Expert Systems with Applications 36 (2009) 11402–11417

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Genetic programming for credit scoring: The case of Egyptian public sector banks

Hussein A. Abdou *

Salford Business School, University of Salford, Salford, Greater Manchester, M5 4WT, UK

ARTICLE INFO

Keywords:
Genetic programming
Credit scoring
Weight of evidence
Egyptian public sector banks

ABSTRACT

Credit scoring has been widely investigated in the area of finance, in general, and banking sectors, in particular. Recently, *genetic programming* (GP) has attracted attention in both academic and empirical fields, especially for credit problems. The primary aim of this paper is to investigate the ability of GP, which was proposed as an extension of genetic algorithms and was inspired by the Darwinian evolution theory, in the analysis of credit scoring models in Egyptian public sector banks. The secondary aim is to compare GP with *probit analysis* (PA), a successful alternative to logistic regression, and *weight of evidence* (WOE) measure, the later a neglected technique in published research. Two evaluation criteria are used in this paper, namely, *average correct classification* (ACC) rate criterion and *estimated misclassification cost* (EMC) criterion with different *misclassification cost* (MC) ratios, in order to evaluate the capabilities of the credit scoring models. Results so far revealed that GP has the highest ACC rate and the lowest EMC. However, surprisingly, there is a clear rule for the WOE measure under EMC with higher MC ratios. In addition, an analysis of the dataset using Kohonen maps is undertaken to provide additional visual insights into cluster groupings.

1. Introduction

Credit scoring models are widely used by financial institutions, especially banks, to assign credit to good applicants and to differentiate between good and bad credit. Using credit scoring can reduce the cost of the credit process and the expected risk of being a bad loan, enhancing the credit decision, and saving time and effort (Lee, Chiu, Lu, & Chen, 2002; Ong, Huang, & Tzeng, 2005). Particularly, with the fast growth in the credit industry and the huge loan portfolio management, credit scoring is regarded as a one the most important techniques in banks and has become a very critical tool during recent decades.

It is believed that the Egyptian banking sector has been "tough" since 1999, and is "expected to remain so" for the performance of the banking sector in Egypt has shown an "ongoing profitability weakness due to revenue pressure" and a high incidence of problem loans (Central Bank of Egypt, CBE, 2006/2007; Oldham & Young, 2004). The Egyptian banking sector is being reformed to deal with this problem, which was approved in September 2004 by the President of the Arab Republic of Egypt. The main objective of this reform plan was to develop a more effective financial instrument, to strengthen the system's infrastructure, and to enhance competitiveness through increased private participation within the overall development strategy. The main pillars of the reforming plan are: firstly, banking sector consolidation and privatization through reducing the number of operating banks; secondly, financial and managerial restructuring; thirdly, solution of the bad loans problem, and finally, updating the supervision sector at the banking sector. This reforming plan also included the privatization of one of the public sector banks (CBE, 2006/2007; Oldham & Benaddi, 2005).

Egyptian banks' lending activities remarkably expanded during the last two decades. Banks' credit activities witnesses an increase, compared with the previous period, of LE28 billion (7.90%) against LE19.90 billion (6.10%) constituting LE381.80 billion or 37.40% of banks total assets and 54.50% of total deposits at the end of December, 2007. Also the pickup in foreign currency loans witnessed an increase by LE17.80 billion (16.90%) constituting LE123 billion at the end of December 2007, as well. Loans and advances exceeding one year, excluding discounts, also expanded; they went up by LE27.60 billion or 7.80%, to LE380 billion at the end on December 2007 (CBE, 2007/2008).

In view of the size of lending activities and to make efficient decisions in the granting of credit for consumer loans, it is posited that different statistical scoring techniques can be beneficially introduced to supplement the judgemental techniques, which currently are based on single numerical evaluation systems and the CBE's own perspective of creditworthiness. Indeed discussions with key banking personnel have suggested that all public sector banks in Egypt are using judgemental techniques in their evaluation process. Therefore, the role that scoring techniques can play is critical in helping to reduce the current and/or the expected risk they face; because of an inadequate risk-reduction through efficient diversification, and to support the banking sector reforming plan as currently applied.

* Tel.: +44 1612 953001; fax: +44 1612 955022.
E-mail address: h.abdou@salford.ac.uk

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.01.076

The categorisation of good and bad credit is of fundamental importance, and is indeed the objective of a credit scoring model (Lee et al., 2002; Lim & Sohn, 2007). The need of an appropriate classification technique is thus evident. But what determines the categorisation of a new applicant? Characteristics, such as marital status, income and age, have been recommended (Chen & Huang, 2003). The classification techniques themselves can also be categorised into conventional methods and advanced statistical techniques. The former include weight of evidence, multiple linear regression, discriminant analysis, probit analysis and logistic regression. The latter comprise various approaches and methods, such as, fuzzy algorithms, genetic algorithms, expert systems, and neural nets (Hand & Henley, 1997).

A few number of studies have investigated the use of WOE measure in this field, also results were comparable with those from other techniques (Abdou, 2009; Bailey, 2001; Banasik, Crook, & Thomas, 2003; Siddiqi, 2006). Furthermore, PA has been investigated, as well, and compared with other statistical scoring models (Banasik et al., 2003; Greene, 1998; Guillen & Artis, 1992); also classification results were very close to other techniques, such as logistic regression and better than discriminant analysis (Banasik et al., 2003). GP models were proposed by Koza (1992) based on Darwin's evolution theory. The use of GP applications is a rapidly growing area (Chen & Huang, 2003), and number of applications has increased during the last couple of decades, such as bankruptcy prediction (Etemadi, Rostamy, & Dehkordi, 2009; McKee & Lensberg, 2002), scoring applications (Huang, Chen, & Wang, 2007; Huang, Tzeng, & Ong, 2006), classification problems (Lensberg, Eilifsen, & McKee, 2006; Ong et al., 2005; Zhang & Bhattacharyya, 2004) and financial returns (Xia, Liu, Wang, & Lai, 2000).

However, unlike other published works, which used other statistical techniques, such as neural networks (Abdou, 2009; Lee & Chen, 2005; Malhotra & Malhotra, 2003; Oldham & Benaddi, 2005; Tsai & Wu, 2008; West, 2000), discriminant analysis and logistic regression (Elliott & Filinkov, 2008; Lee et al., 2002; Desai, Crook, & Overstreet, 1996), the focus of chosen methodologies in this paper is on two types of GP models, namely, the program model and the team model, as well as conventional techniques, such as WOE and PA. WOE has been mainly neglected in published work, yet may have much to offer, whilst PA can be a successful alternative to logistic regression (see below).

Here, the focus of the chosen environment for credit scoring investigation is upon the Egyptian public sector banks. As stated by Oldham and Young (2004), the main problems with the Egyptian banking sector exist in the large public sector banks, whose assets represent more than 50% of the whole system. The author was not aware of any other studies having investigated the use of statistical scoring models in evaluating consumer loans in whole Egyptian public sector banks. Since statistical techniques have not been used in the Egyptian public sector banks, the sample selection bias problem should be less serious compared with other studies, and this highlights the importance of the present study.

This paper is organized as follows: part two covers methodology, including data collection and sampling method; part three explains the empirical results for both the whole sample scoring models and the validated scoring models; part four compares the classification and misclassification results for different techniques; and finally, part five concludes the study results and suggests areas for future research.

2. Research methodology

In this paper, three different credit scoring modelling techniques are used in building the scoring models. The first model is the WOE measure, which is one of the earliest techniques used in credit scoring, which has a few applications in the field (Abdou, 2009; Bailey, 2001; Banasik et al., 2003). The second model is the PA model, which is also usually used with other statistical techniques for comparative purposes (Guillen & Artis, 1992; Pindyck & Rubinfeld, 1997). Finally, GP models are applied as proposed by Koza (1992) as an extension to genetic algorithms, and inspired by the Darwin's evolution theory (Koza, 1994). Here two types of GP models are used, a program model/evolved program, *which is a single program*, and a team model, *which is a combination of single programs*. The advantage of applying the team model is that the currently selected software creates this model in order to produce better results than any of the single program models can achieve.

The proposed models are discussed in Section 2.1 and the evaluation criteria in Section 2.2. The data collection, sampling method and variables' identification are discussed in Section 2.2.2. Data cases in the validated scoring sample are divided judgementally into two samples: a training dataset (67%), and a testing dataset (33%).

2.1. Proposed scoring models

2.1.1. Weight of evidence measure

The WOE measure has a long history in credit scoring models. It focuses on the odds ratio of good scores to bad scores. The information odds (IO) ratio is used to analyse the difference between two distributions without affecting the overall population. Thus, the IO equation is as follows:

$$IO = (\text{Good scores sub} - \text{classification percent})/(\text{Bad scores sub} - \text{classification percent})$$

i.e. the number of good scores within a given category as a percentage of the number of good scores for all categories, divided by the number of bad scores for a given category as a percentage of the number of bad scores for all categories.

WOE can be calculated from the IO using the logarithmic function, which can be considered as raw scores, as follows:

$$WOE = Ln(IO).$$

The information value (IV), *or total strength of the characteristics*, is used to identify the strength of different variables, as an alternative to other statistical tests, such as chi-square. IV can be calculated, from as follows:

$$IV = \Sigma[(G\% - B\%) \times WOE].$$

The importance of IV as a measure can be seen by its provision of the maximum contribution to the attributes that generate the maximum impact on the score. IV is sometimes adjusted by a discretionary multiplicative factor, for example, by multiplying by 100, or calling it Power, and multiplying by 1000 (Bailey, 2001; Siddiqi, 2006).

Bailey (2001) recommends the following values as a guideline:

Less than 0.03	:	poor prediction
From 0.03 to less than 0.10	:	weak prediction
From 0.10 to less than 0.30	:	average prediction
From 0.30 to less than 0.50	:	strong prediction
Over 0.50	:	very strong prediction

Of course, there is a subjective element attached to the categorised definitions. Nevertheless, IV is widely used in industry, because of its predictive capability. Point Score for the WOE, is determined as follows:

$$Point\ Score = \Sigma\{[P/Ln(2) \times R_W] \times [WOE + c]\}$$

where P is the score at which the odds are doubled; R_W is the correlation coefficient (from a multiple regression) between the

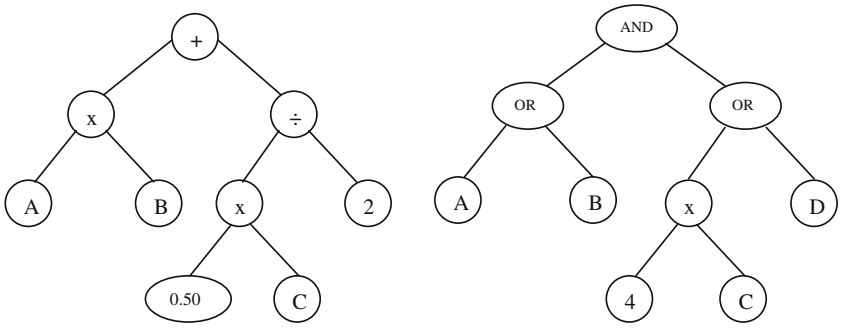

Fig. 1. Two examples of GP trees using simple mathematical operators and using conditional statements.

respective variable and the *WOE*; and *c* is a constant applied to each variable (Bailey, 2001; Siddiqi, 2006).

2.1.2. Probit analysis

Probit, or probability unit, has a long history, having been proposed in the early 1930s (Maddala, 2001; Pindyck & Rubinfeld, 1997). The methodology enables the probability to be estimated of a unit value of a dichotomous variable. These probabilities become coefficient estimates. Under a probit model, a linear combination of the independent variables is transformed into its cumulative probability value from a normal distribution. The software application estimates values for the coefficients in this linear combination, such that the cumulative probability equals the actual probability that the dichotomous variable is one. Hence:

$$Prob(y = 1|V) = \Phi(\alpha + \delta_1 V_1 + \delta_2 V_2 + \cdots + \delta_n V_n),$$

where y is the zero-one dichotomous variable for a given set of value; Φ is the value from the cumulative normal distribution function; α is the intercept term; and δ_i represents the respective coefficient in the linear combination of independent variables; V_i, for $i = 1$ to n, of vector V (see, for example, Banasik et al., 2003; Pindyck & Rubinfeld, 1997).

2.1.3. Genetic programming

GP began as a subset of genetic algorithmic techniques, and can be considered as an extension of genetic algorithms (Koza, 1992). Genetic algorithms transform a dataset according to fitness value, by applying genetic operations. Under genetic algorithms, the solution is in the form of a "string" (Koza, 1992). In GP a set of competing programs are randomly generated by processes of mutation and crossover, which mirror the Darwinian theory of evolution, and the resultant programs are evaluated against each other. Generally, GP generates competing programs in the LISP (or similar) language as a solution output (Koza, 1994; Nunez-Letamendia, 2002).

Koza (1992) proposed GP as one of the most recent and advanced techniques used in credit scoring problems based on simple mathematical operators and/or conditional statements to create a population of randomly generated programs with a fitness value for each one.

The representation of a GP tree can be explained based on "function" and "terminal" sets, the former such as simple mathematical operators (+, −, ×, ÷) and/or conditional statements (If … Then…), and the latter which contains inputs, equations etc in the GP tree (Ong et al., 2005; Teller & Veloso, 2000). An example of these two sets is shown in Fig. 1.

Once the GP population has been created the next procedure normally includes the fitness values and genetic operators, such as a crossover operation (either based on different or based on identical parents to reproduce the children), mutation and reproduction. An example of a crossover GP tree based on different parents is shown in Fig. 2.

Two types of GP models are used in this paper: a program model/evolved program, *which is a single program*, and a team model, *which is a combination of single program models in order to produce better results than any of the single program models*. Discipulus™ Professional Software, which is based on a multi-run linear GP system, is used in developing the GP scoring models. This software applies the GP technique typically utilizing machine-code to develop the programs. The programs that evolve are similar to C++ and other imperative languages, rather than LIPS and such functional programming languages (Koza, 1992; Mukkamala, Vieira, & Sung, 2008). The default search operator of the software applies a 30% block mutation rate, a 30% instruction mutation rate, and a 40% instruction data mutation rate; and a homologous crossover of 95%. From the GP itself the default mutation frequency is 95% and the crossover frequency is 50%.

The current GP program performs a tournament, allowing only a given number of programs, with a random procedure in the selection process and the worst performing programs being replaced (Deschaine & Francone, 2008). Then, GP copies the winner programs into other programs through random crossover, *in which sections of trees are swapped*, and mutation, *in which sections of trees are replaced but not swapped*. Some authors use absolute errors for fitness functions (Huang et al., 2006; Ong et al., 2005), whilst others including the current researcher, use linear combinations of mean square errors and mean classification errors (Koza, 1992; Mukkamala et al., 2008). The fitness function of an evolved program can be calculated as follows:

$$F(ep) = \alpha\left[\sum_{i=1}^{n}(a_i - e_i)^2\right] + \beta(CE)$$

where F is the fitness function; ep is the evolved programme; α is a weighting based on the training sample size and number of outputs; a_i is the actual observation and e_i is the expected (predicted) observation; β is a weighting related to the classification errors in the training sample and CE is the classification error (Golgberg, 1989; Koza, 1994; Mukkamala et al., 2008).

Three samples are used to develop the genetic scoring models: training data (*used for genetic evolution*); validation data (*used for model selection*); and applied/testing data (*played no role in training or model selection*).[1] A GP program/team model is designed using equal training and validation datasets (both samples are combined as a training sample for comparison purposes with other techniques)

[1] The terms "genetic evolution" and validation data for "model selection" have been used by the providers of Discipulus™ Genetic-Programming Software.

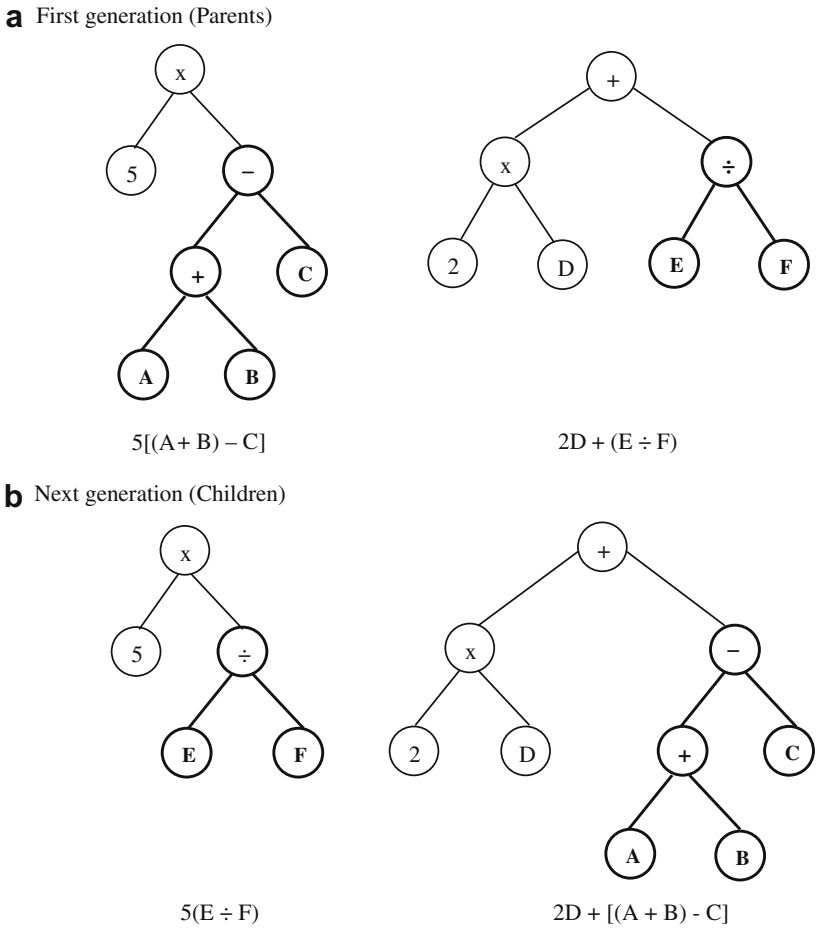

Fig. 2. An example of a crossover GP tree based on different parents (a) first generation (parents) and (b) next generation (children).

Table 1
Classification matrix.

	Predicted observations		
	g	b	
Actual observations			
G	G_g	G_b	TG
B	B_g	B_b	TB
	Tg	Tb	TN

Notations: G = actual good; g = predicted good; B = actual bad; b = predicted bad; G_g = actual good predicted good; G_b = actual good predicted bad; B_g = actual bad predicted good; B_b = actual bad predicted bad; TG = total actual good observations; TB = total actual bad observations; Tg = total predicted good observations; Tb = total predicted bad observations; and TN = total number of observations in the dataset.

and the applied/testing dataset, to see how the model works on data that played no role in building the model. As a part of this GP software design, the team model employs a combination of an odd number of single programs (minimum 1; maximum 9), for each of the proposed samples, i.e. from training, validation and applied, data.

2.2. Evaluation criteria

Applying the public sector banks' dataset, two different evaluation criteria are used: firstly, the ACC rate criterion, as a significant criterion in evaluating the classification capability of the proposed scoring models; and secondly, the EMC criterion, as a crucial criterion to evaluate the overall credit scoring effectiveness and to find the minimum EMC for the suggested scoring models.

2.2.1. Average correct classification rate criterion

The ACC rate is one of most widely used criteria in the area of credit scoring applications in general, and accounting and finance in particular. The ACC rate measures the proportion of the correctly classified cases (good and bad) in a particular dataset (See Table 1).

The majority of credit scoring applications either in accounting and finance or other fields have used the ACC rate as a performance evaluation measure (Paliwal & Kumar, 2009). The idea of correct classification rates came from a matrix, which is occasionally called "a confusion matrix" (Yang, Wang, Bai, & Zhang, 2004), otherwise called a classification matrix. As shown in Table 1, a classification matrix presents the combinations of the number of actual and predicted observations in a dataset. From this matrix a number of useful rates can be calculated, namely: ACC rate, represented by $(G_g + B_b)/TN$; total error rate, *which is a complementary value of the ACC rate*, and given by $(G_b + B_g)/TN$; and other measures, such as the correctly classified good rate (G_g/TG); and the correctly classified bad rate (B_b/TB); Type I error rate (G_b/TG) and Type II error rate (B_g/TB).

On the one hand, in this paper the ACC rate is believed to be an important criterion to be used, especially for new users of credit scoring, such as in the Egyptian public banking sector environment, because it highlights the accuracy of the predictions. On

the other hand, the ACC rate criterion does not accommodate differential costs, to a bank, arising from different types of error. Specifically, it ignores different misclassification costs for the G_b and the B_g observations. For, in the real field it is believed that the cost associated with Type II errors is normally much higher than that associated with Type I errors (Baesens et al., 2003), as will be explained in the next section.

2.2.2. Estimated misclassification cost criterion

The second criterion to be used is the estimated misclassification cost criterion. Few credit scoring applications have used EMC criterion in the field (Lee & Chen, 2005; West, 2000). The reason is that the trustworthy or consistent estimates of the misclassification costs are a complicated and real challenging job to be provided, therefore valid prediction might not be available, as noted by Lee and Chen (2005).

This criterion gives an evaluation of the effectiveness of the scoring models' performance, which can cause a serious problem to the banks in the case of the absence of these estimations, especially with the B_g cases. The following equation, which is similar to that by West (2000), is used in computing the EMC:

$$EMC = C(\mathrm{I}) \times (G_b/TG) \times (TG/TN) + C(\mathrm{II}) \times (B_g/TB) \times (TB/TN),$$

where $C(\mathrm{I})$ is the misclassification cost associated with a Type I error; (G_b/TG) is the probability of a Type I error, expressed as a ratio of number of good credit predicted as bad (G_b) to total good credit (TG); (TG/TN) is the prior probability of good credit, namely, the ratio of the number of total good (TG) to the overall number of observations (TN); $C(\mathrm{II})$ is the misclassification cost associated with a Type II error; (B_g/TB) is the probability of a Type II error, expressed as a ratio of numbers of bad credit predicted as good (B_g) to total bad credit (TB); and (TB/TN) is the prior probability of bad credit, namely, the ratio of the number of total bad (TB) to the overall number of observations (TN).

The previous equation can be re-expressed as:

$$EMC = C(I) \times (G_b/TN) + C(II) \times (B_g/TN).$$

Lee and Chen (2005) stated that "it is generally believed that the costs associated with both Type I, *in which good credit is misclassified as bad credit*, and Type II, *in which bad credit is misclassified as good credit*, errors are significantly different" and "the misclassification cost associated with a type II error is much higher than the misclassification cost associated with a type I error". West (2000) noted that Dr. Hofmann, who compiled his German credit data, reported that the ratio of misclassification costs, associated with Type II and Type I, is 5:1.

In this paper, the emphasise is not only on this relative cost ratio at 5:1, but also it provides a *sensitivity analysis* using higher cost ratios at e.g. 7:1, 10:1, etc. Particularly, it is expected that the higher cost ratio might be more appropriate, especially for an environment such as the Egyptian banking sector. In addition to this, the prior probabilities of good and bad credit are set at 67.43% and 32.57%, respectively, using the actual ratios of good and bad credit in the Egyptian dataset.

EMC can be calculated, in part, from the classification matrix introduced earlier. The probabilities of Type I and Type II error rates can be determined by G_b/TG and B_g/TB, respectively. It is strongly suggested that the lowest EMC might very well be found in a model that does not have the highest ACC rate.

2.3. Data collection and sampling method

In order to build the proposed credit scoring models, a consumer loans' dataset was provided by the Egyptian commercial public sector banks. This consists of 1,262 personal loans with 851 good loans and 411 bad loans. It should be emphasized that this dataset is *pertinent* because of the large number of bad loans (32.57%) compared with good loans (67.43%).

Table 2
List of predictor variables proposed in building the credit scoring models for public sector banks.

Variables/description	Code
X_1 Loan Amount[a]	LOAN AMO
X_2 Loan Duration[a]	LOAN DUR
X_3 Type of Loan	–
X_4 Purpose of Loan	–
X_5 Age[a]	AGE
X_6 Marital Status[a]	DUM (MARR/SING/OTHER[b])
X_7 Gender[a]	GENDER
X_8 Dependants[a]	DEPE
X_9 Profession[a]	PROFE
X_{10} Educational Level[a]	EDUC
X_{11} House Status[a]	HOU STA
X_{12} Telephone[a]	TELE
X_{13} Monthly income[a]	MON INCO
X_{14} Utility Bill	–
X_{15} CBE Report[a]	CBE REP
X_{16} Personal Reputation	–
X_{17} Guarantees[a]	GUAR
X_{18} Field Visit[a]	FIE VISI
X_{19} Feasibility Study[a]	FEASI STU
X_{20} Credit Card Status[a]	CC STA
X_{21} Relation with Other Banks	–
X_{22} Loans from Other Banks[a]	LFOB
X_{23} Car Ownership[a]	CAR OWN
X_{24} Formal Documents	–
X_{25} Customer began to Default	–
Y Loan Quality (dependent variable)	LOAN QUA

[a] Variables finally selected in building the scoring models.
[b] Two dummies were used (married and single); the third (other) being implied.

Table 2 shows the list of predictor variables used in building the proposed credit scoring models for the Egyptian public sector banks. Each bank client is linked with 25 predictor variables in this dataset, besides the independent loan quality variable which is explained by two values, one for good credit and zero for bad credit. Some of these dataset' variables have identical values: such as type of loan, actually all cases in the currently used dataset are personal loans; and utility bills, for all cases in this dataset provided a utility bill when applying for loans and all of them provided formal documents.

Some variables have not been used in other published studies, such as CBE report, field visit, feasibility study and loans from other banks.[2] Indeed, from the review of literature to date, the author was not aware of other studies having used these variables in building scoring models for personal loans. By contrast, some other variables are used in other studies, such as loan duration, gender, telephone and guarantees.

The sampling method used in this paper is based on applying different samples to investigate the capabilities of the scoring models. First, the whole dataset is used under each of the proposed scoring techniques. The reason for this is to investigate the overall capability of different scoring models because of the benefits of the larger dataset. Second, however for the purpose of testing the predictive ability of the scoring models, a simple validation technique has been applied by classifying the dataset into training sample of 67% (846 cases), and testing (validation) sample of 33% (416 cases), that can be used to test the predictive effectiveness of the fitted model. All models, under the validation sampling method, were built using the training sample and were tested using the testing sample. By dividing the dataset, as previously explained, an

[2] For prediction purposes it is difficult to predict when a customer will start to default. Therefore, the X_{25} variable has been excluded from the final list.

Table 3
Classification results for the WOE and PA models using the whole sample.

Observed group model	Predicted group				Observed group model	Predicted group			
	G	B	T	Overall %		G	B	T	Overall %
WOE					WOE_1				
G	290	561	851	34.08	G	283	568	851	33.25
B	7	404	411	98.30	B	7	404	411	98.30
T			1262	54.99	T			1262	54.44
WOE_{T1}					WOE_{T11}				
G	383	233	616	62.18	G	393	223	616	63.80
B	92	319	411	77.62	B	92	319	411	77.62
T			1027	68.35	T			1027	69.33
WOE_{T2}					WOE_{T21}				
G	591	260	851	69.45	G	602	249	851	70.74
B	92	319	411	77.62	B	92	319	411	77.62
T			1262	72.11	T			1262	72.98
PA					PA_1				
G	757	94	851	88.95	G	754	97	851	88.60
B	134	277	411	67.40	B	135	276	411	67.15
T			1262	81.93	T			1262	81.62

Cut-off point 0.50.

investigation can be conducted into whether different results in terms of ACC rates and EMC might be achieved. The earlier procedure, of using 67% data as a training dataset and 33% as a testing dataset is clearly non-random.

3. Empirical results

STATGRAPHICS Plus 5.1, SPSS 16.00, Scorto™ Credit Decision Software and Discipulus™ Genetic-Programming Software were used to run the proposed credit scoring models in this paper. The detailed credit scoring results using the above-mentioned scoring modelling techniques under different sampling methods can be summarized as follows.

3.1. Whole sample credit scoring models:

All models, namely WOE measure, PA and GP, under this section were built using the whole dataset, namely 1,262 cases.

3.1.1. Weight of evidence measure

Following the WOE methodology explained earlier, all the selected nineteen variables, including two dummies, used in building the scoring models, had an acceptable information value (weak, average, strong, or very strong prediction), except four poor predictive variables, which were DUM SING, DUM MARR, HOU STA, and CAR OWN with an IV of 0.0058, 0.0036, 0.0012 and 0.0154, respectively (for IV details for all variables, see Appendix A). However, due to their potential importance and for a fair comparison with other overall sample techniques, it had been decided to keep them in the overall sample models namely, WOE, WOE_{T1}, and WOE_{T2}; but exclude them from the "stepwise" models, namely, WOE_1, WOE_{T11}, and WOE_{T21}.[3]

As shown in Table 3, for the WOE technique a 54.99% ACC rate was achieved with a 50% cut-off point and a maximum of 75.99% ACC (see Appendix B) with a 30% cut-off point. Using weak, average, strong and very strong predictor variables, denoting the model as WOE_1, 15 variables have been selected with a 54.44% ACC rate with a standard 0.50 cut-off point and a 76.94% maximum ACC rate with a 0.30 cut-off point.[4]

As a form of *sensitivity analysis*, further trial-models have been developed, taking into account the 235 cases out of 1,262 cases which had a corporate guarantee, which means there is no such chance of any of them to be defaulted. In practice, using WOE, if there is a corporate guarantee, accept the application. If not, apply the normal scoring procedures. As a result, further trial-models investigated.

As a first trial, it had been decided to take these 235 cases out, because of their observed certainty of repayment, from the total sample, the remaining number of cases being (1,262 – 235=) 1027 cases. Classification results for these models are provided in Table 3 as well. The total ACC rate applying this trial procedure, with a cut-off point of 50%, were 68.35% (using all variables) and 69.33% (using the 15 predictor variables) ACC rates for $WOE_{T1}$and WOE_{T11}, respectively.[5]

As a second trial, the 235 corporate guarantee case-applications were added back to the sample, but without their sub-corporate guarantee scores. The reason for this is that these scores were very high compared with other sub-independent variable scores, which affect the average cut-off score for the overall model. Then those 235 cases reintroduced as part of the overall sample. Results for these models are shown in Table 3, with 72.11% and 72.98% ACC rates for WOE_{T2} (using all variables) and WOE_{T21} (using 15 variables), respectively,[6] with a 50% cut-off point.

3.1.2. Probit analysis

PA credit scoring models were developed to describe the relationship between the dependent variable (LOAN QUA) and nineteen independent variables. Because the *P*-value for the model in the analysis of deviance table (for more details see Appendix C) is less than 0.01, there is a statistically significant relationship between the variables at the 99% confidence level. In addition, the *P*-value for the residuals is greater than or equal to 0.10, indicating

[3] Furthermore, weak and poor predictive variables (nine variables, five of which weak predictors namely, AGE, DEPE, TELE, FEASI STU, CC STA; and four of which are poor predictors namely, DUM SING, DUM MARR, HOU STA, and CAR OWN) had been excluded from the models; the reason for this, was to investigate the effect of excluding these variables on the ACC rare.

[4] Using only average, strong and very strong predictor variables, 10 variables were also selected with a 52.93% ACC rate with a standard 0.50 cut-off point, and a 76.55% maximum ACC rate with a 0.25 cut-off point. Due to the large number of excluded variables, 9 variables, which might affect the quality of the final fitted model, it was decided that weak, average, strong and very strong predictor 15 variables were used for comparison purposes in this section.

[5] Once again with the 10 predictive variables instead, a 65.24% ACC rate was found.

[6] Alternatively, using only the 10 average, strong, and very strong predictor variables, a 69.33% ACC rate was found.

that the model is not significantly worse than the best possible model for this data at the 90% or higher confidence level.

As shown in Fig. 3, there are many cases of a high probability prediction of good credit, which were confirmed as true (the light coloured boxes), for both PA (on the left-hand side) and PA_1 (on the right-hand side). Where the prediction probability exceeds 0.45 and/or 0.50 for both PA and PA_1, there are a few false results (the dark coloured boxes), i.e. bad credits; and vice versa, i.e. for probabilities of good credits less than 0.50 for PA and less than 0.40 for PA_1, there are more false results, than true results, i.e. more bad credits associated with low predictions of good credits, than good credits associated with low predictions of good credits.

The prediction capability for LOAN QUA describes the relationship between different cut-off points and the per cent correctly classified. As shown in Fig. 4, the middle blue line refers to the true correctly classified. The highest orange line at the lower cut-off rates is the true correctly classified set, while the lowest red line at the lower cut-off rates refers to the falsely classified set, in both PA (on the left-hand side) and PA_1 (on the right-hand side).

All selected variables were significant at the 95% confidence level except five variables: AGE, DUM MARR, DUM SING, PROFE, and TELE (the *P*-value of TELE was 0.0603, at the beginning, and 0.0833 after excluding the four insignificant variables). However, due to their potential importance they were kept in the model. Table 3 reveals an 81.93% ACC rate for this model using a 50% cut-off point. Nevertheless, the highest correct classification per cent was found using both 45% and 50% cut-off points, which is 81.93% (see Appendix D). Hence the model was run again, without AGE, DUM MARR, DUM SING, PROFE, and TELE (calling this the PA_1 model). All included variables were significant at a 95% confidence level, and an 81.62% ACC rate was observed with a cut-off of 50% as it is shown in Table 3. That was the highest ACC rate with both 45% and 50% cut-off points.

3.1.3. Genetic programming models

Two types of GP models are used in this section, program model/evolved program (GP_p), *which is single program*, and team model (GP_t), *which is a combination of single program models in order to produce better results than any of the single program models*. Two samples are used to develop the genetic scoring models, comprising training data (*used for genetic evolution*) and validation data (*used for model selection*).

3.1.3.1. Program model. The GP_p model was designed using the whole dataset divided equally between training and validation samples including all the nineteen predictor variables. Again the overall dataset is used in building the proposed program model for a better comparison of results with all other statistical techniques, as discussed earlier.

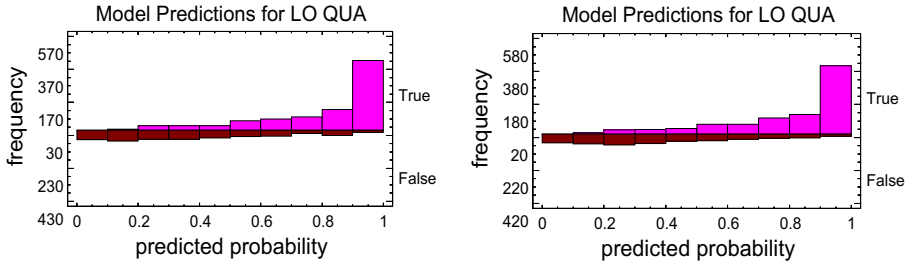

Fig. 3. Model prediction using PA (on the left-hand side) and PA_1 (on the right-hand side) for loan quality.

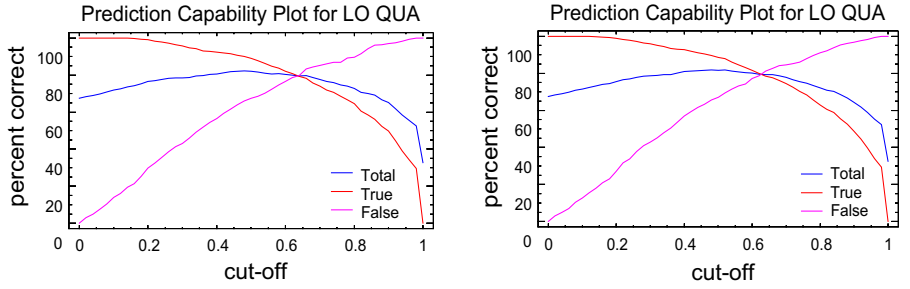

Fig. 4. Prediction capability plot using PA (on the left-hand side) and PA_1 (on the right-hand side) for loan quality.

Table 4
Classification results for the genetic programming, namely, GP_p and GP_t using the whole sample.

Sample model	Training sample				Validation sample				Overall T&V sample			
	G	B	T	Overall %	G	B	T	Overall %	G	B	T	Overall %
GP_p												
G	388	33	421	92.16	394	36	430	91.63	782	69	851	91.89
B	78	132	210	62.86	64	137	201	68.16	142	269	411	65.45
T			631	82.41			631	84.15			1262	83.28
GP_t												
G	381	40	421	90.50	394	36	430	91.63	775	76	851	91.07
B	49	161	210	76.67	54	147	201	73.13	103	308	411	74.94
T			631	85.90			631	85.74			1262	85.82

Table 4 summarizes the best GP_p's classification results. It can be observed that the ACC rates for training and validation samples were 82.41% and 84.15%, respectively. The overall training and validation ACC rate was 83.28%. For the purpose of comparing genetic results with other statistical techniques, both training and validation classification results will be used (*both training and validation samples are used to select the best models*) to produce the whole sample genetic scoring model.

3.1.3.2. Team model. A GP_t model was developed, *as a part of the currently used genetic software design*, following the same procedures as developed using the program model using the whole dataset in terms of training and validation and all the nineteen predictors.

As revealed in Table 4, the ACC rates for the best GP_t model were 85.90% and 85.74% for training and validation samples, respectively. An 85.82% ACC rate was found using the overall training and validation sample, for a team size of seven programs for all samples (training, validation, and training and validation). This achieved a 2.54% increase over the best program model's overall training and validation ACC rate.

As shown in Fig. 5, for a small number of completed runs (on the left-hand side), the overall training and validation performance line of the GP_p (dard red coloured line) is better than the overall training and validation performance line of the GP_t (light green coloured line). For an increased number of completed runs, GP_p and GP_t are changing positions but perform at similar levels. For a high number of completed runs (on the right-hand side), it is clear that the overall training and validation performance line of GP_t is much better than (i.e. fewer missed hits) the overall training and validation performance line of GP_p. This supports the classification results shown in Table 4.

3.2. Sub-sample credit scoring models

The focus here is upon the predictive ability of the scoring models. The main purpose of this section is to investigate whether different results in terms of ACC rates and EMCs, *which will be discussed later*, could be achieved using different sample sizes. It was considered important to make a fair comparison between all proposed scoring models using different statistical techniques, and to reduce and/or avoid sample bias, which might happen in the above whole sample analysis. Hence, a simple validation technique was applied by classifying the whole dataset into a training sub-sample (846 cases) and a testing sub-sample (416 cases) that test the predictive effectiveness of the fitted models. This consists of a 67% training dataset sample and a 33% testing (applied) dataset sample.

3.2.1. Weight of evidence measure

Nineteen variables were used in building the WOE scoring models in this section, based on the training sub-sample only, including five poor independent variables (DUM SING, DUM MARR, HOU STA, TELE and CAR OWN) and three weak independent variables (AGE, DEPE and CC STA). These variables were kept in the final analysis because of their potential importance and for the comparison purposes with other techniques (for more details regarding the IV for all the nineteen variables see Appendix A).

As shown in Table 5, ACC rates were found in the training sub-sample, *for which the data used in building the model*, and testing sub-sample, *for which the data played no role in building the model*, were 52.16% and 55.44%, respectively.

WOE_{T1} and WOE_{T2} trials are developed to test the *sensitivity* of the classification results for WOE scoring models. It can be observed from the results in Table 5 that using WOE_{T1}, *for which all*

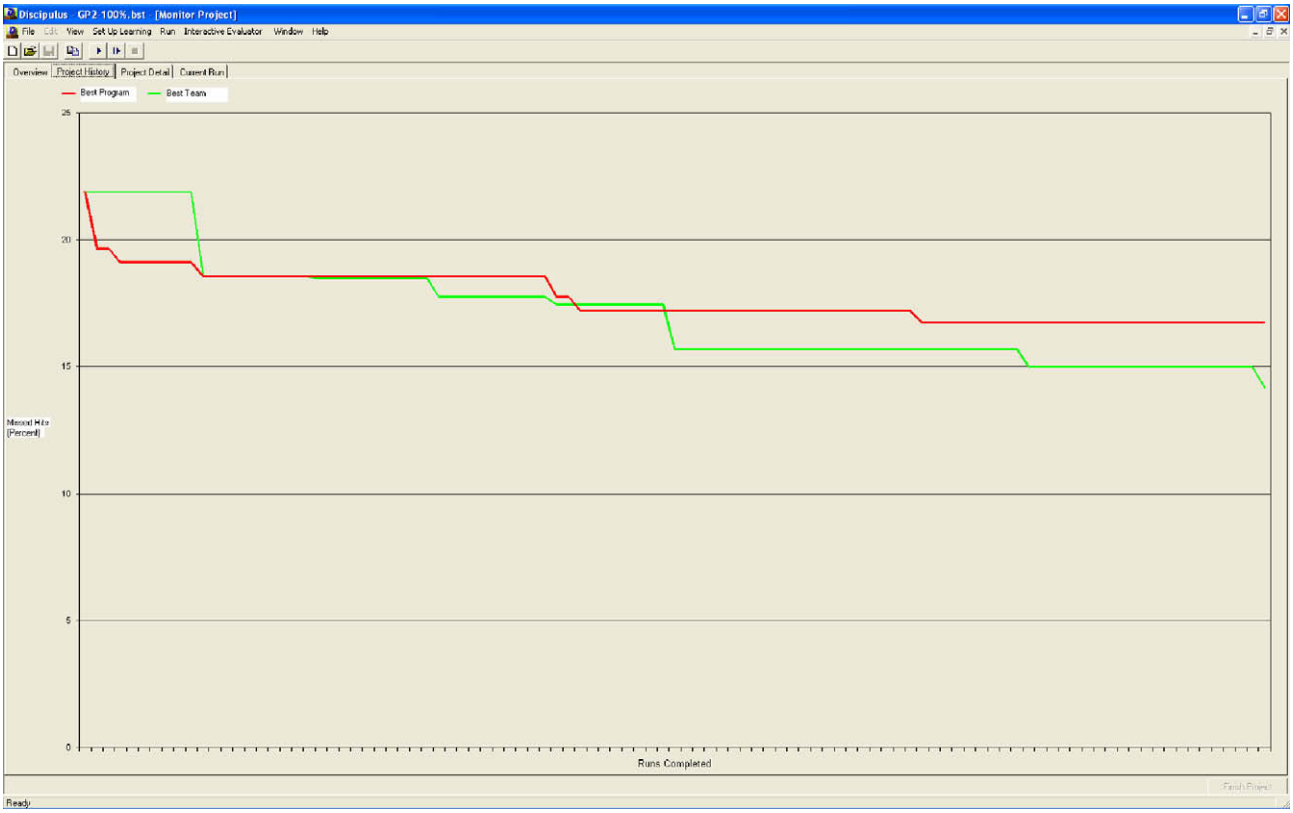

Fig. 5. Genetic best program and best team overall training and validation performance for whole sample.

Table 5
Classification results for the WOEs and PA; predictions (in columns) versus observations (in rows).

SampleModel	Training sub-sample				Testing sub-sample			
	G	B	T	T%	G	B	T	T%
WOE								
G	190	371	561	33.87	91	199	290	31.38
B	6	279	285	97.89	0	126	126	100
T			846	55.44			416	52.16
WOE_{T1}								
G	199	206	405	49.14	155	56	211	73.46
B	50	235	285	82.46	51	75	126	59.52
T			690	62.90			337	68.25
WOE_{T2}								
G	329	232	561	58.65	209	81	290	72.07
B	50	235	285	82.46	35	91	126	72.22
T			846	66.67			416	72.12
PA								
G	499	62	561	88.95	261	29	290	90.00
B	90	195	285	68.42	43	83	126	65.87
T			846	82.03			416	82.69

Cut-off point 0.50.

the 235 corporate guarantee cases in both training sub-sample$_3$(156 cases) and testing sub-sample (79 cases) were excluded from these sub-samples, the ACC rates were 62.90% using the training sub-sample, and 68.25% using the testing sub-sample. The ACC rates for the testing, and the training sub-samples using WOE_{T2}, *for which all the 235 corporate guarantee case scores were not included in the total score but their respective values for the other independent variables were included*, were 72.12%, and 66.67%, respectively.

3.2.2. Probit analysis

Five predictor variables, namely AGE, MON INCO, DUM MARR, PROFE and TELE were not significant at the 90% confidence level. However, due to their potential importance and for the comparison purposes with other techniques, it has been decided to keep them in the final model. Because the *P*-value for the PA model in the analysis of deviance table is less than 0.01, there is a statistically significant relationship between the variables at the 99% confidence level. In addition, the *P*-value for the residuals is greater than or equal to 0.10, indicating that the model is not significantly worse than the best possible model for this data at the 90% or higher confidence level, see Appendix D. As shown in Table 5, an 82.69% ACC rate was found in the testing sub-sample and an 82.03% ACC rate was found in the training sub-sample, using a 50% cut-off point.

3.2.3. Genetic programming models

The same two types of GP models, *which are used in the whole sample models*, are used in this section, namely, the GP_p model and the GP_t model. Three sub-samples are selected to develop the genetic scoring models, namely, sub-samples for the initial training data (*used for genetic evolution*); validation data (*used for model selection*); and applied (testing) data (*played no role in training or model selection*).

Table 7
Classification results for the GP_p and GP_t models; predictions (in columns) versus observations (in rows).

Sample model	Training sub-sample[a]				Applied sub-sample			
	G	B	T	T%	G	B	T	T%
GP_p								
G	497	64	561	88.59	257	33	290	88.62
B	92	193	285	67.72	38	88	126	69.84
T			846	81.56			416	82.93
GP_t								
G	511	50	561	91.09	274	16	290	94.48
B	95	190	285	66.67	51	75	126	59.52
T			846	82.86			416	83.89

[a] Training sub-sample = weighted sum of initial training and validation sub-samples.

3.2.3.1. Program model. The GP_p model, *which is a single program*, was designed to describe the relationship between the independent variables and the dependent variable, using an equally divided dataset between the initial training sub-sample and validation sub-sample datasets, in addition, to the testing sub-sample dataset. Table 6 shows the GP_p (best program model) classification results. The ACC rates for both the initial training sub-sample and validation sub-sample were 82.03% and 81.09%, respectively. It follows that the overall training sub-sample ACC rate (*this rate is a weighted average of the correct classification rates of the initial training and validation sub-samples*) was 81.56, and the testing ACC rate, using GP_p best model, was an 82.93% ACC rate, see Table 7.

3.2.3.2. Team model. The best GP_t model, *which is a combination of single programs in order to produce better results than any of the single programs*, produced an 82.86% ACC rate for the overall training sub-samples. This was a weighted average of the ACC rates for the initial training sub-sample and validation sub-sample at 84.63% and 81.09% ACC rates, respectively, as observed in Table 6. Furthermore, as revealed in Table 7, 83.89% and 82.86% ACC rates were found for the testing (applied) and overall training sub-samples (*for a best team size in the initial training sub-sample of 9 programs, 9 and/or 1 in the validation sub-sample, and 9 and/or 3 in the total training sub-sample, whilst, it was 5 and/or 7 and/or 9 programs in the testing/applied sub-sample*).

As shown in Fig. 6, the overall training performance line for GP_t (light green coloured line) was almost the same as the overall training performance line for GP_p (dark red coloured line) for a

Table 6
Classification results for the GP_p and GP_t models; predictions (in columns) versus observations (in rows).

Sample model	Initial training sub-sample				Validation sub-sample			
	G	B	T	T%	G	B	T	T%
GP_p								
G	239	31	270	88.52	258	33	291	88.66
B	45	108	153	70.59	47	85	132	64.39
T			423	82.03			423	81.09
GP_t								
G	248	22	270	91.85	263	28	291	90.38
B	43	110	153	71.90	52	80	132	60.61
T			423	84.63			423	81.09

Fig. 6. Genetic best program and best team overall training performance for sub-sample.

small number of completed runs (on the left-hand side). At a certain number of completed runs (around one-third of the total completed runs) a switch occurs, at which point the GP_t model performs better than GP_p model. Based on the results revealed in Table 7, the difference between the overall training sub-sample (*both initial training and validation sub-samples*) ACC rates for GP_p and GP_t was 1.30% (i.e. 82.86% for GP_t– 81.56% for GP_p). This outcome enhances the results revealed from Fig. 6.

4. Comparison of results of different credit scoring models and sensitivity analysis of EMCs

In this section, a comparison of different statistical scoring techniques was made, based on the overall sample for the whole sample credit scoring models, and based on the testing sub-samples for the sub-samples credit scoring models. Four different criteria have been used for analytical purposes: firstly, the ACC rate criterion; secondly, EMC with an MC ratio criterion of 5:1; thirdly, EMC with an MC ratio criterion of 7:1; and finally, EMC with an MC ratio criterion of 10:1.

The reason for this was to investigate whether different results may occur by applying different evaluation criteria. Furthermore, *as discussed before*, valid prediction for MCs, associated with both type I and type II errors, might not be available in an environment such as the Egyptian banking sector, and, based on discussions with bank personnel, a high MC ratio might be more appropriate. Tables 8 and 9 summarize different samples' ACC rates, and EMCs under different MC ratios (i.e. 5:1, 7:1, and 10:1 cost ratios).

4.1. Comparison based on whole sample credit scoring models

The classification results for whole sample models are compared for evaluation purposes. Table 8 summarizes the ACC rates and EMCs for the scoring techniques, namely, WOE, WOE_1, PA, PA_1, GP_p, and GP_t. It can be concluded from Table

Table 8
Comparing classification results, errors, and EMCs for the scoring models using whole sample.

Scoring model	Correctly classified results			Error results		EMC	EMC	EMC
	G%	B%	ACC%	Type I	Type II	(5:1)	(7:1)	(10:1)
Whole sample								
WOE[a]	34.08	98.30	54.99	0.6592	0.0170	0.4722	*0.4833*	*0.4999*
WOE_1	33.25	98.30	54.44	0.6675	0.0170	0.4778	0.4889	0.5055
PA	88.95	67.40	81.93	0.1105	0.3260	0.6054	0.8178	1.1363
PA_1	88.60	67.15	81.62	0.1140	0.3285	0.6118	0.8258	1.1468
GP_p	91.89	65.45	83.28	0.0811	0.3455	0.6173	0.8424	1.1800
GP_t[b]	91.07	74.94	*85.82*	0.0893	0.2506	*0.4683*	0.6316	0.8764

[a] Best model amongst all overall sample models based on EMC under MC ratio of 7:1 or above criteria.
[b] Best model amongst all whole sample models based on ACC rate and EMC under MC ratio of 5:1 criteria.

Table 9
Comparing classification results, errors and EMCs for the scoring models using the testing sub-sample.

Scoring model	Correctly classified results			Error results		EMC	EMC	EMC
	G%	B%	ACC%	Type I	Type II	(5:1)	(7:1)	(10:1)
Testing sub-sample								
WOE[a]	31.38	100.00	52.16	0.6862	0.0000	0.4627	*0.4627*	*0.4627*
PA	90.00	65.87	82.69	0.1000	0.3413	0.6232	0.8456	1.1790
GP_p	88.62	69.84	82.93	0.1138	0.3016	0.5679	0.7644	1.0590
GP_t[b]	94.48	59.52	*83.89*	0.0552	0.4048	0.6964	0.9601	1.3557

[a] Best model amongst all testing sub-sample models based on EMC under MC ratio of 5:1 or above criteria.
[b] Best model amongst all testing sub-sample models based on ACC rate criterion.

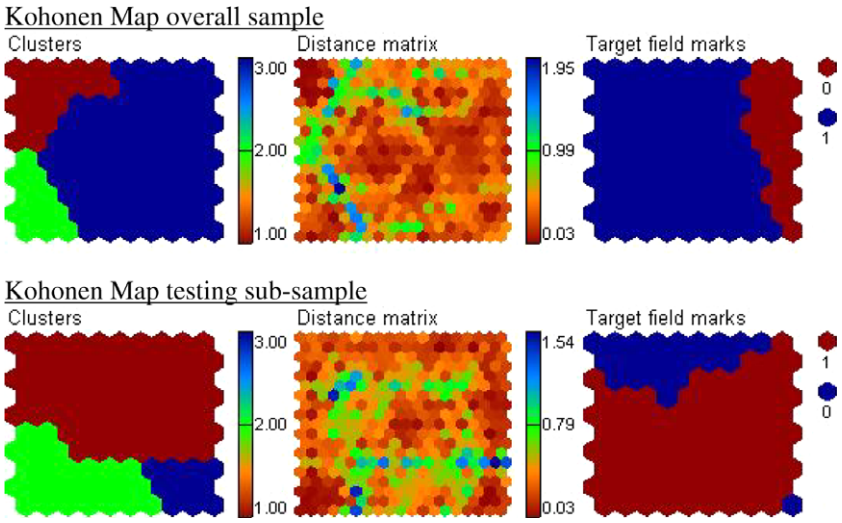

Fig. 7. An analysis of different samples using Kohonen maps. (0 represent the bad cases and 1 represents the good cases.)

8 that GP_t has the highest ACC rate, which is 85.82%, amongst all techniques. All models predict good credit better than bad credit, except the WOE models, namely, WOE, and WOE_1. The reason is that the type I errors in the WOE models are higher than the type II errors. By contrast, PA models predict good credit much better than the WOE models. In addition, the highest correctly classified bad credit was 98.30% for WOE models, whilst the highest correctly classified good credit was 91.89% for GP_p. As shown in Table 8, on average the overall performance of the GP models is better than the overall performance for PA models, but much better than the overall performance for the WOE techniques.

The GP models' type II errors are higher than type I errors. Amongst all models, the lowest EMC with an MC ratio of 5:1 at 0.4683 is for GP_t. That is the chosen model, according to the ACC rate of 85.82% (see Table 8). When using EMC with higher MC ratios, the lowest MCs at 0.4833 and 0.4999 for MC ratios of 7:1 and 10:1, respectively, are for the WOE model. That is not the chosen model according to the ACC rate criterion, which select GP_t at 85.82% ACC rate. Finally, GP_t is the best model according to the ACC rate and the EMC with an MC ratio of 5:1, whilst WOE is the best model under EMC with an MC ratio of 7:1 or above.

4.2. Comparison based on testing sub-sample credit scoring models

The emphasis in this section is upon using a testing sub-sample to test the predictive ability of the scoring models developed in this paper. Table 9 summarizes the ACC rates and EMCs for the scoring techniques, namely, WOE, PA, GP_p, and GP_t. As shown in Table 9, GP_t had the highest ACC rate at 83.89%, amongst all techniques. All models predict good credit better than bad credit, except only one conventional model, namely, WOE. In addition, the highest correctly classified bad credit was 100% for WOE, whilst the highest correctly classified good credit was 94.48% for GP_t. It can be concluded from Table 9, that the average predictive performance of the GP and PA models is better than the average predictive performance of the WOE techniques.

Furthermore, comparing conventional techniques, there is only one model, where the type I error rate exceeds the type II error rate, namely, WOE with the lowest EMC with an MC ratio of 5:1 at 0.4627, among all techniques and an 52.16% ACC rate. Correspondingly, where the type II error rate exceeds the type I error rate, as for the PA and GP models, the lowest EMC with an MC ratio of 5:1 at 0.5679 is for GP_p. This is not the chosen model between PA and GP models, for GP_t had amongst these the highest ACC rate at 83.89%, as shown in Table 9.

Having extended the analysis to include an MC ratio of 7:1, the WOE model is still the best model at the same cost of 0.4627. Finally, the criterion of an MC ratio of 10:1 was also applied. Actually, almost the same result was found using this criterion, namely the WOE was the preferred model amongst all models (see Table 9). Correspondingly, increasing the cost ratios from 5:1 to 7:1 or even 10:1 does not change the decision; and, if WOE model had been neglected from the comparison, GP_p would have been the best under EMC with different MC ratios. But that was not the chosen model under the ACC rate criterion, which was GP_t at 83.89%.

Therefore, the ACC rate criterion led to selecting GP_t; however, this does not provide the lowest EMC, which was found using WOE under different MC ratios.

4.3. Kohonen map analysis

An analysis of the overall sample and testing sub-sample was also investigated using Kohonen maps which were generated to indicate the cluster grouping. A Kohonen map organizes cases according to the topological order within the spatial setting (Melssen, Wehrens, & Buydens, 2006; Yim & Mitchell, 2005). In the process of organising the map an unsupervised neural training takes place. The purpose of its use in this context is to identify visually if the cluster groupings for the different sample sizes bear similarities or not, and to see if a particular sample has more poorly defined cases in terms of good or bad credit. The maps are "self-organized" and not pre-set. The clustering can incorporate several groupings, rather than a binary split. Here three clusters have been used for random data applied under the median clustering method. For a visual analysis of the distances between input and weight vectors, reference is made to the distance matrices; and for an analysis of loan quality as a target, the reader is referred to the target field matrices.

For the overall sample, the light green coloured cluster grouping on the lower left-hand side of the box represents poorly defined cases (neither clearly good nor bad); and the dark blue coloured cluster grouping on the right-hand side represents good cases. However, as the testing sub-sample is used, which comprises 33% of the total dataset, the light green coloured cluster grouping now in the lower left-hand corner has increased to some extent; and the medium dark brown cluster grouping in the upper section of the cluster matrix, which represents good cases, and is slightly smaller than this group size in the overall sample (dark blue on the right-hand side), as shown in Fig. 7.

Indeed, this visual analysis supports the previous comparison between these two samples, concerning the efficiency of the overall sample (under the ACC criterion) and the testing sub-sample (under the EMC with different MC ratios criteria). Correspondingly, the Kohonen analysis indicates that the overall sample is a little better than the testing sub-sample, which might be because of the benefits of the larger dataset. It should be emphasised that results under both samples are close contenders, as demonstrated by the Kohonen visual analysis and credit scoring models using ACC rate and EMC criteria.

5. Conclusion and area for future research

Recently, GP has become one of the most important techniques used in classification and scoring problems, because of its high capabilities in solving different complex problems. Therefore, it can be applied in a wide range of fields, such as, management, finance, banking and logistics.

This paper presents an evaluation of credit scoring models, which are not yet used in practice by the Egyptian public sector banks. Using a consumer loan dataset provided by these banks, an evaluation of credit scoring models is undertaken. To the best of the author's knowledge, no other studies of the Egyptian public sector banks have investigated the use of sophisticated statistical scoring models. The focus in this paper is upon three different statistical scoring techniques, GP, PA and WOE to predict consumer loan quality.

In this paper a range of new variables has been provided which have not been used in other published works, such as CBE report, field visit and feasibility study. Moreover, the ranking of the finally selected model is varies according to the evaluation criteria. Using the ACC rate criterion, GP_t is the preferred model under both whole sample and testing sub-sample. However, using the EMC criterion with an MC ratio of 5:1 the best model for the whole sample was GP_t, whilst the best model for the testing sub-sample was WOE. Finally, using the EMC with an MC ratio of 7:1 or above, the WOE is preferred, for both whole sample and testing sub-sample. Actually, the previous analysis has been extended to include higher MC ratios e.g. 12:1 and 15:1. Under those higher MC ratios the ranking of the decision does not change, and WOE is still the best model. It should be emphasised that the results for a given model under different samples are close, as proved by different evaluation criteria and the analysis of Kohonen maps.

The investigations could be extended to include other financial products, such as, mortgages, house loans and corporate loans. Also, the use of other criteria in evaluating the scoring models, such as, GINI coefficient and area under the ROC curve would be useful. Furthermore, the future plan is to investigate the behaviour of the customers who had defaulted in relation to the timing of the default within the loan period, and determine in particular what variables may affect early default.

Acknowledgement

The author would like to thank Professor John Pointon for comments on the early draft and helpful suggestions on this paper.

Appendix A. IVs for whole sample and sub-sample WOE models

Variables/characteristics	Whole sample			Sub-sample		
	IO	WOE	IV	IO	WOE	IV
LOAN AMO						
Loan Band 1	5.532101	1.710568	0.207487	4.826203	1.574060	0.169058
Loan Band 2	2.414806	0.881619	0.094080	2.407580	0.878622	0.099806
Loan Band 3	1.062515	0.060638	0.000507	0.959596	−0.041243	0.000210
Loan Band 4	0.973840	−0.026508	0.000103	1.104947	0.099797	0.001470
Loan Band 5	1.004559	0.004549	2.52E−06	1.001931	0.001929	4.71E−07
Loan Band 6	0.694786	−0.364151	0.03083	0.720233	−0.328181	0.025450
Loan Band 7	0.430749	−0.842229	0.043161	0.413943	−0.882026	0.048971
Loan Band 8	0.375637	−0.979133	0.040161	0.418371	−0.871388	0.030232
Loan Band 9	0.068994	−2.673729	0.084792	0.050802	−2.979817	0.099243
Loan Band 10	0.087811	−2.432567	0.059388	0.056447	−2.874456	0.085648
$\sum$			**0.560510**			**0.560090**

(*continued on next page*)

Appendix A (*continued*)

Variables/characteristics	Whole sample			Sub-sample		
	IO	WOE	IV	IO	WOE	IV
LOAN DUR						
Duration Band 1	1.7708578	0.571464	0.006431	2.032086	0.709063	0.010271
Duration Band 2	2.6562867	0.976929	0.047243	2.596554	0.954185	0.048108
Duration Band 3	1.3080200	0.268515	0.014489	1.433749	0.360293	0.024675
Duration Band 4	1.2517566	0.224548	0.006740	1.195344	0.178434	0.004158
Duration Band 5	0.8073028	−0.214056	0.020473	0.758600	−0.276281	0.034634
Duration Band 6	0.5083802	−0.676526	0.015375	0.349265	−1.051925	0.038429
Duration Band 7	0.8969280	−0.108780	0.000955	1.270053	0.239059	0.004530
Duration Band 8	0.0804935	−2.519578	0.033821	0.254011	−1.370379	0.007174
Duration Band 9	0.8451821	−0.168203	0.000507	0.653170	−0.425917	0.003628
∑			**0.146035**			**0.175609**
AGE						
Age Band 1	2.2078227	0.792007	0.016292	2.438503	0.891384	0.022496
Age Band 2	1.3374311	0.290751	0.009309	1.593340	0.465832	0.021336
Age Band 3	0.9519236	−0.049271	0.000398	0.971867	−0.028536	0.000130
Age Band 4	0.8418590	−0.172143	0.007220	0.867023	−0.142690	0.004993
Age Band 5	1.0564777	0.054940	0.000604	0.990211	−0.009837	1.993E−05
Age Band 6	0.8774929	−0.130686	0.002766	0.803597	−0.218657	0.008288
Age Band 7	0.9337250	−0.068573	0.000332	1.072490	0.069983	0.000320
Age Band 8	1.8513514	0.615916	0.007655	1.219251	0.198237	0.000763
∑			**0.044576**			**0.058345**
DUM SING						
Single	1.1913043	0.1750488	0.004889	1.201904	0.183907	0.005342
Other	0.9672984	−0.0332482	0.000929	0.966073	−0.034515	0.001003
∑			**0.005817**			**0.006344**
DUM MARR						
Married	0.9702734	−0.0301773	0.000729	0.959596	−0.041243	0.001368
Other	1.1269095	0.1194789	0.002878	1.185383	0.170066	0.005642
∑			**0.003604**			**0.007010**
GENDER						
Male	0.8118970	−0.2083818	0.035287	0.819135	−0.199506	0.032665
Female	2.6975151	0.9923310	0.168040	2.728263	1.003665	0.164330
∑			**0.203327**			**0.196995**
DEPE						
Dependant Band 1	1.137099	0.128480	0.003386	1.075810	0.073074	0.000991
Dependant Band 2	1.462299	0.380010	0.030776	1.545682	0.435465	0.039187
Dependant Band 3	0.906542	−0.098118	0.002722	0.817024	−0.202087	0.012585
Dependant Band 4	0.738930	−0.302551	0.019218	0.849348	−0.163286	0.005524
Dependant Band 5	0.810129	−0.210562	0.003016	0.989305	−0.010753	7.667E−06
Dependant Band 6	0.965922	−0.034672	1.725E−05	0.762032	−0.271767	0.001362
Dependant Band 7	0.482961	−0.727819	0.000916	0.508021	−0.677232	0.001169
∑			**0.060050**			**0.060826**
PROFE						
Public Sector	2.049135	0.717418	0.128192	2.182183	0.780326	0.142419
Private Sector	0.784635	−0.242537	0.043338	0.784166	−0.243135	0.044375
∑			**0.171529**			**0.186794**
EDUC						
Before University	0.523056	−0.648067	0.199292	0.527140	−0.640288	0.197595
University or/and Higher	1.865686	0.623629	0.191777	1.888403	0.635731	0.196189
∑			**0.391070**			**0.393784**

Appendix A (*continued*)

Variables/characteristics	Whole sample			Sub-sample		
	IO	WOE	IV	IO	WOE	IV
HOU STA						
Owned	1.033945	0.033382	0.000587	1.101305	0.096496	0.004905
Rented	0.963483	−0.037200	0.000654	0.897981	−0.107606	0.005470
∑			**0.001242**			**0.010375**
TELE						
Yes	1.092872	0.088809	0.007024	1.059410	0.057713	0.002960
No	0.467126	−0.761155	0.060198	0.625257	−0.469592	0.024081
∑			**0.067222**			**0.027041**
MON INCO						
Income Band 1	1.696556	0.528600	0.034939	1.687357	0.523163	0.035329
Income Band 2	1.400588	0.336892	0.016418	1.354724	0.303598	0.013603
Income Band 3	0.980783	−0.019404	5.897E−05	1.052330	0.051007	0.000393
Income Band 4	1.462299	0.380010	0.015388	1.320856	0.278280	0.009399
Income Band 5	0.990071	−0.009979	1.447E−05	0.812834	−0.207228	0.006124
Income Band 6	0.608173	−0.497295	0.038402	0.615487	−0.485341	0.034050
Income Band 7	0.663048	−0.410907	0.019876	0.747920	−0.290459	0.009249
Income Band 8	0.505959	−0.681299	0.017198	0.603275	−0.505381	0.011256
∑			**0.142293**			**0.119403**
CBE REP						
Positive	0.744058	−0.295636	0.065724	0.753704	−0.282756	0.059623
Not Required	2.692062	0.990307	0.220160	2.465762	0.902501	0.190305
∑			**0.285884**			**0.249928**
GUAR						
Corporate Guarantee	113496.2	11.63952	3.214176	79251.62	11.28038	3.136751
Own Guarantee	0.723856	−0.323163	0.089239	0.721928	−0.325830	0.090604
∑			**3.303416**			**3.227356**
FIE VISI						
Positive	0.654616	−0.423706	0.118212	0.623580	−0.472278	0.145339
Not Required	2.451487	0.896695	0.250174	2.686652	0.988296	0.304138
∑			**0.368387**			**0.449477**
FEASI STU						
Positive	0.772738	−0.257815	0.032076	0.710006	−0.342482	0.057848
Not Required/Available	1.274914	0.242879	0.030217	1.404530	0.339703	0.057379
∑			**0.062293**			**0.115227**
CC STA						
Yes	0.735073	−0.307785	0.031545	0.778345	−0.250586	0.021243
No	1.167156	0.154570	0.015842	1.137275	0.128635	0.010905
∑			**0.047387**			**0.032148**
LFOB						
Yes	2.643796	0.972216	0.451052	2.332280	0.846846	0.348368
No	0.353626	−1.039516	0.482276	0.404870	−0.904190	0.371957
∑			**0.933328**			**0.720325**
CAR OWN						
Yes	0.877689	−0.130463	0.008076	1.008105	0.008072	2.938E−05
No	1.125323	0.118070	0.007308	0.993392	−0.006630	2.413E−05
∑			**0.015384**			**5.352E−05**

IO = G%/B%; WOE = LN (IO) and IV = [(G%-B%) × WOE].

Appendix B. ACC rates with different cut-off points using the whole sample for conventional techniques, namely, WOE, WOE_1, PA, and PA_1

Model cut-off	WOE%	WOE_1%	PA%	PA_1%
0.05	67.43	67.43	69.18	69.33
0.10	67.43	67.43	71.95	71.55
0.15	67.91	67.91	73.85	73.53
0.20	69.41	69.57	76.47	75.52
0.25	74.64	75.28	78.05	77.18
0.30	75.99*	76.94*	78.53	78.61
0.35	70.21	69.81	79.79	79.32
0.40	63.63	63.39	80.67	81.06
0.45	59.11	58.64	81.93*	81.62*
0.50	54.99*	54.44*	81.93*	81.62*
0.55	52.46	52.30	80.51	81.06
0.60	51.27	51.27	80.35	80.11
0.65	51.19	51.19	79.71	79.08
0.70	49.05	49.05	77.65	77.81
0.75	47.39	47.54	75.36	75.20
0.80	45.40	45.48	72.66	72.03
0.85	43.58	43.74	69.65	69.26
0.90	38.51	38.27	65.13	64.42

Percentages in cells refer to the ACC rates under the different cut-offs. The 0.50 standard cut-off rates and the highest rates per model are asterisked.

Appendix C. Statistical analysis using whole sample for conventional models, namely, PA and PA_1

Analysis of Deviance and Likelihood Ratio Tests for PA model:
Analysis of Deviance

Source	Deviance	Df	*P*-value
Model	612.539	19	0.0000
Residual	980.287	1242	1.0000
Total (corr.)	1592.83	1261	

Likelihood Ratio Tests

Factor	Chi-square	Df	*P*-value
AGE	0.678145	1	0.4103
DEPE	11.4276	1	0.0007
LOAN AMO	13.1637	1	0.0003
LOAN DUR	6.36327	1	0.0116
MON INCO	5.25069	1	0.0219
CAR OWN	8.1677	1	0.0043
CBE REP	7.49171	1	0.0062
CC STA	24.8939	1	0.0000
DUM MARR	0.16095	1	0.6883
DUM SING	0.607906	1	0.4356
EDUC	62.0485	1	0.0000
FEASI STU	13.2553	1	0.0003
FIE VISI	18.2358	1	0.0000
GUAR	77.5347	1	0.0000
HOU STA	19.7139	1	0.0000
LFOB	174.033	1	0.0000
PROFE	1.43227	1	0.2314
GENDER	9.52929	1	0.0020
TELE	3.52892	1	0.0603

Appendix C (*continued*)

Source	Deviance	Df	*P*-value
Model	605.527	14	0.0000
Residual	987.3	1247	1.0000
Total (corr.)	1592.83	1261	

Likelihood Ratio Tests

Factor	Chi-square	Df	*P*-value
DEPE	9.98094	1	0.0016
LOAN AMO	13.6965	1	0.0002
LOAN DUR	5.04639	1	0.0247
MON INCO	4.19389	1	0.0406
CAR OWN	7.24711	1	0.0071
CBE REP	7.25321	1	0.0071
CC STA	21.422	1	0.0000
EDUC	70.8587	1	0.0000
FEASI STU	12.4255	1	0.0004
FIE VISI	24.8907	1	0.0000
GUAR	77.176	1	0.0000
HOU STA	21.3615	1	0.0000
LFOB	173.172	1	0.0000
GENDER	9.72741	1	0.0018

Appendix D. Statistical analysis using the training sub-sample for conventional PA model

Analysis of Deviance and Likelihood Ratio Tests for PA model:
Analysis of Deviance

Source	Deviance	Df	*P*-value
Model	407.816	19	0.0000
Residual	673.277	826	1.0000
1081.09	845		

Likelihood Ratio Tests

Factor	Chi-square	Df	*P*-value
AGE	0.716954	1	0.3971
DEPE	6.32275	1	0.0119
LOAN AMO	13.8682	1	0.0002
LOAN DUR	3.7968	1	0.0513
MON INCO	1.54748	1	0.2135
CAR OWN	11.1242	1	0.0009
CBE REP	7.00026	1	0.0081
CC STA	17.2319	1	0.0000
DUM MARR	1.09222	1	0.2960
DUM SING	2.71392	1	0.0995
EDUC	39.9503	1	0.0000
FEASI STU	3.48637	1	0.0619
FIE VISI	13.6749	1	0.0002
GUAR	52.9741	1	0.0000
HOU STA	26.8494	1	0.0000
LFOB	104.324	1	0.0000
PROFE	1.32995	1	0.2488
GENDER	9.17582	1	0.0025
TELE	0.653076	1	0.4190

References

Abdou, H. (2009). An evaluation of alternative scoring models in private banking. *The Journal of Risk Finance, 10*(1), 38–53.

Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

Bailey, M. (2001). *Credit scoring: The principles and practicalities*. Kingswood, Bristol: White Box Publishing.

Banasik, J., Crook, J., & Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society, 54*(8), 822–832.

Central Bank of Egypt (2006/2007). Banking sector reform. Annual report.

Central Bank of Egypt (2007/2008). Banking developments, lending activity. Economic Review 48(2), 22–30.

Chen, M., & Huang, S. (2003). Credit scoring and rejected instances reassigning through evolutionary computation techniques. *Expert Systems with Applications, 24*(4), 433–441.

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research, 95*(1), 24–37.

Deschaine, L. & Francone, F. (2008). Comparison of DiscipulusTM linear genetic programming soft-ware with support vector machines, classification trees, neural networks and human experts. White paper. Available at: <http://www.rmltech.com/> Accessed 10.06.08.

Elliott, R., & Filinkov, A. (2008). A self tuning model for risk estimation. *Expert Systems with Application, 34*(3), 1692–1697.

Etemadi, H., Rostamy, A., & Dehkordi, H. (2009). A genetic programming model for bankruptcy prediction: Empirical evidence from Iran. *Expert Systems with Applications, 36*(2), 3199–3267.

Golgberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning, reading*. Boston, MA: Addison-Wesley.

Greene, W. (1998). Sample selection in credit-scoring models. *Japan and the World Economy, 10*(3), 299–316.

Guillen, M. & Artis, M. (1992). Count data models for a credit scoring system. In *The European conference series in quantitative economics and econometrics on econometrics of duration, count and transition models, Paris*.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523–541.

Huang, C., Chen, M., & Wang, C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*(4), 847–856.

Huang, J., Tzeng, G., & Ong, C. (2006). Two-stage genetic programming (2SGP) for the credit scoring model. *Applied Mathematics and Computation, 174*(2), 1039–1053.

Koza, J. R. (1992). *Genetic programming on the programming of computers by means of natural selection*. Cambridge, MA: MIT Press.

Koza, J. R. (1994). *Genetic programming II automation discovery of reusable programs*. Cambridge, MA: MIT Press.

Lee, T., & Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications, 28*(4), 743–752.

Lee, T., Chiu, C., Lu, C., & Chen, I. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), 245–254.

Lensberg, T., Eilifsen, A., & McKee, T. (2006). Bankruptcy theory development and classification via genetic programming. *European Journal of Operational Research, 169*(2), 677–697.

Lim, M. K., & Sohn, S. Y. (2007). Cluster-based dynamic scoring model. *Expert Systems with Applications, 32*(2), 427–431.

Maddala, G. S. (2001). *Introduction to econometrics*. Chichester: John Wiley & Sons Inc..

Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega, 31*(2), 83–96.

McKee, T., & Lensberg, T. (2002). Genetic programming and rough sets: A hybrid approach to bankruptcy classification. *European Journal of Operational Research, 138*(2), 436–451.

Melssen, W., Wehrens, R., & Buydens, L. (2006). Supervised Kohonen networks for classification problems. *Chemometrics and Intelligent Laboratory Systems, 83*(2), 99–113.

Mukkamala, S., Vieira, A. & Sung, A. (2008). Model selection and feature ranking for financial distress classification. Available at: <http://www.rmltech.com/> Accessed 10.06.08.

Nunez-Letamendia, L. (2002). Trading systems designed by genetic algorithms. *Managerial Finance, 28*(8), 87–106.

Oldham, M. & Benaddi, O. (2005). The Egyptian banking sector: The long road to reform. Special report. Fitch Inc., Fitch Ratings Ltd. and its subsidiaries.

Oldham, M. & Young, M. (2004). Egypt: Difficult times remain for the banking sector. Special report. Fitch Inc., Fitch Ratings Ltd. and its subsidiaries.

Ong, C., Huang, J., & Tzeng, G. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications, 29*(1), 41–47.

Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications, 36*(1), 2–17.

Pindyck, R. S., & Rubinfeld, D. L. (1997). *Econometric models and economic forecasts*. McGraw-Hill/Irwin.

Siddiqi, N. (2006). *Credit risk scorecards: Developing and implementing intelligent credit scoring*. New Jersey: John Wiley & Sons Inc..

Teller, A., & Veloso, M. (2000). Internal reinforcement in a connectionist genetic programming approach. *Artificial Intelligence, 120*(2), 165–198.

Tsai, C., & Wu, J. (2008). Using neural networks ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 34*(4), 2639–2649.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11–12), 1131–1152.

Xia, Y., Liu, B., Wang, S., & Lai, K. K. (2000). A model for portfolio selection with order of expected returns. *Computers & Operations Research, 27*(5), 409–422.

Yang, Z., Wang, Y., Bai, Y., & Zhang, X. (2004). Measuring scorecard performance. *Computational Science, 3039*, 900–906.

Yim, J., & Mitchell, H. (2005). Comparison of country risk models: Hybrid neural networks, logit models, discriminant analysis and cluster techniques. *Expert Systems with Applications, 28*(1), 137–148.

Zhang, Y., & Bhattacharyya, S. (2004). Genetic programming in classifying large-scale data: An ensemble method. *Information Sciences, 163*(1–3), 85–101.