Predictive Model in Electronic Banking Data

Vasilis Aggelis

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

Abstract

Generation of predictive models is an important process in the banking area. Data mining methods are useful in approaching present customers in a closer way increasing also their population. These methods are part of successful banking services campaigns and contribute to the development of new banking products. A specific area exhibiting interesting features is electronic banking (e-banking) and its users-customers. E-banking as an alternative banking service network increases bank profit and reduces the cost while increasing the customer number through a user-friendly interface and application of new services. As a result, it is of major importance for a bank to be in position to predict e-banking data, such as daily use, especially in respect to external factors that potentially influence the market like the stock market rate or/and a particular bank share. Scope of this paper is the presentation of a predictive model, concerning daily use of e-banking services.

Keywords: Data Mining, e-Banking, Predictive model, Linear Regression, Evaluation Chart.

Περίληψη

Η δημιουργία μοντέλων πρόγνωσης είναι μια σημαντική διεργασία στον τραπεζικό χώρο. Οι Data Mining μέθοδοι είναι χρήσιμοι στην καλύτερη προσέγγιση υφιστάμενων πελατών, αλλά και στην αύξηση του αριθμού τους. Οι μέθοδοι αυτές μπορούν να λειτουργήσουν υποβοηθητικά και σε διαφημιστικές καμπάνιες μια Τράπεζας καθώς και σε παραγωγή νέων προϊόντων. Μια εξειδικευμένη περιοχή του τραπεζικού χώρου είναι η ηλεκτρονική τραπεζική και οι πελάτες της. Η ηλεκτρονική τραπεζική έχει ιδιαίτερα χαρακτηριστικά καθώς είναι ένα εναλλακτικό τραπεζικό κανάλι, το οποίο αυξάνει τα κέρδη της Τράπεζας και μειώνει τα λειτουργικά της κόστη. Είναι πολύ σημαντικό για έναν τραπεζικό οργανισμό να είναι σε θέση να προβλέψει παραμέτρους της ηλεκτρονικής τραπεζικής. Στην εργασία αυτή παρουσιάζουμε ένα μοντέλο πρόγνωσης της ημερήσιας χρήσης ηλεκτρονικών υπηρεσιών σε σχέση με τον εξωγενή παράγοντα της τιμής του Γενικού δείκτη του χρηματιστηρίου Αθηνών και της τιμής της μετοχής της Τράπεζας.

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. 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 of e-banking application is studied. We developed this predictive model using the stepwise linear regression method. 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 [D. Foster, R. Stine, 2002., B. Zupan, J. Demsar, M. Kattan, M. Ohori, M. Graefen, M. Bohanec, and J.R. Beck., 2001., A. Raftery, D. Madigan, and J. Hoeting., 1997., P. Laud, and J. Ibrahim, 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 [D. Hand, H. Mannila, P. Smyth, 2001., A. Raftery, D. Madigan, and J. Hoeting, 1997., N.R. Draper, and H. Smith, 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 ["*Clementine 7.0 Users's Guide*", 2002., S.J. Hong, and S. Weiss, 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 [B. Zupan, J. Demsar, M. Kattan, M. Ohori, M. Graefen, M. Bohanec, and J.R. Beck., 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 ["*Clementine 7.0 Users's Guide*", 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 predictive models in e-banking

Scope of this study is the generation of a predictive model concerning the use of e-banking in relation to external financial factors. The significance of establishment of such a model is undoubtful as it helps in prediction, decision making and design of the bank policy, since the bank progress is not influenced from internal situations alone but also from the general status of the national economy.

A crusial measure of national finance market which is influenced from internal and external situations is the Athens Stock Market Rate. A second important measure, which represents the entire state of a bank organization, is its share value. The daily use of e-banking services was computed as the number of daily logins.

For the model purposes, we define one business *Day* as the time measure. This is because only on business dates, Athens Stock Market operates. Response variable (Y) is the *Count of daily Logins* (CoL) in e-banking services. Predictor variables (X_i) are daily *Athens Stock Market Rate* (ASMR) and daily *Bank Share Value* (BSV).

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

Business Day	CoL	ASMR	BSV
…	…	…	…
23/01/2002	279	2,558.88	3.94
24/01/2002	244	2,608.17	3.92
25/01/2002	271	2,623.67	3.82
28/01/2002	315	2,614.91	3.80
29/01/2002	277	2,614.74	3.82
30/01/2002	285	2,606.76	3.84
…	…	…	…

Table 1

In order to generate a prediction like (2.1) the stepwise linear regression ["*Clementine 7.0 Users's Guide*", 2002., S.A. Madeira., 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).

CoL^ = (-0.6526)*ASMR + 186.5*BSV + 1,254.4 (4.1)

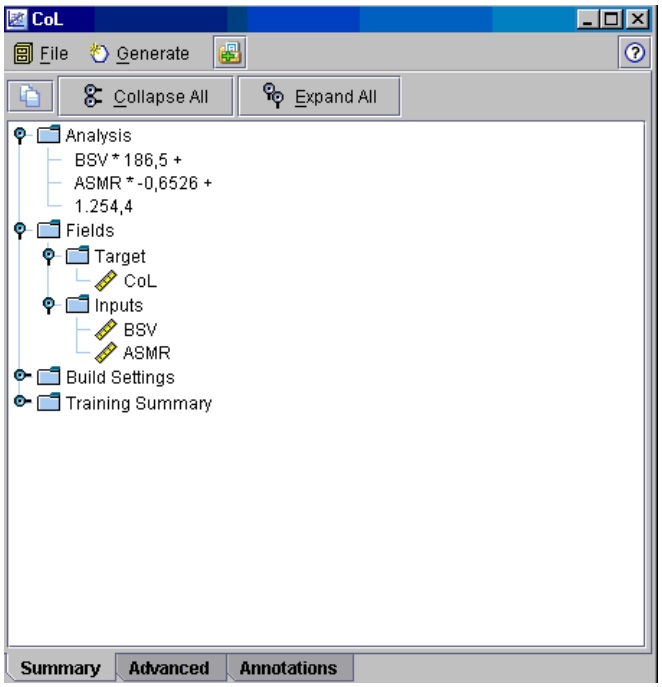

Figure 1

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

As mentioned before the fitness of the model is good, because chart shows that the gain rises steeply towards 100% and then levels off. Using the prediction of the model, calculate the percentage of CoL 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.

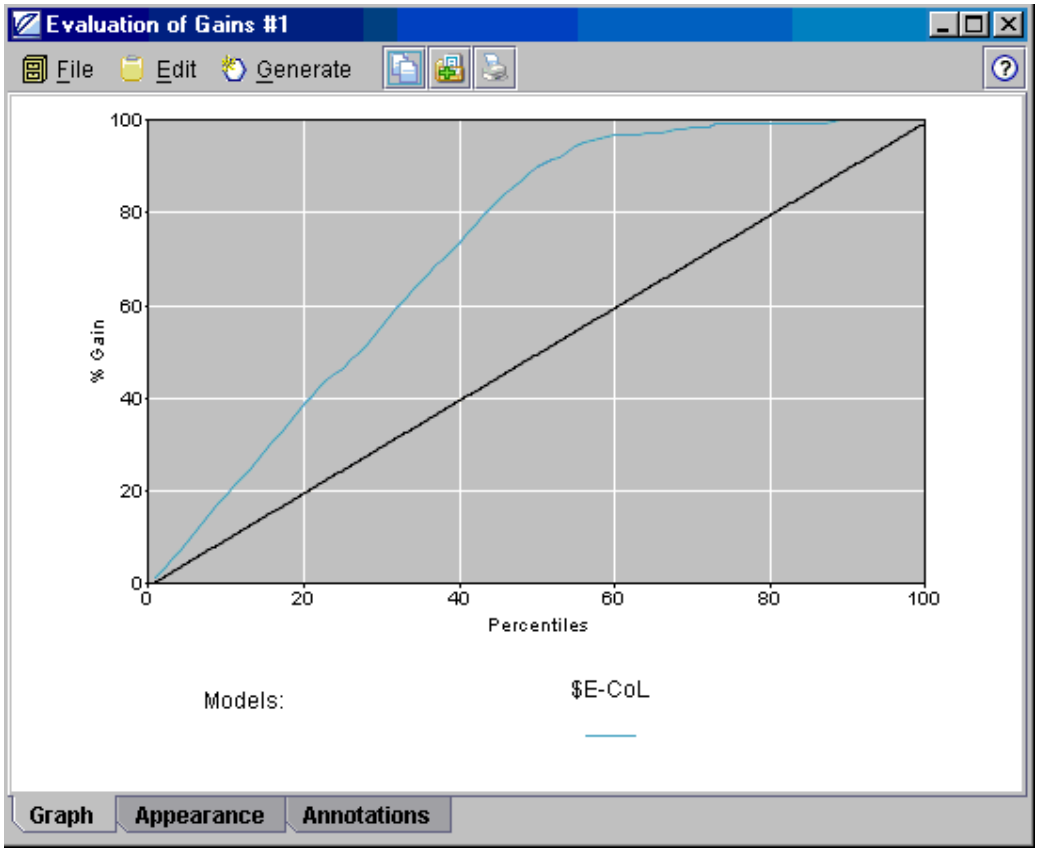

Figure 2

Lift Chart

A second important indication of the model fitness is the 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.

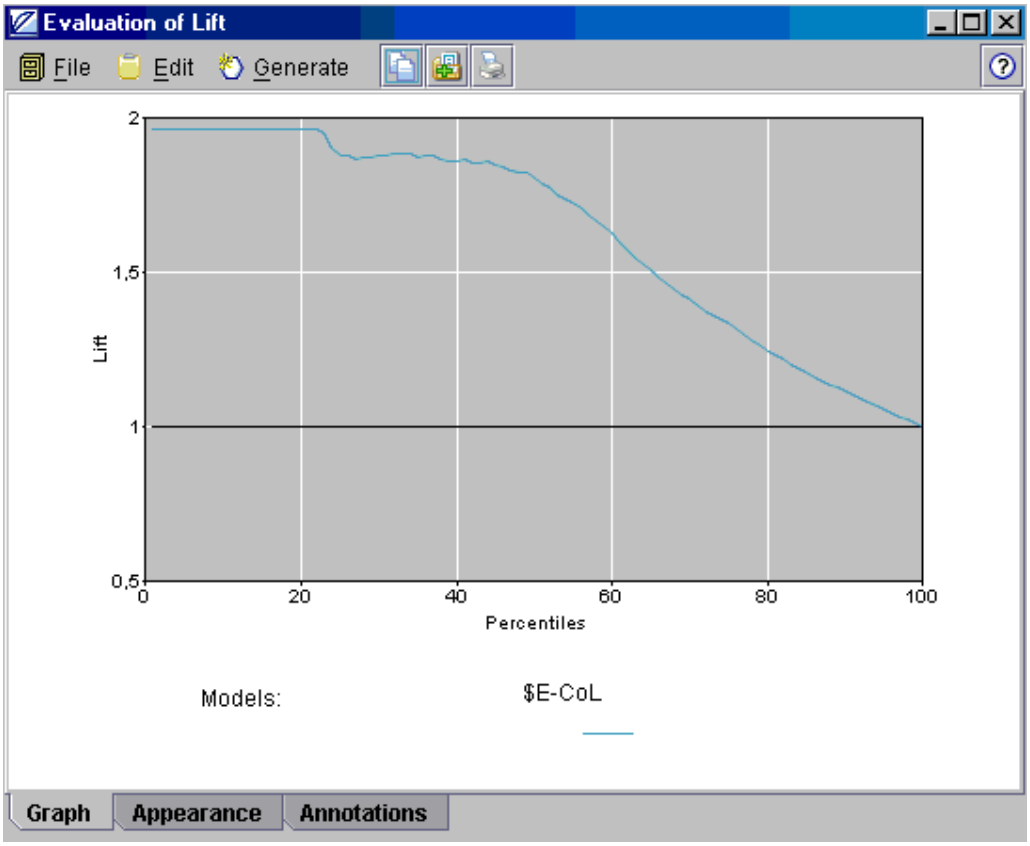

Figure 3

Other measures

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

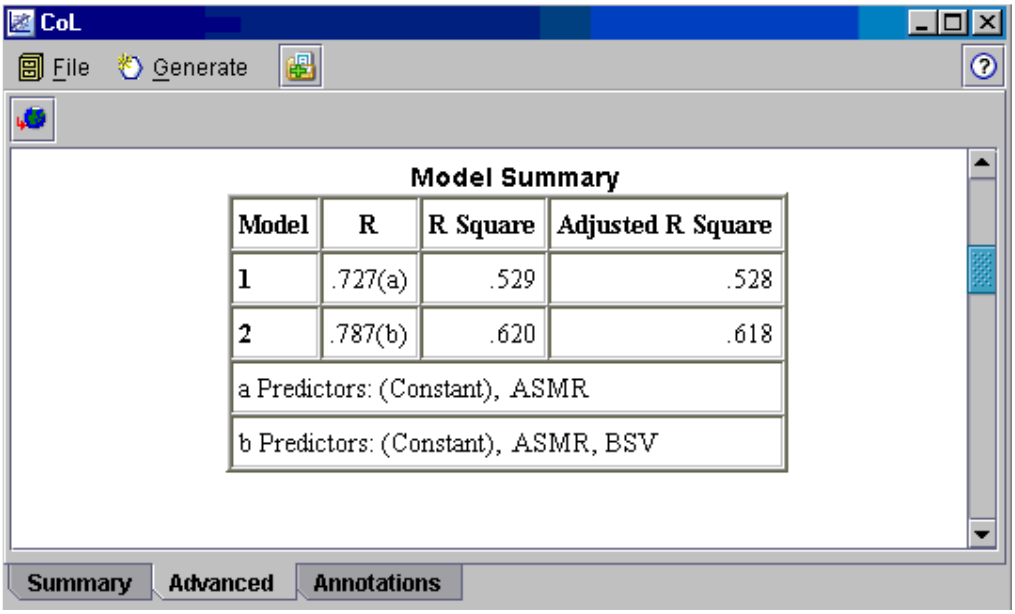

Figure 4

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* [D. Hand, H. Mannila, P. Smyth, 2001., "*Clementine 7.0 Users's Guide*", 2002., K. Joreskog, 1999., N.R. Draper, and H. Smith, 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 second step model is appropriate since it is close to 1. Additionally it can be observed that decrease of ASMR and the increase of BSV, is accompanied by an increase of the count of Logins in e-banking services.

R square is commonly used as measure of a model's goodness of fit. An R square value near 1 indicates a perfect regression. R square value of 0.62 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. The nearest the adjusted R square is to 1, the better the model is.

Adjusted R square value of 0.618 is almost the same with R square indicating therefore an acceptable model.

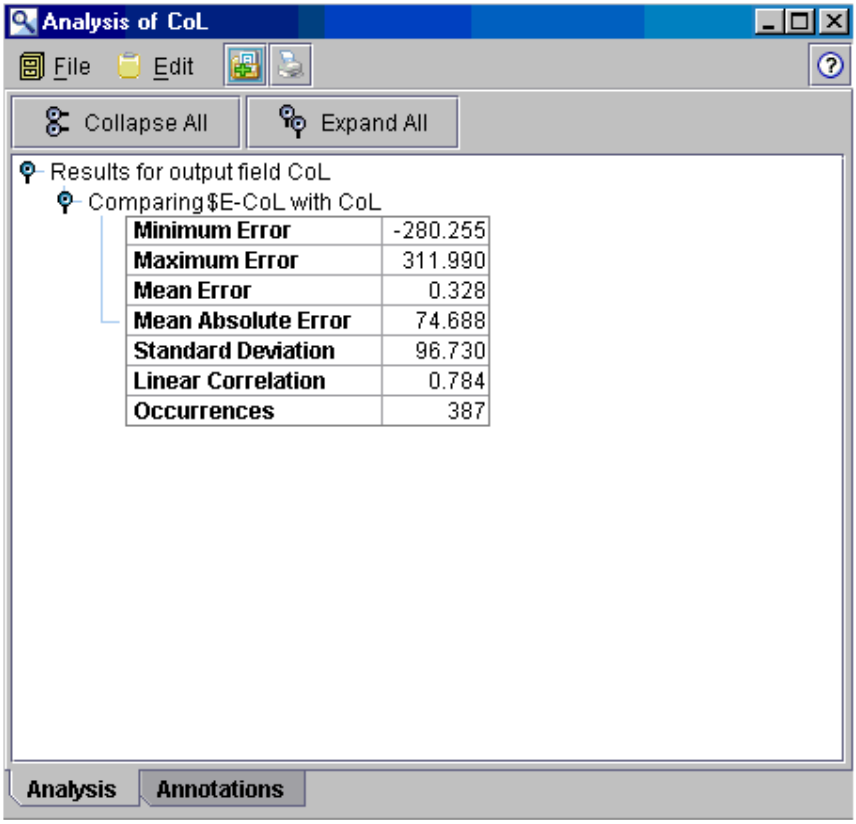

Figure 5

Finally, as can been seen in Figure 5 the level of Linear Correlation of the model is 0.784. 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 development of a predictive model concerning the number of users of e-banking services relatively to the stock market rate and the bank share value is described while experimental results are also supplied.

It is concluded that there exists a strong relation between the use of an alternative bank channel (specifically e-banking) and the status of generally the stock market rate and also the individual bank share values. This trend could be expected since there is a close relation between the level of the stock market rate and that of national economy. An interesting feature rises from the observation concerning the increase of e-banking users with the reduction of ASMR, something emphasized by the long descending trend of the ASMR and the continuously increasing number of Internet users.

Future plans employ the development of predictive models using other external sources either national, such as the inflation rate or/and international such as foreign stock market rates, oil price and others. Apart from this, models concerning other features of e-banking can be developed like the number of transactions, the number of active users and a number of others.

Finally the use of other data mining methods (Neural Networks, Decision trees) for predictive model development is expected to enhance the effectiveness through comparison between different models, yielding information regarding the degree of suitability of each method.

6. References

"*Clementine 7.0 Users's Guide*". Integral solutions Limited, 2002.

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

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

D. Foster, R. Stine, 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.

D. Hand, H. Mannila, P. Smyth "*Principles of Data Mining*". The MIT Press, 2001.
K. Joreskog, 1999 "*What is the interpretation of* R^2*;*", www.ssicentral.com/lisrel/column3.htm.

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

P. Laud, and J. Ibrahim, 1995 “Predictive Model selection”, *Journal of the Royal Statistics Society*.

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

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

S.J. Hong, and S. Weiss. 2000 “Advances in Predictive Model Generation for Data Mining”, *Pattern Recognition Letters Joyrnal*.