KATHOLIEKE
UNIVERSITEIT
LEUVEN

DEPARTEMENT TOEGEPASTE ECONOMISCHE WETENSCHAPPEN

RESEARCH REPORT 0019

WRAPPED FEATURE SELECTION FOR NEURAL NETWORKS IN DIRECT MARKETING

by
S. VIAENE
B. BAESENS
D. VAN DEN POEL
G. DEDENE
J. VANTHIENEN

D/2000/2376/19

Wrapped Feature Selection for Neural Networks in Direct Marketing

S. Viaene[1], B. Baesens[1], D. Van den Poel[2], G. Dedene[1] & J. Vanthienen[1]
[1]*K. U. Leuven, Dept. of Applied Economic Sciences, Belgium*
[2]*Ghent University, Dept. of Marketing, Belgium*

Abstract

In this paper, we try to validate existing theory on and develop additional insight into repeat purchasing behaviour in a direct-marketing setting by means of an illuminating case study. The case involves the detection and qualification of the most relevant RFM (Recency, Frequency and Monetary) features, using a wrapped feature selection method in a neural network context. Results indicate that elimination of redundant/irrelevant features by means of the discussed feature selection method, allows to significantly reduce model complexity without degrading generalisation ability. It is precisely this issue that will allow to infer some very interesting marketing conclusions concerning the relative importance of the RFM-predictor categories. The empirical findings highlight the importance of a combined use of all three RFM variables in predicting repeat purchase behaviour. However, the study also reveals the dominant role of the frequency variable. Results indicate that a model including only frequency variables still yields satisfactory classification accuracy compared to the optimally reduced model.

1 Introduction

It need not be emphasized that customer retention is as least as important as customer acquisition in the current context of competitive markets, not in the least for mail order companies. Service providers are finding themselves in mature markets in which they have to switch their marketing efforts from acquiring new customers towards retention of existing customers [10,21]. Customer relations undoubtedly represent an important opportunity cost: 5% fewer customers may result in profit losses between 25 and 85% [20].

The objective of this paper is, by means of an illuminating marketing case study, to validate existing theory on and develop additional insight into repeat purchase behaviour in a direct marketing setting by using a neural network set-up. For modelling repeat purchase behaviour, several techniques have already been proposed and operationalised. To date, in most research on this topic traditional statistical models are used. Examples include: logit, probit and discriminant analysis (both linear and quadratic). In this paper we will use Multilayer Perceptron (MLP) neural networks as an extension to previous work [24]. As universal approximators, MLPs have shown to be very promising supervised learning tools for modelling non-linear relationships. This, especially in situations where one is confronted with a lack of domain knowledge, which in turn prevents any valid argumentation to be made concerning model selection bias on the basis of prior knowledge.

The empirical study focuses on the *purchase incidence*, i.e. the issue whether or not a purchase is made from any product category offered by the direct mailing company. As to the choice of the independent variables, the set-up is limited to assessing the predictive importance of the traditionally discussed (R)ecency, (F)requency and (M)onetary variables. This choice is motivated by the fact that most previous research cites them as being most predictive and because they are internally available at very low cost [2,4].

The main focus of the research reported on in this paper goes out to assessing the relevance of the RFM variables by means of a wrapped neural network feature selection mechanism. It will be shown that by making use of the discussed feature selection method, model complexity can be significantly reduced without degrading generalisation ability. It is precisely this issue that will allow to infer some very interesting marketing conclusions concerning the relative importance of the RFM-predictor categories. As argued in [24], this has never been thoroughly investigated, let alone in the context of connectionist modelling. Bass and Wind [1] cite the confidentiality of the

data, resulting from the high business value of information in this area, as one of the major reasons for this fact.

This paper is organised as follows. In section 2, we provide a concise overview of response modelling issues in direct marketing and motivate the choices made in the research set-up. Section 3 discusses the empirical setting. In a first subsection the data set is briefly described. The second subsection highlights the initial neural network set-up. In a final subsection, we elaborate on the proposed feature selection method and discuss its application to the marketing case at hand.

2 Response Modelling in Direct Marketing

In this section, we will briefly elaborate on some response modelling issues typical to direct marketing. Hereby, we position both the nature of the problem statement and the chosen variables, in casu the RFM predictors and the response model.

Cullinan [5] is generally credited for identifying the three sets of variables most often used in database marketing modelling: (R)ecency, (F)requency and (M)onetary values [2,13]. Since then, the literature has accumulated so many uses of these three variables, that there is overwhelming evidence both from academically reviewed studies as well as from practitioners' experience that the RFM variables are the most important set of predictors for modelling mail-order repeat purchasing. However, when browsing the vast amount of literature, it becomes evident that only very limited attention has been devoted to selecting the right set of variables to include into the model of mail-order repeat buying. In fact, most studies do not offer a formal justification of their choice of variables, which is therefore often of an ad-hoc nature. Instead, the focus of these articles lies mostly on selecting the appropriate modelling technique.

We will investigate how a wrapped MLP-based feature selection method can assist in determining which of the suggested RFM variables (cf. subsection 3.1) may play a pivotal role in predicting repeat purchase behaviour by mail-order. The adoption of MLPs for modelling purposes is motivated by the fact that they are flexible, non-parametric modelling techniques allowing to perform any complex function mapping with arbitrarily desired accuracy [11].

For mail-order response modelling, several alternative problem formulations have been proposed based on the choice of the dependent

variable. The first category is purchase incidence modelling [3]. In this problem formulation, the main question is whether a customer will purchase during the next mailing period, i.e. one tries to predict the purchase incidence within a fixed time interval. Other authors have investigated related problems dealing with both the purchase incidence and the amount of purchase in a joint model [14,26]. A third alternative perspective for response modelling is to model interpurchase time through survival analysis or (split-)hazard rate models [6,25] which model whether a purchase takes place together with the duration of time until a purchase occurs.

This paper focuses on the first type of problem, i.e. purchase incidence modelling. More specifically, we consider the issue whether or not a purchase is made from any product category offered by the direct-mail company. This choice is motivated by the fact that the majority of previous research in the direct marketing literature focuses on the purchase incidence problem [19,27]. Furthermore, this is exactly the setting that mail-order companies are typically confronted with. They have to decide whether or not a specific offering will be sent to a (potential) customer during a certain mailing period.

3 Empirical set-up and Evaluation

3.1 Data set

From a major European mail-order company, we obtained Belgian data on past purchase behaviour at the order line level, i.e. we know when a customer purchased what quantity of a particular product at what price as part of what order. This allowed us to derive all RFM variables discussed below for a total sample size of 1200 customers of which 37.6 % represented buyers. As a form of pre-processing, the few missing values were handled by the unconditional mean imputation procedure [15]. The (R)ecency, (F)requency and (M)onetary variables have then been modelled as follows.

Recency is operationalised as the number of days since the last purchase [2]. An alternative operationalisation would be the number of consecutive mailings without response [4]. Additionally, we include the log transformation of the recency variable as an input to the MLPs. The choice of this transformation is motivated as a means to reduce the skewness of the distribution of the original recency variable.

Although in most studies no detailed results are reported, the frequency variable is generally considered to be the most important of the RFM predictors [19]. Frequency is usually operationalised as the number of purchases made in a certain time period [2,4]. When considering time interval length, inclusion versus exclusion of returned items and orderline versus order level processing, many combinations of these variables are possible. We decide to consider two levels for each factor, which results in a 2x2x2 design (8 operationalisations) as indicated in Figure 1. In the frequency column *Fr* refers to frequency, *Year* refers to the frequency during the last 12 months, *Hist* refers to the frequency during the whole customer history, *NoRet* refers to the fact that returns are deleted before processing, *Returns* refers to the fact that returns are also included in the count, *Orderlines* refers to the fact that the frequency reflects a count of the number of orderlines and *DiffOrders* refers to the fact that not orderlines, but rather the number of different dates (purchase occassion) on which orders are placed, are counted.

Recency	Frequency	Monetary
- *Recency*	- *FrYearNoRetOrderlines*	- *MonYearNoRet*
- *Log(Recency)*	- *FrYearNoRetDiffOrders*	- *MonHistNoRet*
	- *FrYearReturnsOrderlines*	- *MonMaxNoRet*
	- *FrYearReturnsDiffOrders*	- *MonAvgNoRet*
	- *FrHistNoRetOrderlines*	- *Log(MonYearNoRet)*
	- *FrHistNoRetDiffOrders*	- *Log(MonHistNoRet)*
	- *FrHistReturnsOrderlines*	- *Log(MonMaxNoRet)*
	- *FrHistReturnsDiffOrders*	- *Log(MonAvgNoRet)*

Figure 1: RFM variables included in data set.

Monetary value can either be operationalised as the total accumulated monetary amount of spending by a customer during a certain amount of time [5], as the highest transaction sale or as the average order size [19]. In the monetary column of figure 1, *Mon* refers to monetary value, *Year* refers to the monetary value during the last 12 months, *Hist* refers to the monetary value during the whole customer history, *Max* refers to the highest transaction sale over the whole customer history, *Avg* refers to the average transaction order size over the whole customer history and *NoRet* refers to the fact that returns are deleted before processing. Again, the log transformation is used to reduce skewness.

3.2 Initial experimental set-up

As a pre-processing stage to the induction algorithm, all 18 features are statistically normalised to a mean of 0 and a standard deviation of 1. All MLPs in this study have one hidden layer and 3 hidden units, influenced by theoretical works, which show that a single hidden layer is sufficient to approximate any complex non-linear function with any desired degree of accuracy [11]. Both hidden and output units use logistic activation functions.

The MLPs are trained using Bayesian regularisation for maximal 1000 epochs. This technique is related to a Levenberg-Marquardt optimisation, using Bayesian theory for finding optimal training parameters [7]. In this way, pattern recognition can be handled without too many problems of parameterisation, typical for nonlinear optimization. Setting the weights w and V and the biases b as the unknown vector θ then the objective function

$$F(\theta) = \beta E_D + \alpha E_W$$

is minimised using however a Bayesian formulation to determine optimal parameters α and β [16]. Since Bayesian regularisation inherently counters the effect of overfitting, it avoids the need for a separate validation set to implement early stopping [16].

A 10-fold cross-validation procedure is used in which the data set is split into 10 mutually exclusive folds of equal size. Hence, 10 MLPs are trained using only 9 folds of the data (training folds) and tested on the remaining fold (validation fold). As a result, all observations are used for training and each observation is used exactly once for testing. Performance is then computed by averaging the classification accuracy over all 10 validation folds.

Classification accuracy is measured by means of the Percentage Correctly Classified (PCC) and the Area under the Receiver Operating Curve (AUC). For the PCC, a cut-off value of 0.5 is used. Because an economically optimal cut-off value might be prefered, the AUC is used as an additional criterion since it is not dependent upon the specific cut-off value [22].

Table 1 illustrates the results of the above procedure for the model with all inputs (full model).

Table 1: Mean results for full model

Mean Results	Full Model
PCC_{train}	0.7403
PCC_{val}	0.7133
AUC_{train}	0.7814
AUC_{val}	0.7372
Features	18

3.3 Wrapped feature selection

Feature selection is now implemented using a typical wrapper-approach with a best-first search heuristic guiding the backward search procedure towards the optimal feature set [12]. Starting with the full feature set, all inputs are pruned sequentially, i.e. one by one.

Following Moody et al. [17,18], we perturb each input feature to its mean and compute the impact on the network output by means of the mean squared error on the training data, giving rise to the following Sensitivity Index (SI_i):

$$SI_i = MSE(\bar{x}_i, w) - MSE(x_i, w),$$

where w represents the trained weight vector, x_i the i^{th} feature and $\bar{x}_i$ its average over all data points.

The SI_i are then aggregated over all 10 training folds and the feature with the lowest aggregated sensitivity index is then removed. This amounts to a strategy of constant substitution, treating the input to be neglected as missing by substituting its effect to its mean over the whole sample [23]. Notice that no retraining is needed while computing these sensitivities.

Note that the concept of sensitivity of the model to the presence/absence of a feature, as defined by the above sensitivity measure, does not completely correspond to the concept of causal relevance of a feature within the real, but unknown, functional relationship. Interaction and correlation effects among features tend to obscure a rightful assessment of the causal relevance of a feature.

However, the suggested approach proves to be fairly robust with respect to interaction and correlation. As to the presence of interaction effects, suppose two variables are interacting in a significant fashion. Setting one variable to its mean will destroy the interaction and consequently degrade the network performance, resulting in a large value for the sensitivity index SI_i.

As to the presence of correlation effects, consider for instance the extreme situation in which two variables are perfectly correlated and at least one of them has an inherent significant causal contribution within the functional relationship to be learned. At first sight, the network will seem individually insensitive to either one of these variables since holding either one of them to a constant value loses no information. However, after having removed the first one, the relevance of the other feature increases dramatically. For that reason, re-computing the sensitivity indices in the next step will provide evidence on the significance of the remaining feature. Figure 2 gives an indication of the classification accuracy of the feature selection method as measured by the PCC and AUC on training and validation set at each step of the pruning procedure. Note that all features are pruned sequentially. This gives rise to 19 pruning steps with the last corresponding to an MLP having only the bias term as its input, yielding a classification accuracy of approximately 62.4 % (majority prediction).

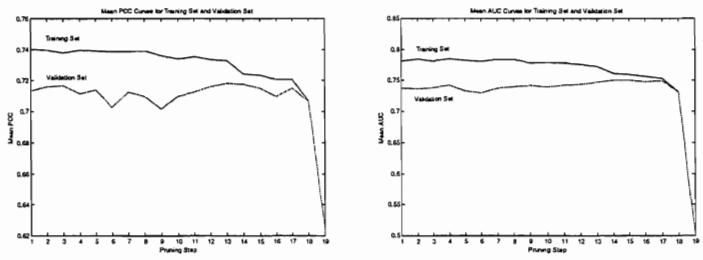

Figure 2: Mean PCC and AUC Curves for Training and Validation Set

Table 2 indicates the order in which the features are pruned using the above discussed sensitivity based method.

Table 2: Order of feature removal

Pruning Step	Feature	Pruning Step	Feature
1	FrHistNoRetOrderlines	10	FrYearNoRetDiffOrders
2	Recency	11	Log(MonAvgNoRet)
3	FrHistReturnsDiffOrders	12	Log(MonYearNoRet)
4	FrHistReturnsOrderlines	13	*MonAvgNoRet*
5	Log(MonHistNoRet)	14	*Log(Recency)*
6	MonYearNoRet	15	*FrYearReturnsOrderlines*
7	Log(MonMaxNoRet)	16	*FrYearReturnsDiffOrders*
8	MonMaxNoRet	17	*FrYearNoRetOrderlines*
9	MonHistNoRet	18	*FrHistNoRetDiffOrders*

The question naturally arises where to situate the cut-off in order to determine the optimal feature subset. In doing so, a trade-off must be evaluated between model complexity and model accuracy also known as the bias/variance trade-off [8,9]. Several criteria have been devised to effectively cope with this trade-off, e.g. Network Information Criterion, Akaike Information Criterion. In this paper, we will determine the cut-off point by means of an Analysis of Variance (ANOVA) approach. The procedure is fairly straightforward. We start by identifying the top of the mean PCC curve on the training set. Naïve reasoning would then go for the feature set at this point as the optimal cut-off. For the RFM case, this would lay the cut-off at pruning step 1. The resulting feature set would then consist of all features. However, the cut-off decision would then be purely based on a mean performance criterion evaluated on the training set. In order to take into account the beneficial effect of reduced model complexity (cf. bias/variance trade-off), we proceed with a sequence of ANOVA tests.

In subsequent steps, we proceed along the mean PCC_{train} curve, starting at pruning step 1 (i.e. maximum mean PCC_{train}) and perform a one-way ANOVA analysis to determine the point at which the mean PCC_{train} value decreases significantly (5% significance level) vis-a-vis the starting point i.e. pruning step 1. This procedure allows to take into account the variance of the PCC_{train} values over all 10 cross-validation runs.

Table 3 presents the results of applying the feature selection procedure described above.

Table 3: Mean results for reduced model.

Mean Results	Reduced Model
PCC_{train}	0,7330
PCC_{val}	0,7183
AUC_{train}	0,7725
AUC_{val}	0,7468
Features	6

Observe from Table 3 how the suggested feature selection method allows to significantly reduce the model complexity (from 18 to 6 variables) without degrading the generalisation behaviour on the validation set (from 0.7133 to 0.7183 in terms of PCC).

Some interesting marketing conclusions can now be inferred. Among the 6 remaining features of the reduced model, predictors of all three feature categories (Recency, Frequency and Monetary) are encountered, suggesting that a combined use of all three variable categories yields the richest model for repeat purchase behaviour. However, it has to be remarked that a feature set consisting of only 4 features, in casu with only frequency variables, (pruning step 15 in Table 2) still yields a mean PCC_{val} of about 71.5 % and a mean AUC_{val} of about 75%. This clearly illustrates the importance of the frequency variables in predicting mail-order repeat-purchase behaviour. This piece of evidence supports the hypothesis that the frequency variable is to be considered the most important of the RFM predictors [19].

4 Conclusions

In this paper, we studied a wrapped neural network feature selection method in a direct marketing setting by means of an illuminating case study. The case involved the detection and qualification of the most relevant RFM (Recency, Frequency and Monetary) features. Results indicate that elimination of redundant/irrelevant features by means of the discussed feature selection approach allows to significantly reduce model complexity without degrading generalisation ability. It is precisely this element that allows to make some very interesting marketing inferences concerning the relative importance of the RFM predictor categories. The empirical findings highlight the importance of a combined use of all three variable categories in predicting mail-order repeat purchase behaviour. However, the results also illustrate the dominant role of the frequency variable. Even a model with only frequency variables still yields a satifying classification accuracy compared to the optimally reduced model.

Bibliography

[1] Bass, F.M. & Wind, J. Introduction to the special issue: empirical generalisations in marketing. *Marketing Science*, 14, G1-G6., 1995.

[2] Bauer, A. A direct mail customer purchase model. *Journal of Direct Marketing*, 2(3), pp. 16-24, 1988.

[3] Bult, J.R. *Target selection for direct marketing*. Phd. Dissertation, Groningen University, 1993.

[4] Bult, J.R. & Wansbeek T.J. Optimal selection for direct mail. *Marketing Science,* 14, pp. 378-394, 1995.

[5] Cullinan, G.J. *Picking them by their batting averages' recency-frequency-monetary method of controlling circulation*, Manual release 2103, Direct Mail/Marketing Association, N.Y., 1977.

[6] Dekimpe, M. G. & Degraeve Z. The attrition of volunteers, *European Journal of Operation Research*, 98, pp. 37-51, 1997.

[7] Foresee, D.F. & Hagan, M.T. Gauss-Newton approximation to bayesian learning. *International Conference on Neural Networks*, 3, pp.1930-1935, 1997.

[8] Friedman, J. On Bias, Variance, 0/1 Loss, and the Curse of Dimensionality, *Data Mining and Knowledge Discovery,* 1, pp.55-77, 1997.

[9] Geman, S., Bienenstock, E. & Doursat, R. Neural Networks and the bias/variance dilemma. *Neural Computation*, 4, pp. 1-58, 1992.

[10] Grant, A.W.H. & Schlesinger L.A. Realize Your Customers' Full Profit Potential. *Harvard Business Review*, Sept-Oct, pp. 59-72, 1995.

[11] Hornik, K., Stinchcombe, M. & White, H. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2, pp. 359-366, 1989.

[12] John, G.W., Kohavi, R. & Pfleger, K. Irrelevant Features and the Subset Selection Problem. *Machine Learning: Proceedings of te Eleventh International Conference*, Morgan Kaufmann Publishers, San Francisco CA, pp. 121-129, 1994.

[13] Kestnbaum, R.D. Quantitative database methods. *The Direct Marketing Handbook*, Nash E.L., pp. 588-597, 1992.

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

[15] Little, R.J.A. Regression with missing x's: a review, *Journal of the American Statistical Association*, 87(420), pp. 1227-1237, 1992

[16] MacKay, D.J.C. Bayesian interpolation. *Neural Computation*, 4, pp. 415-447, 1992.

[17] Moody, J. Note on Generalisation, Regularization and Architecture Selection in Nonlinear Learning Systems. *First IEEE-SP Workshop on Neural Networks for Signal Processing*, IEEE Computer Society Press, Los Alamitos, CA, pp. 1-10, 1991.

[18] Moody, J. & Utans, J. Principled Architecture Selection for Neural Networks: Application to Corporate Bond Rating Prediction. *NIPS4*, pp. 683-690, 1992.

[19] Nash, E.L. *Direct marketing: strategy, planning, execution*, 3rd edition, McGraw-Hill, New York, 1994.

[20] Reicheld, F.F. & Sasser, W.E. Zero Defections: Quality comes to Service. *Harvard Business Review*, Sept-Oct, pp. 301-307, 1990.

[21] Stone, M. & Woodcock, N. *Relationship Marketing*, Kogan Page, 1996.

[22] Swets, J.A. Indices of discrimination or diagnostic accuracy. Their ROCs and implied models. *Psychological Bulletin*, 99(1), pp. 100-117, 1986.

[23] Van De Laar, P., Heskens, T. & Gielen, S. Partial Retraining: A New Approach to Input Relevance Determination. *International Journal of Neural Systems*, 9(1), pp. 75-85, 1999.

[24] Van den Poel, D. *Response Modeling for Database Marketing using Binary Classification.* Phd. Dissertation, K.U. Leuven, Faculty of Economic and Applied Economic Sciences, 1999.

[25] Van den Poel, D. & Leunis, J. Database marketing modeling for financial services using hazard rate models. *International Review of Retail, Distribution and Consumer Research*, 8 (2), pp. 243-257, 1998.

[26] Van der Scheer, H. R. *Quantitative approaches for profit maximization in direct marketing*. Phd. Dissertation, Rijksuniversiteit Groningen, 1998.

[27] Zahavi, J. & Levin, N. Issues and problems in applying neural computing to target marketing. *Journal of Direct Marketing*, 11(4), pp. 63-75, 1997.