3.1.4. In the study of Coussement et al. (2015), the CHAID algorithm as well as CART which is another type of decision trees, perform very good. There can be concluded that these two types are superior to C4.5 (which is a predecessor of C5.0 in section 3.1.4). Also in the study of Olson and Chae (2012) decision trees perform the best on both datasets.

In Table 4 can be seen that for the CART algorithm, there is a huge difference in outcome based on which validation method is chosen. For the other algorithms the numbers are far closer. If we look for example at the top decile lift and bootstrap as performance indicator, CART seems to perform much worse than ANN and LCM and close to LOG. However, when we look at the top decile lift and cross-fold validation, CART performs reasonably close to ANN and LCM and seems to be far better than LOG.

Important advantages are that the model runs fast, complex and non-linear relations can be found and that there is no assumption on the distribution of a feature (Coussement et al., 2015). When the tree gets large, the number of rules may be excessive and it will not be able to interpret it anymore.

### 4.3.3 Artificial Neural Networks

The idea for Artificial Neural Networks (ANN) came from McCulloch and Walter (1943), then called Threshold Logic. They basically modeled it to understand how the brains work. Artificial neural networks try to copy the structure and functioning of the brain, with the idea that if it works in nature, it should also work for computers to learn. In general there are three different layers of neurons: the input layer, a hidden neuron layer and an output layer. In the input layer the independent variables are represented as nodes. The dependent variable (in this case the response) is represented in the output layer. In the layers between, one or more hidden layers of neurons are present. These layers are used to capture the non-linearity that is in the data. A simple illustration of the model is given in Figure 3.

During the training of an ANN, the weights will be determined for each connection, so that the network creates the best distinction between responders and non-responders. The formulas of the layers are given by (Ha et al., 2005):

$$z_j = \sigma(\sum_i w_{ji} x_i) \qquad y = \bar{\sigma}(\sum_j w_j z_j) \tag{3}$$

with $x_i$ being the $i^{th}$ independent variable (feature), $z_j$ the $j^{th}$ hidden neuron, $y$ the output neuron, $w_{ji}$ the weight from the $i^{th}$ node to the $j^{th}$ node, $w_j$ the weight from the $j^{th}$ node to the output node and $\sigma$ and $\bar{\sigma}$ the transfer function of respectively the hidden layer and the output layer. There are several options available as transfer function.

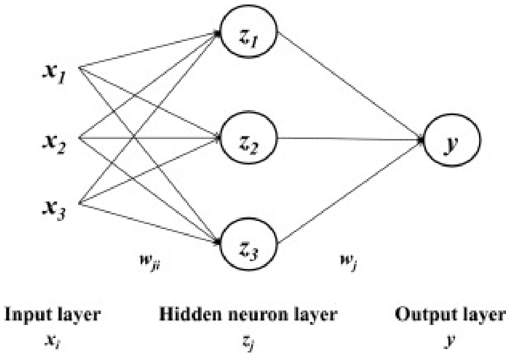

*Figure 3:* Simple illustration of a Neural Network[6]

In the paper of Coussement et al. (2015), neural networks achieve in dataset 1 and 4 the highest score. Next to this, it performs just as good as logistic regression and decision trees on the non-profit organization dataset in the study of Olson and Chae (2012). In another study of Suh et al. (1999), three different algorithms were tested: RFM, neural networks and logistic regression. The dataset they used was the membership and transactions of

[6] from: Improving direct mail targeting through customer response modeling (Coussement et al., 2015)

members of a golf club. The amount of customers (1580) is rather low. In the case study ANN outperforms RFM and logistic regression.

In the study of Kaefer et al. (2005) both Neural Networks as Multinomial Regression is used to classify the so called early purchase classification (EPC). The data they use is gathered from a panel consisting of Consumer Packed Goods (explanation in section 3.1.3) data. In this case the ANN outperforms multinomial regression. They state as well that it is important to observe customers initial and incremental purchases next to the demographical information available.

From Table 4 (research by Cui et al. (2008)) can be concluded that ANN performs the best on the AUROC performance criteria and on Cross-fold Top decile lift. On the other performance criteria it scores second best. Compared to the decision trees, the running time is much higher. Furthermore, the parameter optimization is difficult for ANNs and the interpretation is difficult (black box).

### 4.3.4 Latent Class Models

Latent Class Models (LCM) is a technique mostly used for characterizing persons and is invented by Lazarsfeld and Henry (1968). It is used for analyzing relationships in categorical data and can be used for predicting as well. Relationships among variables are scored at either the nominal or ordinal level of measurement (Lazarsfeld & Henry, 1968). The mathematical formulation is given by (McCutcheon, 1987):

$$\pi_{ijt}^{ABX} = \pi_{it}^{AX} \times \pi_{jt}^{BX} \times \pi_{t}^{X} \tag{4}$$

Where $\pi_{ijt}^{ABX}$ is the probability that a randomly selected case will be located in the $i$, $j$, $t$ cell, $\pi_{it}^{AX}$ is the conditional probability that a case in class $t$ of the latent variable ($X$) will be located at level $i$ of variable $A$, $\pi_{jt}^{BX}$ is the conditional probability of being at level $j$ of variable $B$, and $\pi_{t}^{X}$ is the probability of a randomly selected case being at level $t$ of the latent variable $X$.

In the study of Cui et al. (2008), LCM is used for predicting the response and it works in that case study the best for three out of the six performance measures. This may indicate that LCM can be a useful model for predicting the response. However, it is only mentioned in one article and never tested in other direct marketing applications. Furthermore, it scores high on the simple error rate and the top decile lift, but performs the worst on the AUROC criteria, which is used frequently as a performance criteria in direct modeling studies.

### 4.3.5 Ensemble

Fitting one model on the database may not be optimal. A combination of different models may improve the performance. Suh et al. (1999) for example, tested multiple models both separate and combined to see if this would improve the model. Firstly, they look at the correlation between the three different models. the RFM has a low correlation with both artificial neural networks as logistic regression (respectively 0.32 and 0.27). The correlation between ANN and LOG was rather high (0.62). The combination of ANN and RFM lead to improved results in all cases and the combination of LOG and RFM in approximately 80% of the cases. On the other hand, the combination of ANN and LOG did not improve the results. An ensemble of multiple models does not necessarily lead to an improvement of the performance. In this study the combination of models with a low correlation did lead to an improvement in most cases. Therefore combining multiple models which have a low correlation might be an interesting concept.

Ha et al. (2005) made instead of combining different type of models, multiple ANNs. They applied this on the same dataset that is used in the study of Coussement et al. (2015), one of a gift catalog company (dataset 4 in Table 1). They proposed bagging or bootstrap aggregating in order to prevent an ANN of over-fitting and instability. They compare the new models with a single ANN and LOG. The models are evaluated on the gain. The bagging variant of multiple ANNs performed the best. Logistic regression achieved a high accuracy, but this was due to the case that many customers were classified as non-responders, leading to a potentially huge loss of business opportunities.

## 4.4 Application on Airline

The basis of this research is about which models will be applicable to an airline for response modeling. Using the observations that are made in the sections before (4.2 and 4.3), this section will come to the conclusions.

Although in some articles more different models are used, most of the response modeling models used in direct marketing are logistic regression, decision trees and neural networks. As stated in the begin of the chapter, the response can be measured on different levels. Whereas the open rate and click-through rate are always measured as performance indicator, it should in the end lead to conversion to be successful. Due to the price of a flight ticket as well as that the customer should have time for a (short) holiday, there is almost no impulse buying. Therefore, conversion rates are usually low. There are no papers found in which response modeling is applied on an airline or a comparable sector, but apart from the response indicators, it seems unlikely that the models will perform very differently.

One general note that should be kept in mind, is that the training set of a response model should not only consist of customers previously selected by a response model (Coussement & Buckinx, 2011). The dataset will in that case contain a higher response rate. In this case there is a discrepancy between the prior distributions of the train and test set. In case this will be used, there can be corrected for this effect, by adjusting the posterior probabilities to the real response probabilities. Non-linear probability-mapping is according to them the best performing algorithm for this.

In many response modeling models logistic regression is used to predict the response. Due to the probability which will be given, it is very easy to interpret. Furthermore, the model can run very fast and is available in many statistical packages. Out of the other statistical models, logistic regression performs the best. However, if it is compared to data mining techniques, the algorithm is most of the times outperformed. The importance of features can be easily distracted and when there is nothing done with response modeling yet, it will be a good start.

Decision trees, mainly CHAID and subsequently CART, perform well in predicting the response. The algorithm is relatively fast and non-linear relations can be found. Moreover it is attractive for marketing managers to have somehow an idea of what the algorithm does and what has a high impact on the response. The CART algorithm seems to be sensitive for a different validation method. Haughton and Oulabi (1997) compare the two models in one case study. CHAID can split a parent node into more than two children and uses statistics to check if the split is significant. The two methods handle missing values different and in the article they conclude that both algorithms should be tested and compared for optimal test results. Based on the other studies seen, the preference will tend to the CHAID algorithm.

Then as last Artificial Neural Networks is also tested in many papers about response modeling. Just like decision trees, the algorithm can find non-linear relations. The algorithm performs in many studies comparable to decision trees. The main problem that is foreseen is that the implementation time of such an algorithm is usually very long. Next to this, the parameter optimization to get to the most optimal result may consume much time. From the network cannot be concluded directly which features are important, but suggested is to run one decision tree or regression model to get an idea of this (Ha et al., 2005).

The study of Coussement et al. (2015) is comparing many different algorithms on different datasets. Dataset 2 has the lowest response rate (2.46%) and dataset 3 has the second lowest response rate (5.36%) which will be comparable with the results of a marketing mailing of an airline. Logistic regression performs a little better than the CHAID and CART algorithms on dataset 2, but on dataset 3 the differences are bigger and CHAID performs by far better. The difficulty with the data mining techniques may be that an airline has budget and time constraints, making the implementation of such an algorithm more difficult than using a statistical package (not available for Neural Networks). Furthermore, marketing managers may think that the difference in performance in often small or not so important. However, relatively small differences often have a big impact on the bottom line profit. Reichheld and Sasser (1990) demonstrate that minor increases in the performance results in significant changes in the profit.

# 5 Response Indicators

In order to find out which features are good indicators for predicting the response, this section describes Genetic Algorithms, which can be used for feature selection. Next to this, observations of which features are possible and have a high predicting power are described.

## Genetic Algorithms

In multiple studies, Genetic Algorithms (GA) are used as a method to select or weigh features. This is an optimization techniques based on the biological evolution (Holland, 1975). The algorithm constantly tries different solutions by mixing the elements of better solutions to improve the search results. New solutions are created from the original solutions, based on the theory of the survival of the fittest. According to Goldberg (1989), GA is theoretically robust by using random selections and is applicable for searching in complex spaces. An example of how GA is functioning is given in Figure 4. Here the model created for predicting is an ANN. Then there is interaction between the GA and the ANN. GA searches for a good subset of features and this is passed to the ANN. The ANN calculates the performance of each subset and returns this to the GA. This process continues until the best feature subset is found.

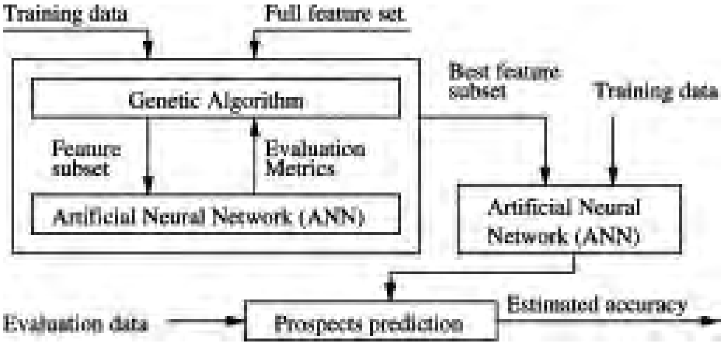

*Figure 4:* Simple illustration of the functioning of the Genetic Algorithm[7]

Chiu (2002) used this theory for feature selection in an insurance company case study. Seven different (socio-)demographical features are input of the GA, leading to one feature having the most impact in predicting: profession of the customer (72%). The study concludes that the GA has a better learning and testing performance compared to a regression model.

Also Kim and NickStreet (2004) used GA for feature selection. As well as the study above, the study is conducted for an insurance company. The algorithm chooses most of the times up to six features that are the most important. This can be an advantage because the Neural Network they use afterwards can be explained to marketing managers and does not turn into a black box. As most important features, multiple insurance-related features came up (last purchase behavior). Also some demographical features as income level and job are selected as important.

There can be concluded that Genetic Algorithms are due to there ability to not getting stuck into a local optimum, is a promising technique for feature selection. The importance of (recent) purchase behavior is mentioned, as well as the use of demographical features. The case studies were both applied to insurance companies, which have compared to an airline more (socio-)demographical information available such as income and (apparently) job information.

[7] from: An intelligent system for customer targeting: a data mining approach (Kim & NickStreet, 2004)

## Other Indicators

Heilman et al. (2003) states that: “Specifically, the longer a consumer has shopped a category, the less likely he or she will be to respond to a direct mailing that attempts to alter his or her behavior.” This is an interesting concept and it sounds logical since the consumers preference is more developed the longer they search. Furthermore, they determined how many previous purchases should be included to increase the accuracy. The analysis are based on panel data and up to 16 previous purchases seems optimal. For the airline industry, the majority of the customers do not fly frequently, so this may implicate that all historical data about customers can be useful.

You et al. (2015) used a RFM model to select the features for a K-means clustering algorithm, of which the important features there were the input for a CHAID algorithm. The objective for the CHAID algorithm is to recognize certain activities who have ordered a high quantity of a certain product. In their research they try to predict the customers satisfaction using RFM. This can then be used for the CHAID algorithm. It is not clearly described what the advantage of this is, but it may be another technique to select features.

# 6 Influence of Timing, Frequency and Content

Whereas the influence of the timing, frequency and content may be difficult to determine, it are important considerations for optimizing the direct marketing strategy. Therefore this short chapter is added, which might be a little less mathematical.

In many articles about response modeling, only one case study is presented and only one email moment is analyzed, so the studies are short-term focused. However, for an airline (and most other companies) it is important to have profitability of customers in the long-term. Piersma and Jonker (2004) do study this long-term relation by optimizing the mailing frequency instead of targeting profitable groups of customers for every new mailing. The objective of direct marketing is to maximize the expected profit. Next to this, the objective can be to minimize the non-response from both the perspective of costs as well as that of the customer: sending unwanted emails can lead to irritation and harm the long-term relation (Roberts & Berger, 1999; Micheaux, 2011). This results in a trade-off in the mailing frequency. The customer relations can be modeled long term and on an individual level. Piersma and Jonker (2004) created a Markov decision model for the mailing frequency which can be used by a decision support system. The direct mailer can control certain parameters as the length of the planning period and the maximum amount of emails that will be sent in the period. The model can then solve to which customer each email should be sent to maximize the long-term profitability.

In a study of Reimers et al. (2016), the importance of permission email marketing is stated, i.e. providing only communications that the customer wants to receive. To determine which content a customer wants to receive, there are two options mentioned. First, the customer should be able to indicate which type of emails he/she wants to receive. Secondly, data mining plays an important role as well. By carefully monitoring the customers purchases and click-through behavior on the site, the content can be modified accordingly. By doing this, the perceived usefulness and enjoyment of the marketing effort increases. Also Patron (2009) emphasizes the importance of data mining for the content a customer should receive. Various automatic mailing options are suggested, such as abandoned shopping basket (customer did not finish the order) but also viewed but not purchased. By better responding to this, the return on investment (ROI) can increase considerably. A case study is presented at a bank where a ROI of 750% was scored by using these types of data mining techniques.

Bult et al. (1997) used logistic regression in order to predict the response rate. Deviating from other research, they did not optimize the target only, but the mail characteristics as well. It is regarding mail characteristics like illustrations, format and kind of paper. This can be applied to email as well, in the sense that the subject line should be catchy and for the content counts the same. Due to the fact that 16 previous mailings were used and four target characteristics, many parameters should be estimated. In the end they predict for 288 different mailings (different content and lay out) including their costs which would be optimal. Some mailing outcomes (13 out of the 52) had only less than 10 households to sent the mail to. This will not be realistic, because the fixed costs will be far too high to set up a different mailing. There should always be made a trade-off between the improvement in response due to personalization and on the other side costs.

Many marketing websites state that next to frequency and content, also timing of the email is important in the emailing success. Most of the times diagrams are shown with the most successful days, based on A/B testing (creating two groups on which a different approach is tested). However, searching in the literature, there is no research related to this conducted. In studies that include the timing, it is most of the times about re-targeting customers instead of a clear preference of the customer leading to a higher open rate and conversion.

# 7 Conclusion

In this paper, the goal was to find which data modeling techniques could be used for optimizing the email marketing strategy for an airline. Since there has been no research conducted specifically for airlines or comparable businesses, this was quite challenging. Many studies are based on mails instead of emails. For email there are compared to mails no distribution costs. However, this does not mean that customers can be bombarded with emails, because as easily as people can opt-in for an email, they should also have the possibility to easily opt-out. Therefore, these mail studies are still relevant.

For the segmentation strategy, two techniques seem to be relevant: K-means clustering and decision trees (CHAID algorithm). In both type of algorithms, all kind of features can be included (when the K-prototype algorithm is used), whereas RFM is limited to transactional data only. For a general segmentation for which for each segment a marketing strategy will be set up, it is important to have not too many clusters. This may be a problem when using a tree algorithm, because many segments will be created. An option can be to put reasonable similar segments together again. For the K-means algorithm, the number of segments can be selected upfront. Disadvantages can be that all features count equally (no significance test), that the algorithm can get stuck in a local minimum and that the running time is long for a large database, which an airline has.

For response modeling it is more difficult to get to the conclusions. There have been conducted multiple comparative studies in the direct marketing setting, but only one (e)mail is tested. Generally, the data mining techniques artificial neural networks and decision trees outperform logistic regression. This is probably the cause of the response of customers not being linear. Decision trees are reasonably fast and provide information on which features are important in predicting the response. Meanwhile, neural networks take a longer time to create and optimize, and the algorithm is a black box. In many studies the two algorithms achieve around the same performance, sometimes neural networks perform better. There should be kept in mind that a small difference in performance could have a high impact on the revenue and profit.

In many articles the importance of recent customer purchases are stated (transactional data). Afterwards demographical features are mostly mentioned. So for the basic selection of features which indicate the response the previous purchases are important as well as demographical information. The demographical information was mostly mentioned in papers researching banking or insurance companies, which have more information available about all the customers. Next to these features, the previous responses specified to type of content (sales, transactional, inspirational etc.) can be included as well as the behavior of the customer on the website. Furthermore, Genetic Algorithms seem to have a good effect on selecting the feature importance.

Unfortunately, most of the articles only focus on one email and not on the long-term relationship with the customer, which will be important for airlines and most other companies. In other articles the importance of this is indicated. The frequency of emails that a customer wants to receive will differ per individual. This is also for the content different. To really optimize the email strategy, both frequency and content are important factors that should be included.

## Further research

There has been done no research yet on the airline industry itself for optimizing the email marketing strategy. Therefore it is difficult to conclude what would really work. However, in the three sectors in which research has been conducted, the same models perform well. It is not assumable that the models will predict very different for the airline industry, but probably other business related features should be included to get to great results. Due to the study of Cui et al. (2008), the importance of choosing the evaluation measure as well as the effect this could have on the results is stated, making it more difficult to make conclusions how models will perform in another business.

A difficulty for the response modeling techniques is that now all articles only have done one case study or experiment, so the long-term effects are not measured. Especially for an airline which sends next to sales, also inspirational and awareness emails, it is important to see which email a customer is interested in. With a continuous model for all emails there might be the risk that the amount of emails will be estimated lower than what the customer wants to receive.

For further research I therefore recommend to make a response model based on airline email data in which the frequency and content are taken into account as features. There can be started with a simple logistic regression as a benchmark. Afterwards data mining techniques such as decision trees or neural networks can further improve the prediction model.

# References

Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009). Np-hardness of euclidean sum-of-squares clustering. *Machine Learning*, *75*(2), 245-249.

Antipov, E., & Pokryshevskaya, E. (2010). Applying chaid for logistic regression diagnostics and classification accuracy improvement. *Journal of Targeting, Measurement and Analysis for Marketing*, *18*(2), 109-117.

Baesens, B., Viaene, S., van den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, *138*, 191-211.

Bucklin, R., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation - an integrated modeling approach. *Journal of Marketing Research*, *29*(2), 201-215.

Bult, J. R., van der Scheer, H., & Wansbeek, T. (1997). Interaction between target and mailing characteristics in direct marketing, with an application to health care fund raising. *International Journal of Research in Marketing*, *14*(4), 301-308.

Chiu, C. (2002). A case-based customer classification approach for direct marketing. *Expert Systems with Applications*, *22*(2), 163-168.

Coussement, K., & Buckinx, W. (2011). A probability-mapping algorithm for calibrating the posterior probabilities: A direct marketing application. *European Journal of Operational Research*, *214*(3), 732-738.

Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Systems with Applications*, *42*(22), 8403-8412.

Coussement, K., van den Bossche, F. A. M., & Bock, K. W. D. (2014). Data accuracy's impact on segmentation performance: Benchmarking rfm analysis, logistic regression, and decision trees. *Journal of Business Research*, *67*, 2751-2758.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B*, *20*(2), 215-242.

Cui, G., Wong, M. L., Zhang, G., & Li, L. (2008). Model selection for direct marketing: performance criteria and validation methods. *Marketing Intelligence & Planning*, *26*(3), 275-292.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining.* Press, Mento Park.

Goldberg, D. (1989). Genetic algorithms in search, optimization, and machine learning. *Reading, Mass. : Addison-Wesley Publishing Company.*

Ha, K., Cho, S., & MacLachlan, D. (2005). Response models based on bagging neural networks. *Journal of Interactive Marketing*, *19*(1), 17-30.

Haughton, D., & Oulabi, S. (1997). Direct marketing modeling with cart and chaid. *Journal of Direct Marketing*, *11*(4), 42-52.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153-161.

Heilman, C. M., Kaefer, F., & Ramenofsky, S. D. (2003). Determining the appropriate amount of data for classifying consumers for direct marketing purposes. *Journal of Interactive Marketing*, *17*(3), 5-28.

Holland, J. H. (1975). Adaptation in natural and artificial systems. *The University of Michigan Press.*

Huang, Z. (1998). Clustering large data sets with mixed numeric and categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283-304.

Jonker, J., Piersma, N., & den Poel, D. V. (2004). Joint optimization of customer segmentation and marketing policy to maximize long-term profitability. *Expert Systems with Applications*, *27*(2), 159-168.

Kaefer, F., Heilman, C. M., & Ramenofsky, S. D. (2005). A neural network application to consumer classification to improve the timing of direct marketing activities. *Computers & Operations Research*, *32*(10), 2595-2615.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*(2), 119-127.

KetchenJr, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, *17*(6), 441-458.

Kim, Y., & NickStreet, W. (2004). An intelligent system for customer targeting: a data mining approach. *Decision Support Systems*, *37*(2), 215-228.

Kubat, M., Holte, R., & Matwin, S. (1997). Learning when negative examples abound. *Proceedings of the 9th European Conference on Machine Learning.*

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Levin, N., & Zahavi, J. (1998). Continuous predictive modeling—a comparative analysis. *Journal of Interactive Marketing*, *12*(2), 5-22.

Liao, S., Chen, Y., & Hsieh, H. (2011). Mining customer knowledge for direct selling and marketing. *Expert Systems with Applications*, *38*(5), 6059-6069.

Ling, X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. *Knowledge discovery and data mining.*

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, *1*, 281-297.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of rfm, chaid, and logistic regression. *Journal of Business Research*, *60*, 656–662.

McCulloch, W., & Walter, P. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*(4), 115-133.

McCutcheon, A. L. (1987). *Latent class analysis.* Thousand Oaks, California: Sage Publications.

Menard, S. W. (2010). *Logistic regression : from introductory to advanced concepts and applications.* SAGE Publications, Inc.

Micheaux, A. (2011). Managing e-mail advertising frequency from the consumer perspective. *Journal of Advertising*, *40*(4), 45-65.

Min, H., Min, H., & Emam, A. (2002). A data mining approach to developing the profiles of hotel customers. *International Journal of Contemporary Hospitality Management*, *14*(6), 274-285.

Neslin, S., Gupta, S., Kamakura, W., Lu, J., & Mason, C. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204-211.

Olson, D. L., & Chae, B. (2012). Direct marketing decision support through predictive customer response modeling. *Decision Support Systems*, *54*(1), 443-451.

Patron, M. (2009). *Your best behavior* (Vol. 21). Haymarket Business Publications Ltd.

Piersma, N., & Jonker, J. (2004). Determining the optimal direct mailing frequency. *European Journal of Operational Research*, *158*(1), 173-182.

Provost, F., Fawcett, T., & Kohavi, R. (2000). The case against accuracy estimation for comparing induction algorithms. *Morgan Kaufman*, 445-453.

Quinlan, R. (1993). C4.5: Programs for machine learning. *Morgan Kaufmann Publishers.*

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections – quality comes to services. *Harvard Business Review*, *68*, 105-111.

Reimers, V., Chao, C., & Gorman, S. (2016). Permission email marketing and its influence on online shopping. *Asia Pacific Journal of Marketing and Logistics*, *28*(2).

Reutterer, T., Mild, A., Natter, M., & Taudes, A. (2006). A dynamic segmentation approach for targeting and customizing direct marketing campaigns. *Journal of interactive marketing volume*, *20*(3-4).

Roberts, M., & Berger, P. (1999). *Direct marketing management.* Prentice Hall, Upper Saddle River, NJ.

Shepherd, D. (1990). *The new dirert marketing.* Homewood, IL: Business One Irwin.

Steinhaus, H. (1957). Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.*, *4*(12), 801–804.

Suh, E. H., Noh, K. C., & Suh, C. K. (1999). Customer list segmentation using the combined response model. *Expert Systems with Applications Volume*, *17*(2), 89-97.

Vattani, A. (2011). k-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry*, *45*(4), 596-616.

You, Z., Si, Y., Zhang, D., Zeng, X., Leung, S. C. H., & Li, T. (2015). A decision-making framework for precision marketing. *Expert Systems with Applications*, *42*(7), 3357-3367.

Zahay, D., Peltier, J., Schultz, D. E., & Griffin, A. (2004). The role of transactional versus relational data in imc programs: bringing customer data together. *J Advert Res*, *44*, 3-18.