© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

A Statistical model to predict gift patterns in large distribution (GDO), industry and trade promotions.

F. Cornalba[1] , P. Giudici[1] , A Lazzarini[2], P. Rebecchi[2]
[1] *University of Pavia, Italy*
[2] *Coro Marketing Activities s.r.l.,Piacenza, Italy*

Abstract

We consider a problem about the way of predicting the exit of gifts in a promotional catalogue for a chain of Hypermarkets located in Northern Italy.
The aim of this paper is to compare experimental and statistical methods to achieve a reliable predictive model.

We realize a predictive statistical model using microsoft Excel. In this way we can obtain a data processing-statistical instrument that allow us to join the valuation of data with the commercial needs of the large-scale retail trade (GDO). We underline the role of performance evaluation data using methods for resolving problems.

1 Introduction

We briefly explain the results of the scientific research that Coro Marketing Activities in Piacenza (Coro deals with the organization and management of fidelity programs related to large distribution (GDO), Industry and Trade) has developed with the Department of Economics and Quantitative Methods of the University of Pavia. For the research a particular convention has been defined and 6 months internship has been put in place nearby the Coro's head-office.
The main aim of the work refers to the definition of a statistical model for the forecasting of outgoing premiums in relation to promotional activities (e.g points collection etc.).

The Company has decided to realise such model using the Excel Software, because it is a very well-known program and additionally it allows to obtain

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

highly technological reporting systems for the elaboration of economical, financial, and technical data (Data Mining).

In our view Data Mining is a process suitable for discovering new and significant correlations, relations, tecniques, "mining" large quantities of data, using tailored tecniques for the identification of useful statistical and mathematical relations.

The necessity to improve the knowledge in this field is linked to the willing of offering a scientific support for an analysis substantially based on experience. Consequently the aim has been to reduce the expected error in the out going premium process inside the fidelity programs and then to limit the risk of both "stock breaks" or excessive not-sold goods in storage, through the application of a statistical model.

The model examines the concepts and tecniques related to the analysis of high number of data. In particular a "Linear Regression Model" is used as an useful instrument to make predictions.

2 Linear regression model

The main goal of the forecasting model is to explain the existing relations between one or more dependent variables (number for outgoing premium) and one or more explicative variables (e.g required average expenses, binary variables...). This is done in order to develop a decision rule which may provide the values of the first variable in relation to the specific values of the latest.
In particular, a linear predictive model is defined in the data mining process on the basis of the following procedures (Giudici, 2001):

1) Choice and control of explanatory variables.

For an appropriate definition of such variables it is necessary to have on disposal a very large storical series of values. Through the analysis of the linear correlation coefficient, it is possible to identify a *correlation matrix* as a method of evaluation the existing correlations among all the possible pairs of variables.

2) Estimation of the parameters of the defined model and calculation of the regression line.

The linear function which better allows to interpolate the considered values is defined through parameters, which are based on historical series of data related to both th explanatory and the dependent variables

3) Dependent variable forecast.

This phase is based on the evolution of the estimated parameters and on the analysis of the obtained values, such as the Excel output, the linear determination coefficient , test t , standardized residuals. This allows to have a knowledge of the future values.

3 Case study

The statistical model, realised using Coro Marketing's historical data and other numerous bibliographic references, has been applied to a pilote project (Cornalba, 2002) .

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

Basically the project corresponds to a points collection lasted 47 weeks, including 4 weeks of clean –up , and to 57 premiums to be paid.
There are 15 explicative variables which have to be considered as significant:

- Required average expenditure for each premium;
- 8 binary variebles, which respectively identify the possible belonging of a product to a certain merchandise area;
- 6 quantitative variables which identify 6 periods of about 7 weeks (useful for the forecasting of the premium outgoing during the activities period, week by week).

On the other hand, the dependent variable is unique and is defined as the forecast of the outgoing premium.

4 Dataset's description

It is necessary, for analysing and building a model, to consult storical data. These data come from numerous references.

5 Applicative phases

There are 4 important phases for the application of the model:
1) Phases of the model for the forecast.
This phase is based on the application of the linear regression model and on the exploration of the data.
The phase of exploration consists in estimating the importance of gathered data, and this exploration preliminary suggests the existence of anomalous data (outliers).
It's very important the analysis of "goodness-of-fit of the regression" which is based on the study of the "linear determination coefficient" called R^2. R^2 is a value included between 0 and 1: as R^2 gets close to 1 , the model results more efficient.
2) Rilevation of the anomalous values: outliers.
It is essential to recognize the so-called outliers, to obtain a more reliable and significant forecast. Consequently the anomalous values must be removed again after doing careful explorative analysis, based on statistical methods and practical experience.
With respect to the below figure, we have defined as outliers values standardised not included in the "confidence interval" [- 1,96; +1,96] which are visible in the scatter plot.
3) Model without outliers.
After purging the model by removing anomalous observations, it is possible to obtain an excellent result, and then reduce the expected error in the outgoing premium process.
4) Interpretation of the results.
The aim of the pilote project has been reached. The margin of error amounts to about 50%; consequently the statistical model is very efficient to reduce the forecast of outgoing premiums.

© 2002 WIT Press, Ashurst Lodge, Southampton, SO40 7AA, UK. All rights reserved.
Web: www.witpress.com Email witpress@witpress.com
Paper from: *Data Mining III*, A Zanasi, CA Brebbia, NFF Ebecken & P Melli (Editors).
ISBN 1-85312-925-9

6 Conclusions

The linear regression model can be considered a very effective method because it respects the characteristics of each shop and of each premium. In fact with the application of statistical methods it is possible to obtain better results than with methods based only on the experience.

The identification of anomalous values, outliers, means that other explicative variables could influence the exit of gifts.

The suggested model could be improved, for instance using more explanatory variables or using other methods for the forecasting, for example neural networks, to capture possible non-linearities but in this case it will be necessary to use more sophisticated softwares.

References

[1] Barry J.M.A and Linof G., *Data mining Techniques for marketing, sales and customer support,* Wiley J.: New York, 1997.
[2] Bolognesi M. G., e Sfogliarini B., La statistica per le imprese, *Tendenze evolutive del sistema distributivo italiano e modificazioni del comportamento di acquisto,* SIS: Torino, pp. 289-302, 1997.
[3] Bonollo G., "*Applicazioni statistiche con Excel*", Franco Angeli., 2001.
[4] Cornalba F., *Modelli statistici per la previsione della distribuzione dei premi nella Grande Distribuzione Organizzata*, Tesi di Laurea Facoltà di Economia Università degli studi di Pavia., 2002.
[5] Giudici P., Tarantola C., Verrecchia F., "*Modelli statistici per le applicazioni di Data Mining*", La Eliotecnica Grafica Ed. Atti del Convegno del 25/10/01: Pavia, 2001.
[6] Giudici P., "*Data Mining*". McGraw-Hill, 2001
[7] Kotler P., *Marketing e management,* vol. 1 e 2, Isedi Petrini,: Torino, 1986.
[8] Paggi S., *Indici statistici di strategia promozionale: alcune proposte metodologiche*, Tesi di Laurea Facoltà di Economia Università degli studi di Pavia, 1999.
[9] Marbach G., *Le ricerche di mercato*, UTET: Torino, 1982.
[10] Meta Group, *Data mining market trends: 1997-1998*, Meta Group, 1998.