DEAL EFFECT CURVE AND PROMOTIONAL MODELS
Using Machine Learning and Bootstrap Resampling Test

Cristina Soguero-Ruiz[1], Francisco Javier Gimeno-Blanes[2], Inmaculada Mora-Jiménez[1], María Pilar Martínez-Ruiz[3] and José Luis Rojo-Álvarez[1]

[1]*Dep. of Signal Theory and Communications, University Rey Juan Carlos, Madrid, Spain*

[2]*Dep. of Signal Theory and Communications, University Miguel Hernández, Elche, Spain*

[3]*Dep. of Marketing, University Castilla La Mancha, Cuenca, Spain*

Keywords: Deal effect curve, Marketing, Multilayer perceptron, Bootstrap resampling, Promotional models.

Abstract: Promotional sales have become in recent years a paramount issue in the marketing strategies of many companies, specially in the current economic situation. Empirical models of consumer promotional behavior, mostly based on machine learning methods, are becoming more usual than theoretical models, given the complexity of the promotional interactions and the availability of electronic recordings. However, the performance description and comparison among promotion models are usually made in terms of absolute and empirical values, which is a limited handling of the information. Here we first propose to use a simple nonparametric statistical tool, the *paired bootstrap resampling*, for establishing clear cut-off test based comparisons among methods for machine learning based promotional models, by simply taking into account the estimated statistical distribution of the actual risk. The method is used to determine the existence of actual statistically significant differences in the performance of different machine design issues for multilayer perceptron based marketing models, in a real database of everyday goods (milk products). Our results show that paired bootstrap resampling is a simple and effective procedure for promotional modeling using machine learning techniques.

1 INTRODUCTION

In the present economic landscape, given by economic instability and changes in the acquisition behavior of consumers, food retailers have turned to the modification of the conventional commercial decisions implemented in their shops (Haluk and Özgül, 2007). Therefore, sales promotion have become in recent years a fundamental tool for retailers' strategies, and the investment in this setting has highly increased in the marketing strategy, with percentage even above 50% (Blattberg and Neslin, 1990). Better understanding of the sales promotion dynamics has received growing attention from machine learning and data mining techniques, which are powerful tools to extract information from examples in past quantitative experience (Leeflang and Wittingk, 2000).

However, operational problems can arise in machine learning promotional modeling, when based on nonlinear estimation techniques, for evaluating and demonstrating working hypothesis (Heerde et al., 2001; Liu et al., 2004; Martínez-Ruiz et al., 2006; Martínez-Ruiz et al., 2006; Wang et al., 2008). First, conventional parametric tests are often not appropriate, because given the heavy tails and heteroscedasticity for the prediction residuals, Gaussianity is no longer a working property for them. Second, special attention has to be paid in order to be sure to be working with hypothesis tests in terms of actual risk comparisons, and not of empirical risk comparisons, to avoid as much as possible the unaware presence of overfitting in the machine learning based models. And third, as an indirect consequence of not having a clear cut-off test, their results cannot always be easily compared across studies, even when they have been made on the same data set.

Therefore, the objective of this work was to propose an operative procedure for model diagnosis in the context of using machine learning techniques for promotional efficiency applications. We used an empirical approach, based on machine learning techniques for analyzing the sales dynamics in a specific product, namely, milk, which is a everyday consumer products, and we analyzed its response to promotional

Soguero-Ruiz C., Javier Gimeno-Blanes F., Mora-Jiménez I., Pilar Martínez-Ruiz M. and Luis Rojo-Álvarez J. (2012).
DEAL EFFECT CURVE AND PROMOTIONAL MODELS - Using Machine Learning and Bootstrap Resampling Test.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 537-540
DOI: 10.5220/0003732705370540

discounts in a retailer environment by means of Multi Layer Perceptron (MLP) neural networks.

2 TEST FOR ACTUAL RISK

Artificial Neural Networks (ANN) are multiparametric nonlinear models, and they are capable of learning from samples and discovering complex relationships among variables which can be hidden in the data volume available in the training set (Bishop, 1995). In the MLP network, one of the mostly used ANN, there is one input neuron for each variable in input pattern $\mathbf{x}$, and as many output neurons as output variables to be estimated (i.e., y can be a multivariable output), hence the number of hidden layers and the number of neurons in each have to be chosen during the design process.

Several merit figures can be used for benchmarking estimation models with machine learning techniques. On the one hand, absolute merit figures give an idea of the actual magnitude of the averaged error, being usual being the Mean Absolute Error (MAE), given by $MAE = \frac{1}{N}\sum_{i=1}^{N}|f(\mathbf{x}_i) - y_i|$ whereas relative merit figures can give a better idea of the amount of variability in the data which has ben actually captured by the model. In this work, a cross-validation technique is used for benchmarking and comparing several model architectures (Haykin, 1999).

A *bootstrap resample* is a data subset is drawn from the observation set according to their empirical *pdf* $\hat{p}_{y,\mathbf{x}}(\mathbf{x},y)$. Hence, the true *pdf* is approximated by the empirical *pdf* estimated from the observations, and the bootstrap resample can be seen as a sampling with replacement process of the observed data, this is, $\hat{p}_{y,\mathbf{x}}(\mathbf{x},y) \mapsto \mathbf{V}^* = \{(\mathbf{x}_i^*, y_i^*); i = 1,...,N\}$, where superscript * represents, in general, any observation, functional, or estimator, that arises from the bootstrap resampling process. A partition of $\mathbf{V}^*$ in terms of resample $\mathbf{V}^*(b)$ is given by $\mathbf{V} = \{\mathbf{V}_{in}^*(b), \mathbf{V}_{out}^*(b)\}$, where $\mathbf{V}_{in}^*(b)$ is the subset of observations that are included in resample b, and $\mathbf{V}_{out}^*(b)$ is the subset of non included observations. A *bootstrap replication* of an estimator is given by its calculation constrained to the observations included in the bootstrap resample. The bootstrap replication of the empirical risk estimator given by a calculation operator t and its weights w is $\hat{R}_{emp}^*(b) = t(\{w\}, \mathbf{V}_{in}^*(b))$. The scaled histogram obtained from B resamples is an approximation to the true *pfd* of the empirical risk. However, further advantage can be obtained by calculating the bootstrap replication of the risk estimator on the non included observations. By doing so, rather than estimating the empirical risk, we are in fact obtaining the replication of the *actual risk*, i.e., $\hat{R}_{act}^*(b) = t(\{w\}, \mathbf{V}_{out}^*(b))$. The bootstrap replication of the averaged actual risk can be obtained by just taking the average of $\hat{R}_{act}^*(b)$ for $b = 1,\ldots,B$. Moreover, the replications of the *pdf* of the model merit figures can provide confidence intervals (CI) for the performance. A typical range for B in practical applications can be from 100 to 2000 bootstrap resamples.

For giving a clear cut-off test allowing us to benchmark the significance of the observed differences between the performance of two different machine learning based promotional models, we use here the bootstrap nonparametric resampling procedure. The use of bootstrap resampling is supported by the previous observation of heavy tails in the residual distribution when using this kind of models, as well as bimodalities, and other non-Gaussian effects (Efron and Tibshirani, 1997). The procedure can be readily adapted in order to benchmark the performance of two different machine learning techniques (or a given algorithm with different settings), by using a *paired* bootstrap resampling, in which the same resamples are to be considered in the benchmarked models. In this work, when resampling two different models $model_1$ and $model_2$, results have been compared according to three different statistics, namely,

$$\Delta MAE = MAE(model_1) - MAE(model_2) \quad (1)$$

$$\Delta CI = \Delta CI(model_1) - \Delta CI(model_2) \quad (2)$$

$$\Delta CI_{sup} = CI_{sup}(model_1) - CI_{sup}(model_2) \quad (3)$$

where CI has been obtained for a 95% level, and $\Delta CI(model_i) = CI_{sup}(model_i) - CI_{inf}(model_i)$. These statistics provide a description in terms of the average magnitude of the error, but also in terms of its scatter. In general, given that it is complicated to obtain closed forms for CI of scatter measurements, bootstrap resampling represents a useful approximation for making it possible.

3 EXPERIMENTAL RESULTS

We used a real database with a everyday consumer products corresponding to the milk product category. Specifically 6 products were analyzed corresponding to 6 different promotional models, as indicated in Table 1. The number of daily sold units were available in the same retailer (supermarket) during one year, excluding the weekends. Up to 304 examples (patterns) were available for each category, corresponding to the days when transactions were recorded in the supermarket. They were aggregated into 43 weeks, in order not to have a too strong effect of weekly seasonality in this database. In all models, there are three kinds

of input variable: price indices $\mathbf{x}^M$, direct discount x^d and baseline BL. Regarding the price indices variable they are defined as $PI(i,t) = \frac{P_{prom}(i,t)}{P_{reg}(i,t)}$, where $P_{reg}(i,t)$ and $P_{prom}(i,t)$ are the regular and promotional price, respectively, of product i at week t. Hence, the price index gives the relative variation between the promotional price and the regular price, and its value is 1 whenever both are equal.

The input pattern was given by concatenation of the three kind of input variables, $\mathbf{x} = [\mathbf{x}^M, x^D, BL]$. The output of each model is given by the sold units for that particular product. Hence, the promotional model can be expressed as $y(i,t) = f(\mathbf{x}^M(i,t), \mathbf{x}^D(i,t), BL(t))$, where $y(i,t)$ is the number of sold units for the $i-th$ product during week t; $\mathbf{x}^M(i,t) = [PI^1(i,t), ..., PI^{n_m}(i,t)]$ is a vector with the price indices of product i during week t, with n_m = 6; $\mathbf{x}^D(i,t)$ is the direct discount dichotomous variable for product i during week t; and $BL(t)$ is the baseline variable at week t. The input metric variables were the same for all the models.

We considered the decision of two different design criteria for MLP-based promotional models. On the one hand, the MLP for estimation problems can give a multiple output, which in principle could benefit from the consideration of joint cross-information among models. However, there is no warranty that a multiple output architecture will work better than a separate MLP single model for each product. On the other hand, another design criterion is the use of different activation functions in the hidden layer nodes, being two widely used forms the linear and the sigmoid logistic activation. Given that there is no theoretical result, this function has to be chosen for each data mining model.

For this purpose, free parameters were tuned in the MLP for the single output set of models using LOO. Then paired bootstrap test was used to check which architecture can be pointed as more convenient, in terms of the previously used merit figures. Table 1 shows the comparison of the number of neurons in the hidden layer (n_0) for both architectures. Note that n_0 has a relevant variation in terms of different products

Table 1: Free parameter tuning in terms of n_0 for MLP with multiple and with single output, for Milk Data Base.

	n_o (MLP)	n_o(MLP_ind)
Model 1 (Asturiana)	17	15
Model 2 (Ato)	17	8
Model 3 (House brand)	17	1
Model 4 (Pascual Calcio)	17	6
Model 5 (Pascual Clasica)	17	8
Model 6 (Puleva Calcio)	17	14

Table 2: Single vs multiple output MLP for Milk products, using MAE merit figure. See text for details.

	MLP_{ind}	MLP	MLP_{ind} vs MLP
Model 1	357.3 ‖ 359.1 [266.2,463.2]	**320.7 ‖ 321.6** **[248.5,401.5]**	66.4 [-193.9,577.1] 37.0 [-50.8,130.3] 65.4 [-192.1,569.9]
Model 2	222.6 ‖ 222.3 [180.8,267.2]	**199.6 ‖ 200.4** **[157.1,247.9]**	26.2 [-114.1,192.1] 28.6 [-17.9,72.9] 50.2 [-69.4,217.6]
Model 3	**135.8 ‖ 136.3** **[105.5,167.3]**	152.5 ‖ 152.5 [119.2,188.6]	-66.9 [-144.6,49.3] -14.4 [-36.0,7.8] -57.5 [-124.6,62.0]
Model 4	**59.6 ‖ 59.7** **[44.2,77.4]**	72.3 ‖ 72.4 [58.5,88.3]	-7.8 [-49.2,18.7] -12.8 [-27.2,1.4] -13.4 [-54.5,7.0]
Model 5	305.2 ‖ 310.8 [227.9,397.5]	**198.7 ‖ 197.8** **[148.5,250.7]**	**258.8 [63.9,528.1]** **103.3 [21.0,186.2]** **267.3 [74.9,544.1]**
Model 6	226.7 ‖ 226.7 [167.3,293.2]	**125.6 ‖ 125.0** **[97.8,155.3]**	**304.5 [98.2,447.1]** **100.7 [37.8,163.1]** **304.5 [108.6,473.2]**

with single output, and also, that n_0 is sensibly larger for multiple output MLP architecture.

Table 2 shows the MAE and the comparison among both schemes, with the sigmoid activation function, for all the products in the data base. Individual models for each product with single output is denoted as MLP_{ind}, whereas multiple output architecture is denoted as MLP. Each cell in the second and third columns contains the empirically estimated actual risk (i.e., averaged from LOO estimation of MAE for each case), together with the bootstrap estimate of the averaged MAE, namely, the mean (upper line, right), and the 95% CI of this sample mean. The apparently best model of both, in terms of empirical LOO-MAE, is highlighted in bold. The comparison between both models is represented in the last column, showing the average and the 95% CI for ΔMAE, ΔCI, and ΔCI_{sup}, in the first, second, and third line of the cell, respectively. In this column, bold is used for highlighting the CI which yield significant differences with respect to the paired bootstrap test, i.e., those statistics for the differential merit figure whose estimated difference does not overlap the zero level.

It can be observed that, for Models 3 and 4, the performance is apparently better when using individual architectures, whereas Models 1, 2, 5, and 6 are better when considering the joint architecture. However, only significant differences are present in Models 5 and 6, both in terms of averaged and scatter MAE, hence the most advantageous situation is to use a multiple output architecture. No significant differences are sustained by the paired bootstrap test for Models 1, 2, 3 and 4. In general terms, we can conclude that, for this Milk Data Base, it is better to con-

Table 3: Sigmoid vs linear MLP for Milk products, using MAE merit figure. See text for details.

	Sigmoid	Linear	Sigmoid vs Linear
Model 1	358.2 || 368.3 [258.0,461.3]	**325.7 || 326.3** **[252.4,414.2]**	31.5 [-66.7,138.9] 140.3 [-128.9,583.3] 141.0 [-118.1,584.0]
Model 2	235.1 || 235.5 [195.2,281.7]	**203.8 || 205.6** **[156.8,256.2]**	32.0 [-4.6,67.5] -36.56 [-219.9,85.8] -25.6 [-185.5,91.3]
Model 3	143.3 || 148.6 [111.9,174.9]	**137.7 || 140.7** **[106.4,171.1]**	6.1 [-14.1,29.1] 2.24 [-50.5,84.3] 12.5 [-36.6,87.0]
Model 4	**56.9 || 58.4** **[40.36,78.2]**	64.7 || 66.4 [48.63,82.6]	-8.9 [-18.8,2.0] 0.79 [-28.6,56.7] 2.4 [-29.0,54.3]
Model 5	347.8 || 360.1 [271.9,426.3]	**191.9 || 190.2** **[154.7,231.2]**	**153.4 [68.1,242.2]** **372.8 [121.1,555.6]** **377.8 [116.5,549.5]**
Model 6	245.5 || 255.2 [190.7,302.3]	**116.3 || 116.1** **[90.4,142.9]**	**127.1 [69.2,187.4]** **335.5 [134.8,445.8]** **346.8 [164.4,465.1]**

sider multiple output MLP architecture in those products showing a more stable behavior, and hence, results can be strongly dependent on the specific product.

For the case of MLP with single output, the increased performance due to the use of linear or sigmoid logistic activation function in the hidden layer nodes, was further analyzed. Table 3 shows the same information than in the previous analysis, for the paired comparison of both models in each product, after having fixed the individual output architecture design parameter, according to the previous result. In this case, the empirical MAE is lower when using linear activation functions in all models except Model 4. Significant differences are supported by the bootstrap test only in Models 5 and 6, but with strong consistency for all the statistics evaluated for the merit figure distribution. In general terms, it can be concluded that using linear activation function in the hidden nodes yields better results consistently in the Milk Data Base analyzed here.

4 CONCLUSIONS

In the present work, a method for giving a systematic statistical comparison between two machine-learning based models in promotional sales modeling has been presented. The method has its bases on the principles of the estimation of statistical descriptors of actual risk and their *pdf*, by means of bootstrap resampling, and on the use of the increment in the merit figure. The consideration of paired differences gives a suitable approach for controlling the standard error of the statistical description of the merit figures, hence allowing clear cut-off tests.

ACKNOWLEDGEMENTS

This work was supported by Research Project from Fundación Ramón Areces.

REFERENCES

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford University Press, Oxford.

Blattberg, R. and Neslin, A. (1990). *Sales Promotion: Concepts, Methods and Strategies*. Prentice-Hall.

Efron, B. and Tibshirani, R. J. (1997). *An introduction to the bootstrap*. Chapman & Hall, New York.

Haluk, M. and Özgül, E. (2007). The relationship between marketing strategies and performance in a economic crisis. *Journal of Marketing Practice: Applied Marketing Science*, 25(4):326–342.

Haykin, S. (1999). *Neural networks. A comprehensive foundation*. Presntice-Hall, New Jersey.

Heerde, H. J. V., Leeflang, P. S. H., and Wittink, D. R. (2001). Semiparametric analysis to estimate the deal effect curve. *Journal of Marketing Research*, 38(2):197–215.

Leeflang, P. S. H. and Wittingk, D. (2000). Building models for marketing decisions: past, present and future. *International Journal of Resarch in Marketing*, 17(2-3):178–185.

Liu, B. H., Kong, F. S., and Yang, B. (2004). PEPP: Profits Estimation in Prices Promotion. In *Proc of 2004 Intern Conf Mach Learn Cybern*, volume 2, pages 1146–1151, Australia.

Martínez-Ruiz, M. P., Mollá-Descals, A., Gómez-Borja, M. A., and Rojo-Álvarez, J. L. (2006). Evaluating temporary retail price discounts using semiparametric regression. *Journal of Product & Brand Management*, 15(1):73–80.

Martínez-Ruiz, M. P., Mollá-Descals, A., and Rojo-Álvarez, J. L. (2006). Using daily store-level data to understand price promotion effects in a semiparametric regression model. *Retailing and Consumer Services*, 3(13):193–204.

Wang, T., Li, Y. Q., and Zhao, S. F. (2008). Application of SVM Based on Rough Set in Real Estate Prices Prediction. In *4th Int Conf on Networking, Comm and Mob Comp*, pages 1–4, Dalian.