Improving response prediction in direct marketing by optimizing for specific mailing depths

D. Van den Poel, A. Prinzie & P. Van Kenhove

Department of Marketing, Ghent University, Belgium.

Abstract

Response modeling is a very important application field of classification methods in direct marketing because the success of a direct-mail campaign is highly dependent on who is being targeted. To date, standard classification models are applied to predict future purchasing behaviour for the complete customer file. In practice, however, companies use mailing budgets, i.e. only a subset of customers will be sent mail. Just those customers with sufficiently high-expected response rates are mailed to. The percentage of the total population that will actually receive the mailing is referred to as mailing depth. Hence, the real classification problem is not to classify all potential recipients as well as possible, but rather to find those customers, within the budget limitation, with the highest probability of response. Therefore, we propose an innovative alternative route to improved overall performance by tailoring the classification method to fit the problem at hand. We adapt binary logistic regression by iteratively changing the true values of the dependent variable during the maximum-likelihood estimation procedure. Those customers who rank lower than the cutoff in terms of predicted purchase probability, imposed by the mailing-depth restriction, will not contribute to the total likelihood. We illustrate our procedure on a real-life direct-marketing dataset comparing traditional response models to our innovative approach optimising for a specific mailing depth. The results show that for mailing depths up to 48% our method achieves significant and substantial profit increases.
1 Introduction

Response modeling is a very important application field of classification methods in direct marketing because the success of a direct-marketing campaign is highly dependent on who is being targeted. Basically, response modeling involves prediction of future purchasing, based on behavioral (often recency, frequency and monetary value (RFM)) and socio-demographic characteristics of the customer just before the direct-marketing action.

Generally, the issue of individual-level response modeling in database marketing is approached using ‘standard’ classification techniques (e.g. logistic regression and decision trees) to separate responders from non-responders. In this paper we take a fundamentally different approach to the same issue. We want to improve the predictive performance of a response model by tailoring the parameter estimation towards the target group only. To date, standard classification models are applied to predict future purchasing behavior for the complete customer file. However, largely because of budgetary reasons, companies only mail to a subset of the total customer population (typically referred to as mailing depth). Only those customers with sufficiently high-expected response rates are mailed to. Hence, the real classification problem is not to classify all potential recipients as well as possible, but rather to find those customers, within the budget limitation, with the highest probability of response. Therefore, we propose an innovative alternative route to improved overall performance by tailoring the classification method to fit the problem at hand.

The substantive relevance of an improved response prediction comes from the fact that even small improvements in response rates by using improved models can be commercially very rewarding. We refer to [1] for a quantitative example.

2 Methodology

2.1 Introduction

We propose an innovative alternative route to improved overall performance by tailoring a classification method to parameter estimation for the target group only. We adapt a classification method by iteratively changing the true values of the dependent variable during the maximum-likelihood estimation. By setting the real y-values of people not belonging to the target group to zero during the maximum-likelihood estimation of the parameters, we estimate the influence of the independent variables on the target customers only. This results in a response prediction tailored to the target group.
2.2 A classification method tailored to prediction of response probabilities for a target group

In this case, we choose *binary logistic regression* as classification method. This choice is justified by the fact that several authors (e.g. [2]) have shown that logit modeling may even outperform more sophisticated methods.

The probability of person \( n \) to opt for alternative \( i \) (e.g. response) is given by:

\[
P_n(i) = \frac{1}{1 + e^{-\beta(x_n - x_i)}}.
\]

We denote \( \beta = [\beta_1, \beta_2, ..., \beta_K]' \) as the vector of \( K \) unknown parameters. The \( \beta \)-parameters are typically estimated with Maximum Likelihood (ML).

For convenience, we analyze the log likelihood:

\[
L(\beta_1, ..., \beta_K) = \sum_{n=1}^{N} \left\{ y_n \cdot \log \left( \frac{1}{1 + e^{-\beta(x_n - x_i)}} \right) + (1 - y_n) \cdot \log \left( 1 - \left( \frac{1}{1 + e^{-\beta(x_n - x_i)}} \right) \right) \right\}.
\]

We solve for the maximum of \( L \) by differentiating the right hand side of eqn (2) with respect to each of the \( \beta \)'s and setting the partial derivatives equal to zero.

It is during this ML estimation of the \( \beta \)'s that we tailor the classification method towards an improved response prediction for the target group of the mailing campaign. We adapt *binary logistic regression* by iteratively changing the true values of the dependent variable (represented by \( y_n \) in eqn (2)). In the first iteration the response probabilities \( P_n(i)' \) are estimated based on the initial beta's \( \beta_0 \), which are derived from applying the traditional binary logit model to the estimation sample.

\[
P_n(i)' = \frac{1}{1 + e^{-\beta_0(x_n - x_i)}}.
\]

These first response probabilities are sorted in order to identify the customers to be mailed considering a given mailing depth (i.e. the target group). Given this mailing depth we determine a cutoff \( p \)-value (cutoff_1). For example, for a mailing depth of 10% and a mailing list of 50,000 customers, the target group counts 5,000 customers. By ranking the customers of the mailing list according to their first response probabilities, we select those 5,000 customers with the highest estimated first response probabilities as targets. We define the \( p \)-value of the last customer in the target group as the cutoff value \( \text{cutoff}_1 \). The value of the actual response variable \( y_{real} \) is set to zero for customers with a first response probability \( P_n(i)' \) lower than \( \text{cutoff}_1 \). Then the beta’s \( \beta'' \) are estimated given these adapted \( y \) values \( y_{adapted} \).
In the second iteration of the ML estimation the response probabilities \( P_n(i)^2 \) are estimated based on the \( \beta^d \) parameters. Again, the values of the actual responses \( y_{\text{real}} \) are set to zero for individuals with a new response probability \( P_n(i)^2 \) lower than the newly determined cutoff (i.e. cutoff2). The new estimation of the beta’s \( \beta^2 \) is based on the adapted \( y \) values \( y_{\text{adapted}} \).

In the case of small mailing depths, a second alternative procedure is proposed. We intervene during the maximum-likelihood estimation of the beta-parameters in a gradual way. For small mailing depths, only very few cases are left to estimate the beta-parameters on. This lack of a certain minimum number of observations makes the estimation of the parameters unstable. Therefore, during the first iterations of the ML estimation, beta parameters are estimated on a larger group than the target group. In practice, this is done by limiting the number of cases for which the real \( y \) values are set to zero during the first iterations of the estimation process. The number of cases for which the real \( y \) value is changed to zero is defined by following formula:

\[
Y_{\text{reset cases}} = \min\left( \frac{\text{number of non_targets} \times \text{iteration constant}}{\text{number of non_targets}} \right).
\]

The denominator is an integer that you can choose freely, however different from zero. For example, given a mailing depth of 10% and a mailing list of 50,000, the number of non-targets is 45,000. Suppose the constant to be five. During the first iteration of the ML estimation, the \( y \) values of 45,000 x 1/5 observations, i.e. 9,000 cases will be set to zero if their first purchase probability is lower than cutoff1. Indirectly, this means that the beta estimation in the first iteration is not limited to the 5,000 targets as proposed in the initial procedure, but is extended to 41,000 (50,000 – 9,000) cases. It is only after four iterations that the beta estimation is limited to the target group only.

In order to identify the adaptation rate for which the predictive performance reaches its maximum, we estimate the model for a specific small mailing depth with a constant (cf. eqn (5)) ranging from 2 to 10.

By iteratively changing the true values of the dependent variable (gradual or not) during the ML estimation of the beta’s, the purchase probability estimation is tailored to the target group.
2.3 A specific classification performance measure: the target PCC

The performance test for the adapted binary logit model also differs from the performance criteria used for classic classification methods such as, amongst others, the percentage correctly classified (PCC). The PCC is based on some descriptive statistics of the confusion matrix for binary classification [3].

Table 1: Confusion matrix.

<table>
<thead>
<tr>
<th>Predicted Status</th>
<th>Buyer</th>
<th>Non-buyer</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Status</td>
<td>True Positive (TP)</td>
<td>False Negative (FN)</td>
</tr>
<tr>
<td>Non-Buyer Status</td>
<td>False Positive (FP)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>

The PCC measures the overall accuracy as follows:

\[ PCC = \frac{TP + TN}{TP + FP + TN + FN} \]  

(6)

To date the goodness of fit of a classification model was measured by calculating the PCC on the whole mailing list (e.g. 50,000 customers). In our case, we are only interested in the predictive performance of the classification model on the target group (e.g. 5,000 customers for mailing depth 10%). In order to test the goodness of fit of both models on the target group, only two descriptive statistics of the confusion matrix can be used. After all, every target is predicted as a buyer. Within the target group we calculate the number of true positive (TP) cases. As final criterion we use following ratio:

\[ TARGET_{PCC} = \frac{TP}{TP + FP} \]  

(7)

2.4 Testing the statistical difference in performance between the adapted and classic binary logit model

In order to test the statistical difference between the performance measures of the adapted binary logit model and the traditional one for a specific mailing depth, we use a bootstrap procedure [4]. We use 100 bootstrap samples to approximate the empirical distribution of the target PCC for both models. We look for each mailing depth if the confidence interval for the target PCC of the classical binary classification model overlaps with the confidence interval of the target PCC of the adapted binary classification model. We also calculate the difference between the performance measures of both models over the 100 bootstrap samples for a specific mailing depth. The distribution of the target performance difference (referred to as TPD) is analyzed for each of the mailing depths.
2.5 Validation of results

The results are validated on a validation sample in order to know whether the target performance difference between the adapted binary logit model and the classic model found in the training samples, is generalizable to the total population. The following procedure was used for each mailing depth:

1. Run the classic binary logit model on the validation sample using the beta estimates from the same model on bootstrap sample $r$ (R: $1 \rightarrow 100$). The target PCC of the traditional binary logit model on the validation sample is referred to as target $PCC_{vadap_{r}}$.

2. Run the adapted binary logit model on the validation sample using the beta estimates from the same model on bootstrap sample $r$ (R: $1 \rightarrow 100$). The target PCC of the adapted binary logit model on the validation sample is referred to as target $PCC_{vadap_{r}}$.

3. Calculate the difference in target performance between both models:

$$TPD_{v, r} = target \ PCC_{vadap_{r}} - target \ PCC_{vtrad_{r}}.$$

4. Repeat step 1 to 3 for all 100 bootstrap samples.

5. Analyze the distribution of the 100 target performance differences, i.e. $TPD_{v, 1} \rightarrow TPD_{v, 100}$.

3 Direct-marketing application

3.1 Introduction

In this section we apply our new alternative classification procedure on a real-life direct-marketing dataset from a major European mail-order company and compare it with a traditional response model. We use a binary classification model with as predictors recency, frequency and monetary value (RFM), calculated on a four-year period from July 1993 to June 1997. Cullinan [5] is generally credited for identifying the three RFM variables. Since then the literature has accumulated so many uses of these three variables, that there is overwhelming evidence both from academically reviewed studies [1] as well as from practitioners’ experience that the RFM variables are the most important set of predictors for modelling mail-order repeat purchasing. We define recency as the number of days since the last purchase. Frequency is operationalized as the number of purchases (excluding returns) made in the four-year period July 1993 – June 1997. Finally, monetary value is defined as the total net monetary amount spent during the last 12 months, i.e. sales from past orders minus the monetary value of returns and refunds. In this study, we focus on purchase incidence modelling. We try to build a model that predicts whether a customer (cf. repeat purchase modeling) will make a purchase from any product category during the next mailing period July 1997 - December 1997. The population class proportion of buyers for our bootstrap samples is on average 54.94% buyers and 45.06% non-buyers.

Out of the total dataset, we have randomly drawn an estimation sample and a holdout or validation sample of each 50,000 observations. The holdout sample
only consists of observations not used during model estimation. The 100 bootstrap samples are drawn from the estimation sample.

3.2 Results

3.2.1 Testing the statistical difference in performance between the adapted and classic binary logit model

We check for each mailing depth if the confidence interval for the target PCC of the classical binary classification model (cf. benchmark target PCC) overlaps with the confidence interval of the target PCC of the adapted binary classification model. For mailing depth 5%, there is no overlap (cf. Figure 1). Hence, a statistical test for the difference between the adapted and classic binary logit model at mailing depth 5% is not even necessary. We conclude that at a mailing depth of 5% our adapted model outperforms the benchmark.

Figure 1: Performance of the classic and adapted binary logit model for mailing depth 5%.

We also calculate the difference between the performance measures of both models over the 100 bootstrap samples for a specific mailing depth. The distribution of the target performance difference (referred to as TPD) is analyzed for each of the mailing depths.

Firstly, we compare the TPD between the non-gradual/gradual adapted binary model and the classic binary logit model for small mailing depths. As the attentive reader might remember, we proposed to intervene during the maximum-likelihood estimation of the beta parameters in a gradual way for small mailing depths. From Figure 2 we can conclude that the gradual adapted binary logit model outperforms the non-gradual model. The average differences between the gradual adapted and the classic binary logit model for a specific mailing depth are larger than the average differences between the non-gradual adapted model and the classic one for the same mailing depth. For a mailing depth of 1%, both the gradual and the non-gradual adapted binary logit model were very unstable.
Secondly, we make a distribution analysis of the TPD between the adapted and classic binary logit model for all mailing depths. Remember that for mailing depths 2 to 7 we opt for the gradual adapted model, whereas for larger mailing depths we choose the non-gradual one. It seems that our adapted binary logit model is especially well fitted to response predictions at smaller mailing depths (see Figure 3). We define the ‘average TPD’ as the sum of the differences between the target performances of the adapted versus the classic models, run for each of the 100 bootstrap samples at a specific mailing depth, divided by 100. At a mailing depth of 5%, the adapted binary logit model reaches its maximum average TPD of 2.37% and also the highest maximum TPD of 3.68%. The highest minimum TPD is 1.32% for mailing depth 8%.

Up to a mailing depth of 20%, there are substantive TPDs between the model with beta-parameter estimation tailored to the target group only and the classic...
binary logit model. The average TPD between the gradual/non-gradual adapted binary logit model and the classic model for mailing depths smaller than 21% is 1.50% on average (over mailing depths 2 to 20%), whereas the highest maximum TPD is 3.68% and the smallest TPD is 1.27%. Between mailing depths 21 and 48% the TPD is rather small, i.e. 0.21% on average (over mailing depths 21 to 48%). For mailing depths larger than 48% the average TPD is also small, but negative.

3.2.2 Validation of results
We want to validate the higher target performance of the adapted binary logit model compared to the performance of the classic binary logit model found on the training samples for mailing depths smaller than 48%. We choose to perform the validation for mailing depth 5%. We wonder if the large performance difference between the adapted and classic binary logit model at a mailing depth of 5% found at the bootstrap estimation samples (highest average TPD with 2.37% and highest maximum TPD of 3.68%), will be repeated for the validation sample.

Firstly, we look if the confidence interval for the ‘target PCC’- measure for the classic binary classification model overlaps with the confidence interval of the ‘target PCC’- measure of the adapted binary classification model. As Figure 4 shows, there is no overlap. Consequently, we can conclude that the TPD between the adapted and the classic binary logit model is valid for mailing depth 5%.

![Figure 4: Performance of the classic and adapted binary logit model for mailing depth 5.](image)

Secondly, we analyze the distribution of the TPD between both models for mailing depth 5%. As the minimum TPDs differ from zero (see Table 2), we can reject the null hypothesis. Moreover, because the minimum differences are higher than zero, we can conclude that for mailing depth 5%, the adapted binary logit model has a higher target performance than the classic binary logit model.
Table 2: TPD for mailing depth 5% on the 100 bootstrap estimation samples and the validation sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>min. diff</th>
<th>max. diff</th>
<th>mean. diff</th>
<th>std. dev. diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimation</td>
<td>0.16</td>
<td>3.68</td>
<td>2.36</td>
<td>0.51</td>
</tr>
<tr>
<td>Validation</td>
<td>2.28</td>
<td>3.44</td>
<td>3.13</td>
<td>0.16</td>
</tr>
</tbody>
</table>

4 Discussion

In this paper we have shown that it is possible to improve response prediction by optimizing for a specific mailing depth. For mailing depths up to 48% the performance of the adapted binary logit model is significantly higher than the classic binary logit model. Substantial statistical significant target performance differences have been found for mailing depths up to 20%, on average 1.50% and maximum 3.68%. From a managers perspective, these small improvements in response prediction may result in substantial profit increases.

As our alternative estimation method is especially highly performing at small mailing depths, a great opportunity lies in the application of the model for customer acquisition purposes, where generally only a very small selection of the total list of possible prospects can be mailed to. Another direction for further research exists in the optimization for specific mailing depths of other classification methods than binary logistic regression, like non-linear classification methods (e.g. artificial neural networks).

References