

EVALUATION OF EMPIRICAL ATTRIBUTES FOR CREDIT RISK FORECASTING FROM NUMERICAL DATA

Stelios E. Papadakis^{1*} and Christos Lemonakis²

^{1,2}Dept. of Business Administration, T.E.I. of Crete, Lakonia, Aghios Nikolaos, Crete, Greece

*e-mail: spap@staff.teicrete.gr

ABSTRACT

In this paper, we evaluate 35 features which are empirically utilized for forecasting the credit behavior of the borrowers of a Greek Bank. These features are initially selected according to universally accepted criteria. A data set of historical data (observations) was collected from the database of a Greek bank. Based on those data, we performed extensive data analysis by using non parametric models. Our data analysis revealed that building a simplified model by using only 3 out of the 35 initially selected features can achieve the same or slightly better forecasting accuracy when compared to the forecasting accuracy achieved by a model which uses all the 35 features. Extensive interpretation of the results is provided and the experimentally verified claim that universally accepted criteria can't be globally used to achieve optimal results is discussed.

Keywords: Credit Risk, Computational intelligence.

1. Introduction

Banking activity is exposed to various risks. Understanding and quantifying these risks is crucial for bank management, as well as, for the stability of the whole economy [14]. Banks tend to lend to firms with high credit quality and not to lend to low credit quality firms. So the most important factor in determining lending practices is credit risk [6], [11].

According to [8], research considering credit risk has been one of the most active areas of recent financial research, with significant efforts deployed to analyze the meaning, role, and influence of credit ratings. Credit risk analysis has attracted much attention from financial institutions due to the recent financial crises and regulatory concerns of Basel II [19]. Furthermore, business competition for obtaining more market share and profit become more and more aggressive in recent years, some institutions take more risks to achieve competitive advantage in the market. Consequently, many financial institutions suffered a great loss from a steady increase of defaults and bad loans from their counterparties. However, more and more adult population use credit products, such as mortgages, car and house loan, credit card, etc., from banks or other financial institutions. Therefore, an effective credit risk analysis model has been a crucial factor for describing thoroughly the real credit risks of the selected a bank's loans portfolio.

Generally, the procedures for customer credit risk analysis can be simply viewed as two stages. First, when applicants apply for credit, the lenders must make a decision whether or not to grant the credit and how much to grant. The traditional method of making such decisions is based on experience of previous lending decision. However, with the increase of the number of applicants and the intense competition in credit industry, this traditional method cannot meet the demands of both economic and efficiency aspects for financial institutions.

Nowadays, credit scoring is a widely used technique that helps the lenders to make such credit granting decisions. Its main idea is to evaluate the probability that how likely the applicant will default according to the characters recorded in the application form with a quantitative model based on information of the past applicants, and the accept and reject decision is taken by comparing the estimated default probability with a proper threshold. In the second stage, the lenders need to make the decisions how to deal with the existing customers. When and how to increase and reduce the customers' credit? If the customer starts to fall behind in his repayments (i.e. overdue debts), when and what actions should be taken? How the customer should be treated with regard to retain debt's viability? Techniques that help with these decisions are called as behavior scoring.

The principle of this approach is exactly the same as credit scoring but using more information which describes the customer's performance during some previous observation periods.

The paper is organized as follows. In Section 2, the framework of key steps in a Bank's loan decision making process is described as well as best lending practices to highlight risks undertaken in a loan's evaluation stage. In Section 3, areas of business analysis are set out, to find out customers' financial position and their «particular» operational characteristics. In Section 4, the data sample is described, with the variables used and their specific attributes. In Section 5, the selection of significant inputs is set, as well as the implementation of various selected attributes approaches made, i.e. Pearson's r , Spearman's ρ , Kendall's τ , PCA-SPPS and PCA-[16] and multivariate attributes evaluation. Also, discussion and interpretation of the results are presented. Finally, in Section 6, conclusions are presented as well as further research.

2. Lending decisions based on best-practice

Borrowers such as Small Business Enterprises (SMEs), on a basic model of rating, are divided into two general categories: I) the consistent and II) the borrowers with overdue and problems in repaying their loan obligations. The adoption of this approach, which is consistent with the principles of Basel II, the Standardized Approach and the Internal Ratings Approach for the measurement of credit risk, places emphasis towards the calculation of the expected probability of default (Probability of Default-PD) for each of the categories of loans, taking into account customers' historical default data. The main point, however, for each bank is to find out the right lending decision in evaluation of credit provided to legal persons. The loan evaluation process is a common process, regardless of the bank organization. However, nowadays, due to the economic crisis, further clarifications of individual characteristics are set in identifying those elements that require further evaluation.

Table 1 describes some basic steps of the lending decision procedure, taking into account best practices in the banking industry worldwide. Each bank, initially receives (Step 1) customer's loan application describing the purpose of lending and its characteristics, namely the repayment rate, collaterals provided, etc.

Table 1: Key steps in a Bank's loan decision - best practices

Steps- Procedures	Description	Potential Risks
Step 1 Select the appropriate loan product	Choosing an appropriate loan product that meets customer requirements	Incorrect loan product that does not cover the actual customer needs. Wrong product pricing and increased probability of default
Step 2 Customer's Rating	The bank carries out an initial assessment of the of the borrower's creditworthiness.	Failure to take the real customer's financial status. It results in overestimation or underestimation of the customer's economic capacity.
Step 3A Positive lending decision (+)	The bank decides to provide credit to customer, based on the evaluation of its overall financial status	Possible failure to take adequate guarantees The bank incorrectly determines the individual loan's characteristics, such as the interest rate, the repayment period, etc.
Step 3B Negative lending decision (-) Reassessment	The bank re-evaluates the second stage due to low credit borrower's quality, inappropriate or little collaterals, etc.	Underestimation of the borrower's real financial status.
Step 4A positive lending decision (+) collaterals' strengthening	Conduct borrower's reassessment in the second stage, with a requirement for collaterals' strengthening	Select loan volume, taking into account the personal and collateral guarantees and not the primary ability to repay the loan.
Step 4B Negative Judgment (-)	Reject loan's request	Rejection of a good loan's request, which would have been paid without problems. Loss of bank's profits and adverse consequences for the borrower himself.

Table 1 highlights the risks undertaken by the bank in the evaluation stage, since the repayment of the credit is purely a quantitative-financial, as well as, a qualitative customer's analysis issue (eg.: solvency, borrower's market position, other loans' quality characteristics, etc.) The final credit decision must be varied in acceptable limits carefully defined per bank for each category of loan product. The described framework decision on a bank's loan, which takes into account the best global practice, is for the bank to adopt a flexible strategy that provides adequate and comprehensive information for customer's credit, minimizing credit risk.

3. Customers' financial status analysis

The preparation of information for the decision of credit undertaking is probably the most expensive part of the process of the credit. Table 2 shows the results of assessing the customer's creditworthiness that determines the level of the credit risk undertaken by the bank. The main issue for the credit decision is to determine the client's financial position.

The use of the standard assessment model, depicted in table 2, has common features with the systems operating in banks nowadays. For example, to determine the position of an SME there is a need for using various characteristics (factors), which show its financial position and its «particular» operational characteristics. These factors are grouped in a practical way in nine areas of the baseline analysis (Table 2), i.e. four quantitative and five qualitative characteristics.

Table 2: Areas of business analysis

Number	Quantitative Factors	Qualitative Factors
1	Profitability	Firm's position in the market (in case of business loans)
2	Financial liquidity	Management Quality
3	Business activity	Reliability of businessmen
4	Level of Short and long term debt	Nature of Business
5	-	Historical customer's credit data

The quantitative characteristics using only financial data for calculating relevant financial indicators of profitability, liquidity, activity and customer's bank lending. On the other hand, qualitative factors analysis is called assessment of objective factors. These qualities do not show causal relationships; these factors are measured only in a subjective manner, based on solid criteria.

4. Problem definition

A Greek bank provides loans as products to borrowers. The bank keeps historical records of the behavior of past borrowers. Each record corresponds to a specific borrower and includes a number of measured attributes of him and the payoff status of the loan he received. These attributes, namely candidate attributes or candidate inputs are in advanced decided by an expert in the field of application. This decision is based either on expert's experience and/or his intuition and involves the factors affecting the output of the system according to his intuition/experience. In our case, the predetermined attributes are listed in Table 3.

4.1 Sample and core characteristics

We use quantitative and qualitative data from 150 borrowers of a Greek banking institution in Retail and Corporate Banking. The data used include in particular:

- **The Credit Risk Rate (CRR)** that takes values 0 for low, 1 for acceptable, 2 for acceptable with caution, and 3 for high credit risk.
- **The Exposure at Default (EAD)** and is defined as the gross exposure upon default of an obligor
- **Probability of default (PD)** describing the likelihood of a default over a particular time horizon. It is the likelihood that a borrower will be unable to meet its debt obligations.
- **Expected Loss (EL)** is the average credit loss that a bank would expect from an exposure or a portfolio over a given period of time. Expected Loss (EL) is estimated as the product of Exposure at Default (EAD), Probability of default (PD) and Loss Given Default (LGD). Herein we take a regulatory estimate for LGD equals to 45% under the Basel Committee (2006) requirements.
- **Variables x5-x24** are financial accounts and indicators (quantitative variables), taken out from borrowers' financial statements.
- **Qualitative data**, a binary variable that takes value 1 for satisfactory and 0 for non-satisfactory qualitative characteristics
- **Customer's Characterization**, a variable that takes discrete values 1 for very credible, 2 for credible, 3 for satisfactory, 4 for adequate, and 5 for inadequate customers.
- **The weighted average of annual interest rate**
- **Collaterals** taken for loans' guarantees in euros
- **The Loan to Value Ratio (LTV)** is used by banks to express the ratio of a loan to the value of an asset purchased
- **Obligor type**, that takes the value of 0 for Retail, and 1 for Corporate Borrower
- **Collateral type** that takes the values of 1 for urban property, 2 for commercial property, 3 for other property, and 4 for None
- **Loans' Maturity** that takes the values of 0 for short-term, 1 for mixed term and 2 for long term debt
- **Firms' Exporting Activity**, that takes the value 0 for domestic activity, and 1 for domestic and exported activity
- **Firms' Sectors** that take the values 1 for Agricultural, 2 for Constructions, 3 for Hotels & Tourism, 4 for Manufacture, 5 for Other Industries, 6 for Retail Commerce, 7 for Services, and 8 for Wholesale.
- **Firm's Legal Form**, that takes the values 1 for Sole Proprietorship, 2 for Cooperative, 3 for Municipal Public Companies, 4 for Private Limited Company, 5 for Limited Partnership, 6 for Limited Liability, and 7 for Unlimited Companies.

Table 3: Description of the measured attributes

Attribute	Description
x_1	Credit Risk Rate (CRR), 0=low, 1=acceptable, 2=acceptable with caution, 3=high.
x_2	Exposure at Default (EAD).
x_3	Probability of Default (PD).
x_4	Expected Loss $EL = EAD \cdot PD \cdot LGD$ regulatory estimate for $LGD = 45\%$ Basel Committee (2006) [1]
x_5	Fixed Assets (FA).

x_6	Equity Capital (EQU).
x_7	Bank Loans (BLs).
x_8	Total Sales (TS).
x_9	Cash Deposits (CD).
x_{10}	Gross Profit Margin (GPM).
x_{11}	Net Profit Margin (NPM).
x_{12}	Return on Total Assets (ROA).
x_{13}	Return on Equity (ROE).
x_{14}	Asset Turnover Ratio (ATR).
x_{15}	Sales to Equity Ratio (SER).
x_{16}	Fixed-Asset Turnover Ratio (FATR).
x_{17}	Debt-to-Equity Ratio (DER).
x_{18}	Liquidity Ratio (LR).
x_{19}	Acid Liquidity Ratio (ALR).
x_{20}	Inventory Turnover (IT).
x_{21}	Working Capital Turnover Ratio (WCTR).
x_{22}	Debt-to-Equity Ratio (DER).
x_{23}	Debt Ratio (DR).
x_{24}	Debt to Capital Ratio (DCR).
x_{25}	Qualitative elements (1=satisfactory qualitative and 0=non-satisfactory qualitative).
x_{26}	Customer's Characterization (1=very credible customer, 2=credible, 3=satisfactory, 4=adequate, 5=inadequate).
x_{27}	Average rate of interest.
x_{28}	Collaterals (in euros).
x_{29}	The Loan-to-Value (LTV) Ratio
x_{30}	Obligor type, 0=Retail, 1=Corporate.
x_{31}	Collateral_type (1=urban property, 2= commercial property, 3= other property, 4=None).
x_{32}	Loans' Maturity (0=short-term, 1=mixed term 2=long term).
x_{33}	0=domestic activity, 1=domestic and exported activity.
x_{34}	Sectors: 1=Agricultural, 2=Constructions, 3=Hotels \& Tourism, 4=Manufacture, 5=Other Industries, 6=Retail Commerce, 7=Services, 8=Wholesale.
x_{35}	Firm's Legal Form, 1=Sole Proprietorship, 2=Cooperative, 3=Municipal Public Companies, 4=Private Limited Company, 5=Limited Partnership, 6=Limited Liability, 7=Unlimited Company.

The aforementioned attributes are summarized in Table 3. The sample consists of 150 borrowers, solely firms that keep “C” category accounting books (in the Greek National Accounting System) and categorized based on Basel II criteria, as exposures to retail or corporate banking; in that case such loans are covered by mortgages on commercial or residential property and cash collaterals. The CRR is the result of calculating the degree of credit risk using a commercial software and takes values between 0.07 (AA +) and 1 (CCC) of Standard &

Poor's rating scale. These values correspond to the respective levels of credit risk assessment classification ranging from «Minimum» to «High» credit risk respectively. Regarding the used financial ratios and variables taken into account the average values as derived from firms' published financial statements for the three consecutive years 2009-2011.

Regarding the qualitative customer data (QUAL), exported from the total «financial» status of the borrower. We enter the value 1 for the category "Satisfactory Quality Data" when all firms' qualitative data are positively evaluated (eg: quality of cooperation, professionalism, successors continuity, market position, etc.) and account for more than 60% of the total information taken and we put the value 0 where customer's quality data range in less than 60% of the total information.

Also, regarding the due borrowers we take into account debts during the last year (2011) of the study in relation to the reference period (01/01/2009 - 12/31/2011). We enter a value of 1 for non-existence of a customer's overdue debt as well as other information taken from the Tiresias system, the main Greek Default Financial Obligations & Mortgages and Prenotations to Mortgages System that contains data concerning bounced checks, unpaid bills of exchange, mortgages and prenotations to mortgages. In contrast, we enter 0, where borrowers show due debt greater than 90 days or unfavorable data in Tiresias system.

Finally, with respect to the variable associated with the «maturity» in firms' loans we take into account the repayment period and a customers' separation in those with short-term, mixed term and long term lending. Thus, loans with annual recycling capital and interest (e.g. credit limits using overdrafts) are accounted for as short-term lending and we enter the value 0. Where borrowers receive long-term funding (usually more than two years), we enter the value 2 and in intermediate cases, where borrowers have and short and long-term debt, we enter the value 1. Long-term loans are considered as the highest ones in terms of credit risk by the bank, due to time horizon of repayment that may be changed significantly during the years to come.

The usefulness of economic fundamentals and financial indicators of a company as explanatory variables for the assessment of credit risk has been shown in various studies, including [7],[2],[9] among others. The choice of these variables was based on this literature and the validity of those financial indicators. The methodological framework for the variables draws evidence from both the hybrid creditworthiness model of [2] and from key characteristics of Risk Calc and KMV EDF Risk Calc (v.3,1) Software of [13].

4.2 Mathematical formulation

The Bank aims to predict the payoff status of a future borrower, provided that his attributes are given. This problem can be formulated as a classification problem. Each past borrower is an instance consisting of an input vector of his attributes and a label $\{\pm 1\}$ which denotes that the borrower was trustworthy (+1) or not (-1). A function that assigns a label to an input vector can be constructed from the given database, by computational intelligence techniques. From that point of view the prediction of the behavior of a new customer can be formulated as a typical classification problem according to the following mathematical formulation.

Let $\mathbf{X} \subseteq \mathbb{R}^m$ be the set of ordered vectors $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,m}]$ $i = 1, 2, \dots, n$. Each vector corresponds to a particular borrower, encoding his measured attributes as real numbers. If n borrowers are available and m attributes per borrower are recorded, then the dimensions of each vector equals m and the cardinality of $\mathbf{X}$ equals. Let $\mathbf{L} = \{\pm 1\}$ the set of labels that encodes the behavior of the customers. We assume that a label +1 denotes a trustworthy borrower, which in turn means that the borrower fully repaid the entire amount, timely. Controversially, a label -1 denotes that the respective borrower was inconsistent in repaying the loan. The modeling process aims to build a function $f: \mathbf{X} \rightarrow \mathbf{L}$, which assigns a label $\ell \in \{\pm 1\}$ to a given input datum $\mathbf{x} \in \mathbf{X}$. In matrix notation a two dimensional $n \times m$ matrix $\mathbf{X}$ stores the attribute values of historic customers, while an $n \times 1$ matrix $\mathbf{L}$ stores the respective labels. The whole historic data set can be stored in a matrix $\mathbf{D} = [\mathbf{X} | \mathbf{L}]$ consisting of the attributes matrix $\mathbf{X}$ augmented by $\mathbf{L}$. Each row of $\mathbf{D}$ corresponds to a specific borrower encoding both his attributes and his behavior. In terms of computational intelligence a model, considered as black-box can be used to implement the function f as its transfer function. The model accepts an input datum $\mathbf{x}_i$ and produces an output value $\tilde{y}_i \in \{\pm 1\}$.

The aim of the modeling is $\tilde{y}_i = \ell_i \in \{\pm 1\} \forall i$. In that respect the model reliably identifies the input-output relation of the given dataset. As a result, we can make the assumption that the model can be used to predict the label (output value) of a new unseen datum (input vector), correctly.

5. Attribute evaluation

5.1 Data pre-processing

A meaningful pre-processing step is the normalization of the given dataset. The Normalization ensures that the contribution of each input to the computation of metrics is irrespective of its actual range; it might be useful both in the data analysis and in the modeling performance [15]. In general, data normalization is an affine transformation (a linear combination plus a constant term) of each attribute value $x_{i,j}$ from its actual range $[x_j^+, x_j^-]$ into $[0,1]$. If the domain of normalization is the range $[0,1]$, then the transformation can be computed by Eq. (1)

$$\hat{x}_{i,j} = \frac{x_{i,j} - x_j^-}{x_j^+ - x_j^-} \in [0,1] \quad (1)$$

In order to simplify the mathematical notation we will use the symbol $x_{i,j}$ instead of $\hat{x}_{i,j}$ to denote the normalized value of the respective attribute for the rest of the paper.

After normalizing the data, the evaluation of candidate attributes which will be used as inputs to the model, follows. Since the initial set of attributes (i.e. candidate inputs) is intuitively/empirically selected by a human expert, the possibility of wrong decisions always exists. A human expert may select inputs which are either redundant or mutually dependent on one another.

5.2 Feature evaluation approaches

The input selection task involves the selection of those of the candidate inputs which significantly affect the output of the system. The selection of significant inputs is based on the collected observational data and it is usually carried out by statistical processing (filter based input selection) or by non-parametric models employed as wrappers. Moreover, a third category of input selection approaches is the embedded approaches where the identification of significant inputs follows from the model's construction process.

It is empirically recognized [10],[12] that the wrapper based methods provide more exact solutions than the filter based ones. However, wrapper based methods are model depended and their results depend on the particular model, applied as the wrapper. If a specific feature doesn't significantly affects the performance of the selected wrapper, then this feature is considered as meaningless and is eliminated. However, this assumption implies that the wrapper adequately identifies the dataset, which is not always true. Another drawback of wrapper methods is the lack of interpretation of why a particular feature is rejected or not. This drawback is more intense if the input output relation is non-linear for a particular input. Filter based methods are considered as less accurate, but they have the advantage of providing model independent and easily interpretative results. This aspect is important in financial applications where the results of any processing should be explanatory.

5.3 Filter based attribute evaluation

5.3.1 Pearson's correlation Coefficients

A widely used filter based method is the calculation of Pearson's coefficients. A Pearson's coefficient captures the linear correlation between two random variables. Although a Pearson's coefficient is limited to the calculation of the linear correlation between two random variables, it releases the advantage of straightforward and easily interpretative results, even for people who are not experts in the particular domain of application. We compute the Pearson's coefficient r_j for each attribute x_j by Eq. (2)

$$r_j = \frac{\sum_{i=1}^N (x_{i,j} - \bar{f}_j)(\ell_i - \bar{\ell})}{\sqrt{\sum_{i=1}^N (x_{i,j} - \bar{f}_j)^2 \cdot (\ell_i - \bar{\ell})^2}}, r_j \in [-1,1] \subseteq \square \quad (2)$$

Where $\bar{f}_j$ is the mean value of the j^{th} feature (i.e. the j^{th} column of $\mathbf{D}$); $\ell_i \in \{\pm 1\}$ is the label of instance i ; and $\bar{\ell}$ is the mean value of labels (i.e. the last column of $\mathbf{D}$). A value of r_j around ± 1 denotes a strong linear interdependence between the attribute x_j and the output. Controversially, a value around zero denotes linear independence. The sign of r_j denotes whether the linear relation is ascending or descending, respectively. Since we are only interested in the magnitude of the dependence the sign in Eq. (2) can be omitted by taking the

absolute value $|r_j|$ of r_j . The set of the candidate attributes includes thirty five attributes Table 3, intuitively selected by banking experts.

Pearson's approach requires that the data are normally distributed, besides the assumption of an existing linear relation between the random variables being probed. We use the normality test introduced in [5],[3] to test the null hypothesis: $H_0 = \{\text{The sample comes from a normal distribution}\}$. The p -value, which express a two-sided chi squared probability for the hypothesis test, is computed for each attribute (random variable). The p value was less than 0.05 for all attributes except from $x_{32}, x_{27}, x_3, x_{30}$ for which the p value was 0.96, 0.96, 0.78 0.07, respectively. The null hypothesis was rejected for most of the attributes and hence, Pearson's approach should not be directly used to decide the significance of attributes.

5.3.2 Spearman's rank correlation coefficients

An alternative non-parametric statistic is the Spearman's ρ rank-ordered correlation statistic [4], which measures the monotonic relationship between two random variables. Although, the existence of a monotonic relation between the random variables is an underlined assumption of Spearman's approach, however, this approach is less strict than Pearson's correlation coefficients.

When no duplicate values exist between the random variable x_j and the labels L , then Spearman's ρ can be computed by the following Equation.

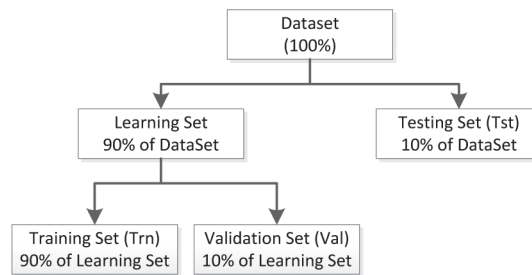
$$\rho_j = \frac{6 \sum_{i=1}^n [R_x(x_{i,j}) - R_L(\ell_i)]^2}{n(n^2 - 1)} \quad (3)$$

Where n is the number of instances; j denotes the j^{th} attribute; $R_x(x_{ij})$ denotes the rank of attribute value x_{ij} when sorted in ascending order; and $R_L(\ell_i)$ denotes the rank of ℓ_i when sorted in ascending order after the sorting of x_{ij} . If duplicate values exist, then the Eq. should be used on the ranked values of $x_{i,j}, \ell_i$, instead of Eq. for computing ρ_j [4].

The next issue is to decide a threshold of significance, below which an attribute is rejected as meaningless and above which the attribute is selected as significant. We follow two approaches to decide on the significance of an attribute. The first approach ignores the assumptions of Pearson's and Spearman's approaches and is based on the well-known student's statistical test to check the null hypothesis: $H_0 = \{\text{The probed random variables are by chance correlated}\}$. We check for two levels of significance: for $\alpha = 0.05$ and for $\alpha = 0.01$. Each attribute is evaluated according to its p -value, which roughly indicates the probability that an uncorrelated system produces datasets, which have a correlation at least as extreme as the one computed from the given data sets. The attributes with $p \leq \alpha$ are selected as important ones because the null hypothesis is rejected with high probability for them. Next, the selected attributes are ranked in descending order according to their $|\rho|$ value.

In the second approach we use the above mentioned statistical methods only as a ranking tool because of their implied assumptions. The second approach exploits the ranking which the statistical tests provide, but employs a non-parametric, non-linear model to evaluate the significant attributes. Tenfold cross-validation of the model is performed as follows: The initial data set is divided into a learning set, including the 90 percent of initial data, and a testing set, including the rest 10%. Next, the learning set is subdivided into the training set, including the 90 percent of the learning set and the validation set including the rest 10% of the learning set (Fig. 1).

Fig. 1: Tenfold cross-validation



The attributes are sorted in descending order according to their correlation to the output and then they are progressively inserted to a support vector machine model, as inputs. A tenfold cross validation is performed on the learning set for each new attribute and the average success classification rate on the validation set is monitored.

Table 4: The selected attributes presented in descending order according to their $|r|$ value when using Pearson's coefficients; Spearman's $|\rho|$ correlation coefficients; and Kendall's $|\tau|$. In every case a student's t-test was performed to check the null hypothesis for significance levels $p=0.05$ and $p=0.01$. Additionally, a selection based on tenfold cross validation on the learning set was performed. The average tenfold success classification rate on the testing set is given in the last column of every case. It is clear that the cross validation based selection delivered more representative attributes.

α	Selected Attributes	Average Tenfold Tst Success Rate
Pearson's r		
0.05	$x_{26}, x_{33}, x_{25}, x_{23}, x_5$	67.33%
0.01	x_{26}, x_{33}	70.00%
Cross-validation	x_{26}, x_{33}, x_{25}	71.33%
Spearman's ρ		
0.05	$x_{26}, x_{33}, x_{25}, x_{23}, x_{22}, x_3$	65.33%
0.01	x_{26}, x_{33}	70.00%
Cross-validation	x_{26}, x_{33}, x_{25}	71.33%
Kendall's τ		
0.05	$x_{26}, x_{33}, x_{25}, x_{23}, x_{22}, x_3$	65.33%
0.01	x_{26}, x_{33}	70.00%
Cross-validation	x_{26}, x_{33}, x_{25}	71.33%
PCA-SPPS		
$\Lambda=0.72$	$x_2, x_5, x_4, x_7, x_6, x_{19}, x_3, x_{18}, x_{28}, x_1, x_{13}, x_{14}$	62.00%
Cross-validation	$x_2, x_5, x_4, x_7, x_6, x_{19}, x_3, x_{18}, x_{28}$	70.00%
PCA-[16]		
$\Lambda=0.72$	$x_{25}, x_{32}, x_{26}, x_{10}, x_{15}, x_{14}, x_{35}, x_{27}, x_8, x_7, x_{16}, x_{12}$	68.00%
Cross-validation	$x_{25}, x_{32}, x_{26}, x_{10}, x_{15}, x_{14}, x_{35}, x_{27}$	70.00%
All Attributes		68.66%

The subset of attributes which provides the maximum average success classification rate on the ten validation sets is selected. Finally, a tenfold cross-validation is performed on the initial data set and the average success classification rate on the testing set is used as the final criterion for the selection. We highlight that the data of the testing set (Tst) were neither used in the construction of the svm [18], [17] model nor in the statistical tests, applied.

5.4 Multivariate attribute evaluation

Principal components analysis (PCA) in SPSS software is a variable-reduction technique that shares many similarities to exploratory factor analysis. Its aim is to reduce a larger set of variables-attributes into a smaller set of 'artificial' variables, called 'principal components', which account for most of the variance in the original

variables. PCA is predominantly used in an exploratory way. If one is interested in reducing the observed variables down to their principal components while maximizing the variance accounted for in the variables by the components, then he should be using PCA.

Except from the assumptions that specific filter based approaches requires, these approaches have the additional drawback that the attributes are evaluated one by one. It includes the risk of sub-optimal solutions because one attribute itself may be characterized as non important when evaluated alone but it might be important when evaluated with another one, jointly. On the contrary, an attribute when evaluated alone may be important while the same attribute might be not important in conjunction with another one. For that reason multivariate statistics exist in the field of multivariate analysis. Factor analysis is a method based on Principal components analysis (PCA). PCA projects the original input space to a new space of orthogonal variables which conveys the same information as the original ones. Factor analysis extracts a subset of specific variables (usually fewer), from the original set. The eigenvalues of the covariance matrix of the original data are computed at first. Next, the eigenvalues are sorted in descending order and are normalized such that their sum equals 1. The normalized eigenvalues are named latent variables. Consider m eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_m$, sorted in descending order ($\lambda_i > \lambda_j \quad \forall i > j$). The criterion for selecting the q most important variables is that the cumulative sum of the selected eigenvalues is up to a predefined threshold $\Lambda \in [0,1]$. That is, select the first q eigenvalues such that:

$$\frac{\sum_{i=1}^q \lambda_i}{\sum_{j=1}^m \lambda_j} \leq \Lambda \quad (4)$$

Which means that the total variance the selected variables express is up to $\Lambda \cdot 100\%$ of the total variance the original variables explain. Selecting the q most important eigenvalues which satisfy Eq. , does not provides direct information on which of the original values are important. In order to identify the q most important original variables we examine the absolute values of the coefficients of the respective q eigenvectors as in [16]. For $\Lambda = 0.95$, we got $q=17$. Actually, PCA computes new features by rotating the original axis, thus transforming the original space to a new orthogonal feature space, linearly. If the q most important eigenvectors are stored into an $m \times q$ matrix V , the transformation of the original $n \times m$ input space X is calculated by, which is an $n \times q$ matrix. Each original datum stored in row X_i in the original m -dimensional space is linearly mapped to row $\tilde{X}_i$ in the new q -dimensional feature space where $q < m$. We performed tenfold cross-validation by using a svm model on $\tilde{X}$. The model had q inputs, while the label of each datum was preserved.

The average success classification rate on the rotated testing data was 68.66% for $q=17$, which was the same as the one achieved when included all the m variables. We conclude that although in our problem linear PCA (and subsequently factor analysis) performed significant dimensionality reduction ($\square 50\%$); however, it failed to identify the best attributes. It failed both quantitatively, in terms of average success classification rate, and qualitatively in terms of which of the original variables were exactly the most important ones.

5.5 Discussion and interpretation of the results

The average tenfold cross-validation success classification rate and the respective attribute for $\alpha = 0.05$ and $\alpha = 0.01$ are summarized in Table 4. The selection based on the statistical test is computationally more efficient and more intuitive. However, it is based on assumptions that we don't priory know whether they are satisfied or not. Moreover, the α value affects the final outcome and it is an extra parameter being decided. Cross-validation is more computationally expensive but provides more accurate results since it is independent from the assumptions of the statistical tests.

It is obvious that results are very promising regarding credit analysis perspectives and key variables' selection, that remain important for a credit officer's thorough decision whether one can proceed to customer's lending or not. First of all, we see that Pearson's r is surprisingly good, nearly optimal results when compared to cross-validation, despite the fact that most of the attributes, considered as random variables, do not follow the normal distribution.

In the optimum set of results, three out of thirty five attributes are selected in cross validation option with average Tenfold Tst Success Rate at 71.33 percent. This is a very constructive result, where one can get. Only three out of thirty five attributes are enough for taking out the most influential information needed for banking authorities, in order to take proper lending decisions.

Another core conclusion is that alpha parameter on statistical methods is difficult to be decided; generally a value at 0.05 provides more attributes outcomes than required in any selected attribute of the research, while

a at 0.01 delivers less attributes than actually required. This issue is solved by the use of cross validation estimator performance. Furthermore, all statistical methods appeared more effective and robust than the simplest of the true eigenvector-based multivariate analyses (PCA). Moreover, all statistical methods detect significant attributes, as it was verified by cross validation; also all methods failed to detect the optimal attribute set. Generally the PCA method provides more attributes (nine attributes for both in PCA-SPPS and PCA-[16] options for cross validation analysis) and less representative, i.e. seventy percent for average tenfold success rate in both cases, while it is not robust.

The average tenfold success rate for all attributes lies at 68.66 percent. Optimal subset of attributes delivered the most accurate results in terms of generalization are in Pearson's r and Spearman's ρ options for the x26, x33 and x25 attributes. It is in -any terms- visible that qualitative attributes are those that express best sample's credit quality and provide maximum success rate in any case. This is very close to what Greek banking market experts' support that except from borrowers' core financial positions, factors such as quality of cooperation as well as good credit history records with the bank are essential for credit quality's assessment. Also, borrowers (SMEs) with exporting activity, apparently, tend to acquire higher credibility rates than those with solely domestic activity.

6. Conclusion

Traditional practices rely too much on credit quality indicators such as delinquency, nonaccrual, and risk rating trends. Banks have found that these indicators do not always provide sufficient information for a borrower's credit quality. Qualitative criteria are essential for the credit quality assessment. In our research a micro-analysis is made, taking under consideration a loan portfolio with reference to Greek firms. This part identifies the main elements, both quantitatively and qualitatively, that play major part in taking good lending decisions within a banking institution. Using several computational intelligence techniques in a data set from a Greek banking institution we find very thorough results for the bank management, relative to mitigate credit risk in loans portfolios.

More specifically, it is revealed that building a simplified model by using only 3 out of the 35 initially selected features one can achieve the same or slightly better forecasting accuracy when compared to the forecasting accuracy achieved by a model which uses all the 35 features (qualitative and quantitative ones). This is the main contribution of this study; only two firms' qualitative attributes (i.e. 1. Customer's Characterization and 2. Customer's Qualitative elements) as well as one attribute relative to firms' exporting or domestic activity that can provide adequate information for credit officers to mitigate bank's credit risk.

From the experimental results we observe that many of the intuitively selected attributes are redundant, while the generalization performance of many classifiers by using the selected attributes is rather poor. This observation leads us to conclude that the initially selected attributes are not representative enough and that the decision on the behavior of a feature borrower should be also based on other representative attributes, that take into account more custom made and focused characteristics, corresponding to focused entrepreneurial environment for the selected firms, in the sub sector that they enlisted. Furthermore, a richer set of training instances might lead to more accurate results.

In any term, it is generally acceptable that there is no a global credit quality system that fits for all cases, nearly for every loan's portfolio selected. This is a major conclusion that may be enriched in future research towards the understanding of a better and more conclusive segmentation of banks' loan portfolios, based upon certain and robust banking and market oriented features.

References

- [1] Basel Committee on Banking Supervision (2006a). Observed range of practice in key elements of Advanced Measurement Approaches (AMA). Available: <http://www.bis.org/publ/bcbs131.htm>
- [2] A. Benos and G. Papanastasopoulos, "Extending the Merton model: A hybrid approach to assessing credit quality," *Mathematical and computer modelling*, vol. 46, pp. 47-68, 2007.
- [3] K. Bowman and L. Shenton, (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2 . *Biometrika*, vol. 62, pp. 243-250, 1975.
- [4] G. W. Corder and D. I. Foreman, (2009). *Nonparametric statistics for non-statisticians: a step-by-step approach*: John Wiley & Sons, 2009.
- [5] R. B. d'Agostino, "An omnibus test of normality for moderate and large size samples," *Biometrika*, vol. 58, pp. 341-348, 1971.
- [6] K. Daniels and G. G. Ramirez, "Information, credit risk, lender specialization and loan pricing: Evidence from the DIP financing market," *Journal of Financial Services Research*, vol. 34, pp. 35-59, 2008.
- [7] M. Doumpos and C. Zopounidis, (2001). Assessing financial risks using a multicriteria sorting procedure: the case of country risk assessment. *Omega*, vol. 29, pp. 97-109.
- [8] A. Duff and S. Einig, (2009). Credit ratings quality: The perceptions of market participants and other interested parties. *The British Accounting Review*, vol. 41, pp. 141-153.

- [9] J. Fernandes, (2000). Corporate credit risk modeling: Quantitative rating system and probability of default estimation," Available at SSRN 722941.
- [10] M. A. Hall and L. A. Smith, (1998). Practical feature subset selection for machine learning.
- [11] C.-L. Huang, M.-C. Chen, and C.-J. Wang, (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, vol. 33, pp. 847-856.
- [12] R. Kohavi and G. H. John, (1997). Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, pp. 273-324.
- [13] Moody's, Moody's, Risk Calc and KMV EDF Risk Calc (v.3,1).
- [14] P. Sieczka and J. A. Hołyst, (2009). Collective firm bankruptcies and phase transition in rating dynamics," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 71, pp. 461-466.
- [15] J. Sola and J. Sevilla, (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *Nuclear Science, IEEE Transactions on*, vol. 44, pp. 1464-1468.
- [16] F. Song, Z. Guo, and D. Mei, (2010). Feature selection using principal component analysis, in *System Science, Engineering Design and Manufacturing Informatization (ICSEM)*, 2010 International Conference on, 2010, pp. 27-30.
- [17] V. Vapnik, (2000). *The nature of statistical learning theory*: Springer Science & Business Media.
- [18] V. N. Vapnik and S. Kotz, (1982). *Estimation of dependences based on empirical data* vol. 41: Springer-Verlag New York.
- [19] L. Yu, X. Yao, S. Wang, and K. K. Lai, (2011). Credit risk evaluation using a weighted least squares SVM classifier with design of experiment for parameter selection. *Expert Systems with Applications*, vol. 38, pp. 15392-15399.