Predictive Models and Knowledge Management in e-Banking Data

Vasilis Aggelis
University of Patras
Department of Computer Engineering and Informatics
Rio, Patras, Greece
Vasilis.Aggelis@egnatiabank.gr

Dimitris Christodoulakis
University of Patras
Department of Computer Engineering and Informatics
Rio, Patras, Greece
dxri@cti.gr

ABSTRACT

Knowledge Management exercises significant influence in establishment and development of a company. A modern such approach is data mining. Data mining is the search for relationships and patterns that exist in data sets, but are "hidden" among the vast amounts of data. These relationships and patterns represent valuable knowledge about the data set. Through data mining methods, certain variables obtained from observation can be represented by means of various models like neural networks and decision trees. Many companies and organizations use nowadays such tools, that contribute to the more effective control and exploitation of their knowledge and information. In the present paper prediction models, a popular data mining method is studied, by the use of an example coming from the real world and specifically electronic banking. It is demonstrated that prediction models contribute to the more efficient knowledge management in electronic banking sector.

Keywords: Knowledge Management, Data Mining, Predictive Models, Linear Regression.

1. Introduction

Banking or financial data treatment is generally conducted using several data mining methods such as Linear Regression, Neural Networks and Decision Trees aiming at the development of patterns, rules, predictive models and finally forecasting. These methods produce interesting as well as useful results enhancing the more efficient knowledge management. However, not all kinds of results lead to rigid conclusions.

From this point of view the data miner and the judgment of the user are essential in evaluating the results and especially the predictive models efficiency. Therefore the co-operation between people expert in data mining and others with good knowledge of the data sets is important leading to proper evaluation of the predictive model. In the banking area this combination is definitely necessary due to the singularity of bank data as well as bank market rules.

A specific kind of bank service is the application of services through internet (e-banking). This alternative channel is relatively new making relevant feature extraction very important. Since future tendencies suggest the increase of its use, a bank should be naturally concerned with enlargement of its customer share in this specific area.

In this paper the development and evaluation of a predictive model for e-banking applications is studied. For the present case study the software SPSS Clementine 7.0 was used. A general description of predictive models follows in section 2, while in section 3 the procedure of predictive model development in e-banking is presented. Experimental results are discussed in section 4 while section 5 contains final conclusions and future work plans.

2. Predictive models basics

A model is an abstract representation of a real-word process. A typical form of a model is Y=aX+b, where Y, X are variables and a, b are parameters. In a predictive model [Foster, D., and Stine, R. (2002), Zupan, B., Demsar, J., Kattan, M., Ohori, M., Graefen, M., Bohanec, M., and Beck, J.R. (2001), Raftery, A., Madigan, D., and Hoeting, J. (1997), Laud, P., and Ibrahim, J. (1995)], one variable is expressed as a function of the others. This permits the value of the response variable to be predicted from given values of the others (the predictor variables). The response variable in general predictive models is often denoted by Y, and the p predictor variables by X_1, …, X_p. The model will yield predictions, $y^ = f(x_1, \ldots, x_p;\theta)$ where y^ is the prediction of the model and θ represents the parameters of the model structure. When Y is quantitative, this task of estimating a mapping from the p-dimensional X to Y is known as regression.

Prediction models [Hand, D., Mannila, H., and Smyth, P. (2001), Raftery, A., Madigan, D., and Hoeting, J. (1997), Draper, N.R., and Smith, H. (1998)] in which the response variable is a linear function of the predictor variables, yields prediction:

$$Y^ = a_0 + \sum_{j=1}^{p} a_j X_j \qquad (2.1)$$

Where $\theta = \{a_0, \ldots, a_p\}$. We have used Y^ rather than simply Y on the left of the expression because it is a model, which has been constructed from the data. In other words, the values of Y^ are values predicted from the X, and not values actually observed.

An easy way to evaluate and compare predictive models in order to choose the best is the use of evaluation charts. These charts show how models perform in predicting particular outcomes. They work by sorting records based on the predicted value and confidence of the prediction, splitting the records into groups of equal size (quantiles), and then plotting the value of the business criterion for each quantile, from highest to lowest.

Outcomes are handled by defining a specific value or range of values as a hit. Hits usually indicate success of some sort or an event of interest.

There are five types of evaluation charts [SPSS (2002), Hong, S.J., and Weiss, S. (2000)], each of which emphasizes a different evaluation criterion.

Gains Charts: Gains are defined as the proportion of total hits that occurs in each quantile. Gains are computed as (number of hits in quantile/total number of hits) X 100%.
Lift Charts: Lift compares the percentage of records in each quantile that are hits with the overall percentage of hits in the training data. It is computed as (hits in quantile/records in quantile) / (total hits/total records).
Response Charts: Response is simply the percentage of records in the quantile that are hits. Response is computed as (hits in quantile / records in quantile) x 100%.
Profit Charts: Profit equals the revenue for each record minus the cost for the record. Profits for a quantile are simply the sum of profits for all records in the quantile. Profits are assumed to apply only to hits, but costs apply to all records. Profits and costs can be fixed or can be defined by fields in the data. Profits are computed as (sum of revenue for records in quantile – sum of costs for records in quantile).
Return On Investment Charts: Return on investment is similar to profit in that it involves defining revenues and costs. Return on investment compares profit to costs for the quantile and is computed as (profits for quantile/costs for quantile) x 100%.

Evaluation charts can be also be cumulative, so that each point equals the value for the corresponding quantile plus all higher quantiles. Cumulative charts usually convey the overall performance of models better, whereas non-cumulative charts often excel at indicating particular problem areas for models.

The interpretation of an evaluation chart [Zupan, B., Demsar, J., Kattan, M., Ohori, M., Graefen, M., Bohanec, M., and Beck, J.R. (2001)] depends to a certain extent on the type of chart, but there are some characteristics common to all evaluation charts. For cumulative charts, higher lines indicate better models, especially on the left side of the chart. In many cases, when comparing multiple models the lines will cross, so that one model will be higher in one part of the chart and another will be higher in a different part of the chart. In this case, you need to consider what portion of the sample you want (which defines a point on the *x* axis) when deciding which model to choose.

Most of the non-cumulative charts [SPSS (2002)] will be very similar. For good models, noncumulative charts should be high toward the left side of the chart and low toward the right side of the chart. (If a non-cumulative chart shows a sawtooth pattern, you can smooth it out by reducing the number of quantiles to plot and reexecuting the graph.) Dips on the left side of the chart or spikes on the right side can indicate areas where the model is predicting poorly. A flat line across the whole graph indicates a model that essentially provides no information.

Gains charts: Cumulative gains charts always start at 0% and end at 100% as you go from left to right. For a good model, the gains chart will rise steeply toward 100% and then level off. A model that provides no information will follow the diagonal from lower left to upper right (shown in the chart if Include baseline is selected).

Lift charts: Cumulative lift charts tend to start above 1.0 and gradually descend until they reach 1.0 as you go from left to right. The right edge of the chart represents the entire data set, so the ratio of hits in cumulative quantiles to hits in data is 1.0. For a good model, lift should start well above 1.0 on the left, remain on a high plateau as you move to the right, and then trail off sharply toward 1.0 on the right side of the chart. For a model that provides no information, the line will hover around 1.0 for the entire graph. (If Include baseline is selected, a horizontal line at 1.0 is shown in the chart for reference.)

3. Generating predictions in e-banking data set

In the case of this study, as stated above, the objective is the production and test of predictions about the volume of e-banking transactions in relation to the active users. The term financial transactions stands for all payment orders or standing orders a user carries out, excluding transactions concerning information content like account balance, detailed account transactions or mini statement. The term «active» describes the user who currently makes use of the electronic services a bank offers. Active users are a subgroup of the enlisted users.

One day is defined as the time unit. The number of active users is the predictor variable (Count_Of_Active_Users) while the volume of financial transactions is assumed to be the response variable (Count_Of_Payments).

A sample of the above data set is shown in Table 1

Transaction Day	Count_Of_Active_Users	Count_Of_Payments
...	...	...
27/8/2002	99	228

28/8/2002	107	385
29/8/2002	181	915
30/8/2002	215	859
...	...	...

Table 1

The date range for which the data set is applicable counts from April 20th, 2001 until December 12th, 2002. Data set includes data only for active days (holydays and weekends not included), which means 387 occurrences.

In order to generate a prediction like (2.1) the stepwise linear regression [SPSS (2002), Madeira, S.A. (2002)] method was used.

The Stepwise method of field selection builds the equation in steps, as the name implies. The initial model is the simplest model possible, with no input fields in the equation. At each step, input fields that have not yet been added to the model are evaluated, and if the best of those input fields adds significantly to the predictive power of the model, it is added. In addition, input fields that are currently in the model are reevaluated to determine if any of them can be removed without significantly detracting from the model. If so, they are removed. Then the process is repeated, and other fields are added or removed. When no more fields can be added to improve the model, and no more can be removed without detracting from the model, the final model is generated.

4. Experimental Results

The stepwise method builds in two steps the following prediction (Figure 1).

Count_Of_Payments ^ = (3.403)* Count_Of_Active_Users – 28.52 **(4.1)**

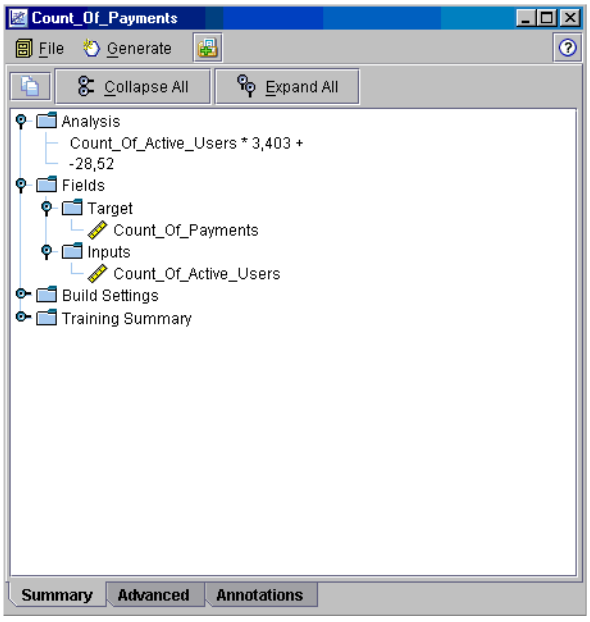

Figure 1

In order to evaluate and test the appropriateness of the model (4.1) evaluation charts were used along with some indicative measures such as R, R-square, Adjusted R-Square and Linear Correlation.

Examples of evaluation Charts (Figure 2 to Figure 3) are shown below.

Gains Chart: Chart shows that the gain rises steeply towards 100% and then levels off. Using the prediction of the model, calculate the percentage of Count_Of_Payments for the percentile and map these points to create the lift curve. An important notice is the greater the area is between the lift curve and the baseline, model is better.

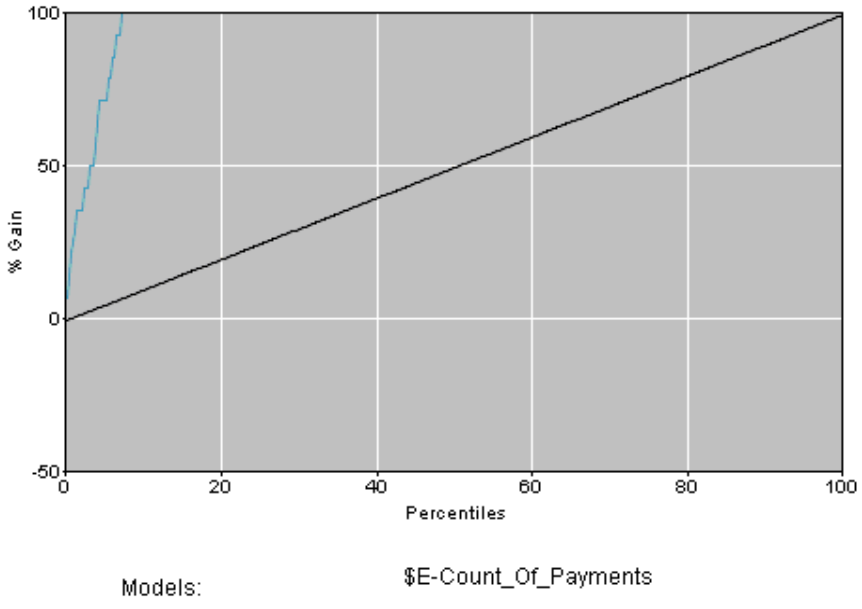

Figure 2

Lift Chart: As can be seen, Chart starts well above 1.0 on the left, remains on a high plateau as we move to the right, and then trails off sharply towards 1.0 on the right side of the chart. Using the prediction of the model shows the actual lift.

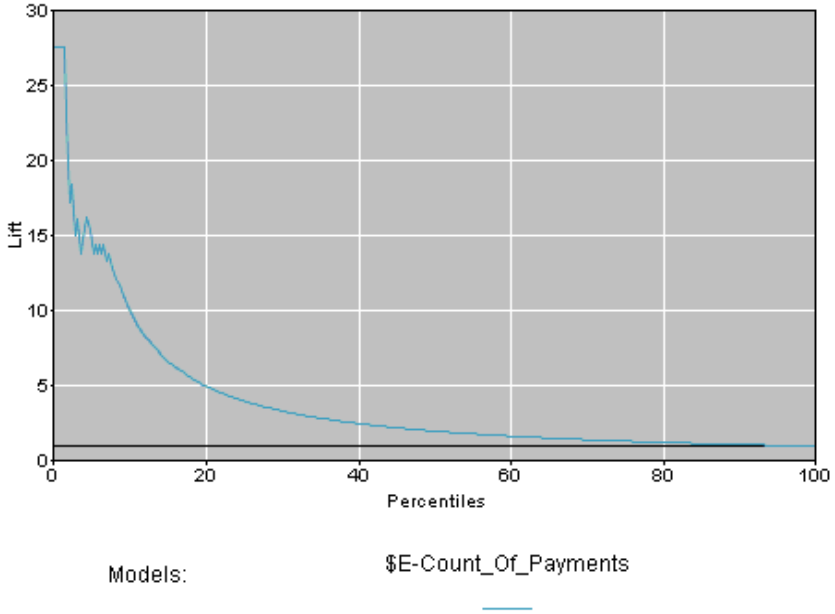

Figure 3

Other measures of the suitability of the models are supplied in Figure 4.

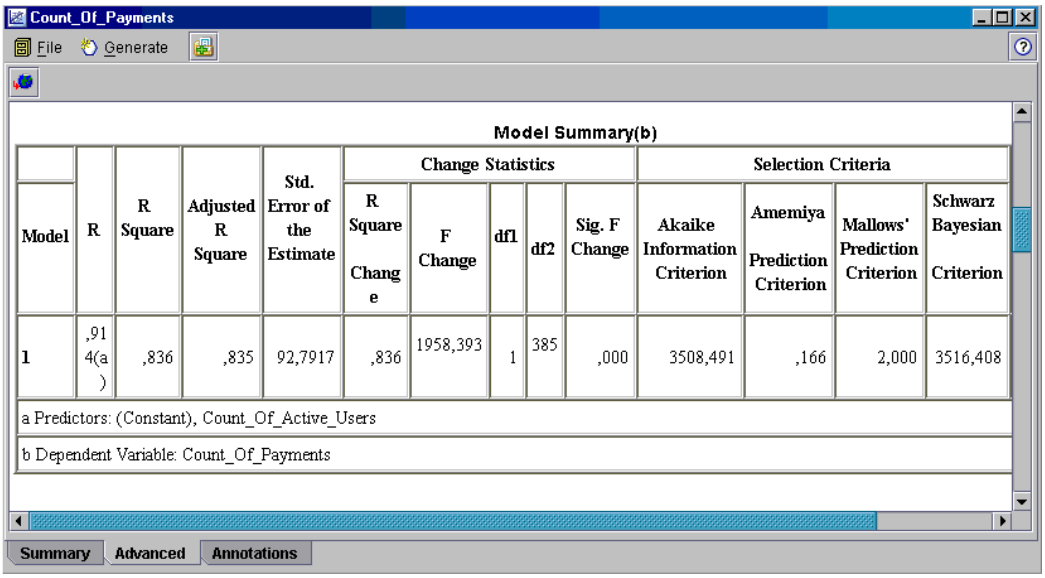

Figure 4

The degree to which two or more predictors (*X* variables) are related to the response (*Y*) variable is expressed in the correlation coefficient *R*, which is the square root of *R-square* [Hand, D., Mannila, H., and Smyth, P. (2001), SPSS (2002), Joreskog, K. (1999), Draper, N.R., and Smith, H. (1998)]. To interpret the direction of the relationship between variables, one should look at the signs (plus or minus) of the regression or parameters (θ). If a parameter is positive, then the relationship of this variable with the dependent variable is positive; otherwise in case the parameter is negative so is the relationship.

As can be seen in Figure 4 the value of R concerning the first step model is appropriate since it is close to 1. Additionally it can be observed the increase of Count_Of_Active_Users, is accompanied by an increase of the count of Payments in e-banking services.

R square is commonly used as measure of a model's goodness of fit. R square value of 0.836 is considered satisfactory and indicates an acceptable model, bearing in mind that:

- R square is a non-descending function of the number of predictor variables present in the model; that is, adding more historical data and predictor variables (X's), has almost constantly an increasing effect on R square. This is because the addition of predictor variables to the model reduces the prediction errors.
- R square assumes that the data set being analysed is the entire population while in fact, it represents only a sample of the population.

Adjusted R square measures the proportion of the variation in the response variable due to the predictor variables. Unlike R square, adjusted R square accounts for the degrees of freedom associated with the sums of the squares. Therefore, even though the residual sum of squares decreases or remains constant as new predictor variables are added, this is not the case for the residual variance. This is the reason, adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square.
If adjusted R square is significantly lower than R square, this normally means that some predictor variables are missing. The absence of these variables causes the improper measurement of the variation in the dependent variable.
Adjusted R square value of 0.835 is almost the same with R square indicating therefore an acceptable model.

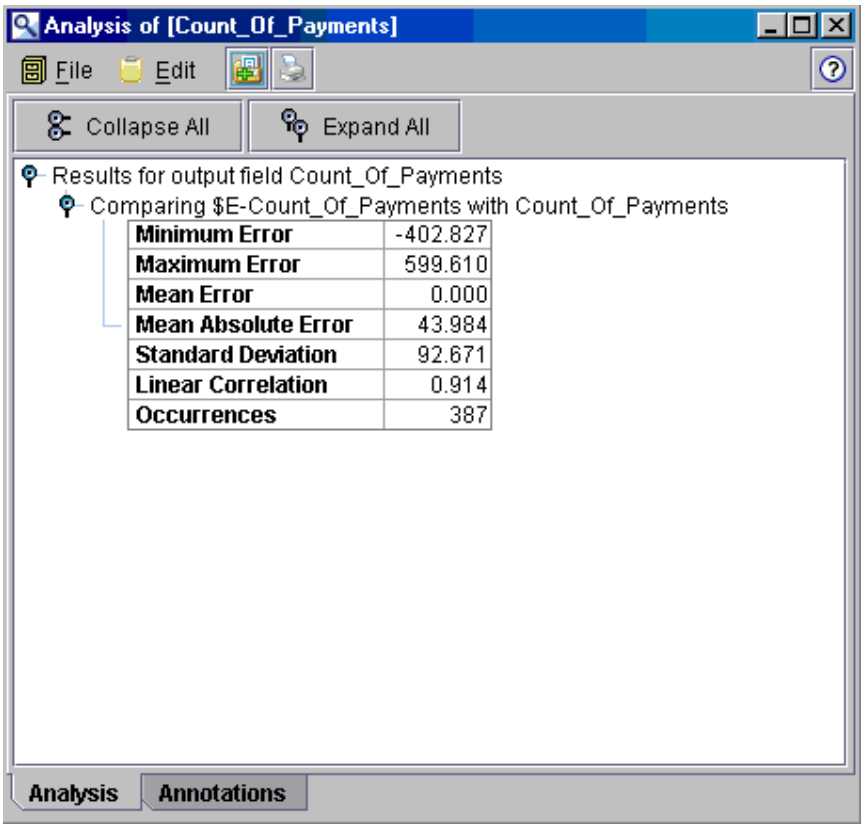

Figure 5

Finally, as can been seen in Figure 5 the level of Linear Correlation of the model is 0.914. Since this value approaches unity it indicates a strong positive relation, such that high predicted values are associated with high actual values and vice versa.

5. Conclusions and future work

In this study the establishment of a prediction model concerning the number of payments conducted through Internet as related to the number of active users is investigated along with testing its accuracy.

Using the Linear Regression method it was concluded that there exists strong correlation between the number of active users and the number of payments these users conduct. It is clear that the increase of the users' number results in increase of the transactions made. Therefore, the enlargement of the group of active users should be a strategic target for a bank. It has been proposed by certain researches that the service cost of a transaction reduces from €1.17 to just €0.13 in case it is conducted in electronic form. Therefore, a goal of the e-banking sector of a bank should be the increase of the proportion of active users compared to the whole number of enlisted customers from which a large number do not use e-services and is considered inactive. Therefore, knowledge management assisted by prediction models leads to valuable conclusions and contributes to higher bank profitability.

Future work includes the creation of prediction models concerning other e-banking parameters like the transactions value, the use of specific payment types, commissions of electronic transactions and e-banking income in relation to internal bank issues. Also extremely interesting is the case of predictive models based on external financial parameters.

The prediction model of this study could be determined using also other methods of data mining. Use of different methods offers the ability to compare between them and select the more suitable. The acceptance and continuous training of the model using the appropriate Data Mining method results in extraction of powerful conclusions and results.

6. References

Draper, N.R., and Smith, H. (1998) "*Applied Regression Analysis*," John Wiley & Sons, Inc.

Foster, D., and Stine, R. (2002) "Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy", *Center for Financial Institutions Working Papers from Wharton School Center for Financial Institutions, University of Pennsylvania.*

Hand, D., Mannila, H., and Smyth, P. (2001) "*Principles of Data Mining*". The MIT Press.

Hong, S., and Weiss, S. (1999) "Advances in Predictive Model Generation in Data Mining", *Proceedings 1st International Workshop Machine Learning and Data Mining in Pattern Recognition.*

Hong, S.J., and Weiss, S. (2000) "Advances in Predictive Model Generation for Data Mining", *Pattern Recognition Letters Journal.*

Joreskog, K. (1999) "What is the interpretation of R^2;", *http://www.ssicentral.com/lisrel/column3.htm.*

Laud, P., and Ibrahim, J. (1995) "Predictive Model selection", *Journal of the Royal Statistics Society.*

Madeira, S.A. (2002) "Comparison of Target Selection Methods in Direct Marketing", *MSc Thesis, Technical University of Lisbon.*

Raftery, A., Madigan, D., and Hoeting, J. (1997) "Bayesian Model Averaging for Linear Regression Models**",** *Journal of the American Statistical Association.*

SPSS (2002) "*Clementine 7.0 Users's Guide*". Integral solutions Limited.

Zupan, B., Demsar, J., Kattan, M., Ohori, M., Graefen, M., Bohanec, M., and Beck, J.R. (2001) "Orange and Decisions-at-Hand: Bridging Predictive Data Mining and Decision Support", *Workshop Integrating Aspects of Data Mining, Decision Support and Meta-Learning.*