Application of profit-based credit scoring models using R

Selçuk BAYRACI (selcuk.bayraci@cybersoft.com.tr)
R&D Centre, C/S Information Technologies,Istanbul, TURKEY

ABSTRACT

In this study, we applied a profit-based scoring system with using 10 different statistical and machine learning algorithms on a consumer credit data of a Turkish commercial bank. RStudio environment and R packages have been used in data cleaning, feature selection and model implementation processes. The results of the study reveal that artificial neural networks model seems to be superior to other techniques in terms of profit maximization.

***Keywords*:** Data analytics; Credit scoring; Banking; Risk management.
JEL Classifcation: C01, C40, C87

1. INTRODUCTION

Commercial banking sector plays a vital role in the economic and social welfare in developing countries as they are main providers of the required funds to manufacturing, agricultural, commercial and service industries. Therefore, building and sustaining a stable and transparent financial system is crucial for the well-being of an economy. Building solid and reliable credit scoring systems is of the utmost importance for the financial institutions as one of the most important threats to a country's financial stability is delinquency and defaults of consumer credits.

With the hasty growth in credit industry and the management of large credit portfolios, credit scoring models have been widely used in the credit application processes. The credit scoring models are developed to classify loan applicants as either a good credit group (accepted) or a bad credit group (rejected) with their associated characteristics or based on the data of the previous defaulted and non-defaulted credits. In measuring credit risk, a plethora of academic research has proposed a variety of theoretical, statistical and machine learning models such as discriminant analysis, linear regression, logit analysis, probit analysis, decision trees, Bayesian classifiers, k-nearest neighbours, support vector machines, and artificial neural networks (Hand and Henley, 1997; Abdou and Pointon, 2011; Lessman et al., 2015; Louzada et al., 2017).

Since 2004, with the global implementation of guidelines issued by the Basel Committee on Banking Supervision within Basel II Accord; large financial institutions started to develop advanced techniques in an attempt

to measure credit risk arising from important aspects of their business lines. The implementation of Basel II also causes several technical issues about the development and calibration of credit risk models. Additionally, commercial banks are private entities, meaning that they are "profit-seeking" organizations; they also have to optimize their "risk-return" trade-off when making a decision in a credit granting process. The accuracy of the classifier does not necessarily yield higher profits for the banks. As Lessman et al. (2015) conducted a comprehensive study on the eight real-world credit scoring datasets by using 41 different classification algorithms, and found out that the profitability of the scorecard does not depend on the accuracy of the classifier.

Therefore, considering risk/return relationship; we applied a profit-based classification measure, recently proposed by Verbraken et al. (2014), with using a number of statistical and machine learning algorithms on classification of the retail loans in a Turkish commercial bank's portfolio. The rest of the paper is organized as follows. Section 2 summarizes the previous studies in the field. The credit scoring techniques are briefly presented in Section 3. Section 4 gives the data properties. Model development and validation processes are summarized in section 5 and section 6, respectively; and section 7 concludes.

2. LITERATURE REVIEW

Since, there exists a vast number of literature devoted to credit scoring, we will review only credit scoring studies used profit-based techniques. To the best of our knowledge, Andreeva et al. (2007) first actualized profit-based credit scoring by survival analysis and logistic regression. They developed an algorithm that calculates the present value of net revenue from a revolving credit account by using the data from a store card designed for purchasing white durable goods in Germany. The empirically document that profit-driven model outperforms the conventional logistic regression but only by margin.

Finlay (2008) proposed a measurement which compares the continuous customer worth against the binary repayment behaviour via linear and logistic regression methods. He shows that incorporating a measurement of worth which estimates of payments and costs, into credit scoring systems significantly improves the accuracy rates of the standard models. In a follow up study, Finlay (2010) estimated a profit-based measurement which is calculated as difference between the net revenue and loss from liquidated debt. In addition to linear and logistic regressions; genetic algorithms and neural networks were also implemented in the classification process. The results of this study are in line with the previous one as scoring functions developed to optimize profit contribution perform better than the credit scoring methods.

For a credit card data, Stewart (2011) designed a profit-based scoring system which treats credit card spending as revenue and charged-off as cost.

The empirical findings of the study reveal that a profit-based scoring system taking risk and spending into account outperforms risk-only credit scoring models. So et al. (2014), in a similar vein, utilized a logistic regression model to analyse the credit card data from a Hong-Kong bank. Their proposed model includes the probability that the applicants will be granted credit card, depending on the interest rate charged and risk profile of the applicant. They argue that, the model leads to more accurate profitability estimates than models that ignore risk/return trade-off.

Barrios et al. (2013) developed a relative profitability measure which is calculated by dividing the credit lifetime value by the remaining debt. They employ linear and logistic regressions on a credit data from a Colombian lending institution. The results suggest their methodology yields higher portfolio returns than traditional scorecards.

Verbraken et al. (2014) introduced a performance measurement based on the expected maximum profit on two datasets of credits granted to micro-entrepreneurs originated from a government organization. Logistic regression and artificial neural networks are chosen as classification methods. The implementation of the expected maximum profit measure into classification algorithms significantly increases the profitability of the credit scoring systems.

In a recent study, Serrano-Cinca and Guiterrez-Neto (2016) developed a profit scoring system for the credit data from a peer-to-peer (P2P) lending platform. They investigated the profitability of investing on P2P loans, measured by internal rate of returns (IRR). They argue that, P2P lending market is not efficient and investors can generate abnormal returns by using data mining methods which can identify the most profitable loans. They statistically show that a profit scoring system using multivariate linear regression outperforms the results obtained from a traditional logistic regression based credit scoring system.

3. PREDICTION MODELS

When modelling delinquency and defaults of consumer credits, it is accepted to label one of the categories as good (non-default) and the other one as bad (default) and score these as 1 and 0 consequently. Thus, a usual credit data set contains a series of ones and zeros as the dependent variable. Connected with each Y, there will often be observations on a set of independent variables $X1, X2, \ldots Xp$.

Since Altman (1968) introduced the Linear Discriminant Analysis (LDA) to predict defaults, several researches have been proposed using different parametric, semiparametric and non-parametric models to improve

Altman's results. For this study, we consider the following models to predict the loan defaults.

3.1 Discriminant Analysis

Discriminant analysis which was introduced by Fisher (1936) is a classification technique that divides groups, which is aimed at the known population classification, in accordance with the discriminant criteria of that significant different groups have minimum variation inside the group, to seek the optimal weight value (w_i) , for the linear combination;

$$Z_i = w_0 + \sum_{i=1}^{m} w_i X_i \tag{1}$$

where Z_i is a discriminant score, w_0 is the intercept term, and $b_i(i = 1, \ldots, m)$ represents the estimated regression coefficient associated with the corresponding discriminant variables $X_i(i = 1, \ldots, m)$.

3.1.1 Linear Discriminant Analysis (LDA)

Suppose we observe two independent multivariate samples drawn from a multivariate normal distribution where p quantitative predictors have been observed for two cases n_1 and n_2, the LDA model assumes that both populations are multivariate normal with means μ_1 and μ_2 and common variance matrix Γ. The LDA rule classifies a p-dimensional vector X to class 2 if;

$$X^T \hat{\Gamma}^{-1} (\hat{\mu_2} - \hat{\mu_1}) > \frac{1}{2}\mu_2^T \hat{\Gamma}^{-1} \hat{\mu_2} - \frac{1}{2}\mu_2^T \hat{\Gamma}^{-1} \hat{\mu_1} + \log\pi_1 - \log\pi_2 \tag{2}$$

the prior probabilities of class memberships π_1 and π_2 are generally estimated by the class proportions in the training set.

3.1.2 Quadratic Discriminant Analysis (QDA)

The optimality of the LDA classification rule is valid under the assumption of multivariate normality and the homoscedasticity of the covariance matrix Γ. In the presence of deviation from homoskedasticity assumption, quadratic discriminant functions are required, and hence the QDA rule leads;

$$\text{argmax}\delta_i(X).\ \delta_i(X) = -\frac{1}{2}\log\left|\hat{\Gamma_i}\right| - \frac{1}{2}\left(X - \hat{\mu_i}\right)^T \hat{\Gamma}^{-1} \left(X - \hat{\mu_i}\right) + \log\pi_i \tag{3}$$

3.2 Generalized Linear Models (GLM)

As noted by Vojtek and Kocenda (2006), distribution of the credit information data generally does not obey the normal law, and this fact may theoretically pose a problem when using a discriminant analysis model. One way to solve the problem with non-normality of data is to use an extension of the DA model that allows for some parametric distribution. In this case a suitable extension is a generalized linear model (GLM). Introduced by McCullagh and Nelder (1989), GLMs provide a unified framework to model response from any member of the exponential family distributions, such as Gaussian, Binomial, or Poisson. In GLM, the dependent variable *Y* is related to the linear combination of predictor $\beta_1 * X1 + \ldots + \beta_m * X_m$ in the form of

$$G(E(Y|X)) = G(u) = \beta_0 + \sum_{i=1}^{m} \beta_i X_i, \quad (4)$$

where u is the mean of dependent variable *Y* and G (.) is a monotonic differentiable function known as Link Function.

3.2.1 Logistic Regression

For Logistic Regression, Generalized Linear Model can be expressed as

$$Logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{m} \beta_i X_i, \quad (5)$$

where (u) is $p = Prob(Y = 1|X)$ and *G(.)* is *Logit*(.) function in this case.

3.2.2 Probit Regression

Probit model assumes a normal distribution of the error term; Therefore, Probit Regression takes the form:

$$Prob(Y = 1|X) = \Phi^{-1}(p) = \beta_0 + \sum_{i=1}^{m} \beta_i X_i \quad (6)$$

where $\Phi^{-1}(p)$ is used to designate the Probit transformation of the predicted values - the link function. The -1 superscript refers to the inverse of the CDF to correspond with the cumulative probability that *Y* is equal to 1. The following formula describes the normal CDF

$$\Phi = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X} \exp(-0.5Z^2) dZ \quad (7)$$

3.2.3 Poisson Regression

Poisson regression model presumes that probability of credit default is drawn from a Poisson distribution with parameter λ;

$$Prob(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, y_i = 0,1 \tag{8}$$

The parameter λ can be estimated via a log-linear function as;

$$\log\lambda_i = \beta_0 + \sum_{i=1}^{m} \beta_i X_i \tag{9}$$

where X_i is a vector of explanatory variables, and β is a set of coefficients which is estimated using maximum-likelihood.

3.3 **Generalized Additive Models (GAM)**

In GLM, the relationship between response and predictors is assumed to be linear. However, a potential risk of such assumption is model misspecification. While the effects of predictors are often nonlinear in real world, it is always challenging to find an appropriate functional form of the partial effect of predictors on the response variable. As a result, GLM might not always be able to provide an appropriate fit for the data of complex structures (Liu and Cela, 2007).

Proposed by Hastie and Tibshirani (1987), Generalized Additive Model (GAM) relaxes the linearity assumption of GLM and assumes that the dependent variable Y is dependent on the univariate smooth terms of predictors rather than predictors themselves. Therefore, the functional form of GLM $\sum_{i=1}^{m} \beta_i X_i$ is replaced by the additive form $\sum_{i=1}^{m} s_i (X_i)$. The linear regression step in GLM is replaced by a nonparametric additive regression step, where the data is used to determine the appropriate smooth function s. This is done through iterative smoothing operations and allows for various non-linear effects of the explanatory variables. The logistic additive model, when applied to binary response data, takes the form (Berg, 2007);

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{m} s_i (X_i) \tag{10}$$

For this study, we use a semi-parametric form of Logistic GAM which can be expressed as;

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + X_1^T\beta + \sum_{i=1}^{m} s_i (X_{2i}) \tag{11}$$

where X_1 and X_2 denote categoric and numeric variables respectively.

3.4 k-Nearest Neighbours (k-NN)

The k-nearest neighbor method is a non-parametric statistical approach which evaluates the similarities between the pattern identified in the training set and the input pattern. It is based on choosing a metric on the space of applicants and takes k-nearest neighbor of the input pattern that is nearest in some metric sense. A new applicant will be classified in the class to which the majority of the neighbors belong (Kocenda and Vojtek, 2011). When conducting the k-NN algorithm, the choice of the metric is crucial. A commonly used metric is the standard Euclidean norm given by;

$$\rho_1(x,y) = [(x-y)^T(x-y)]^{0.5} \tag{12}$$

The choice of the number of nearest neighbors chosen *k* determines the bias/variance trade-off in the estimator. Starting with *k*=1, and each time by incrementing *k* to allow for one more neighbors; we estimated the expected profit in each case. The *k*-value that gives the maximum expected profit is selected.

3.5 Classification and Regression Trees (CART)

Classification and regression trees (CART), which is a nonparametric approach and consists of several layers of nodes: the first layer consists of a root node and the last layer consists of leaf nodes. CART models are found to be useful in producing accurate predictions by easily interpretable rules.

The algorithm of the CART model is explained by Kocenda and Vojtek (2009) as;

"The root node contains the entire training set; the other nodes contain subsets of the training set. At each node, the subset is divided into two disjoint groups based on one specific characteristic x_i *from the measurement vector. The split into two groups is defined by the following inequality: if* x_i *is an ordinal variable, then the split occurs when* x_i>t *for some constant t that characterizes a splitting rule. It follows that an individual j is classified into the right node if the previous statement is true; if not, the individual j is classified into the left node. A similar rule applies when* x_i *is a categorized variable".*

The density parameter is used to select the size of the CART trees. The density parameter or cost complexity factor controls the size of the tree by requiring the split at every node to decrease the overall lack of fit by that factor.

3.6 Artificial Neural Networks (ANN)

The most common type of neural networks which is called multilayer perceptron (MLP), has three layers of units: input layers, hidden layers, and output layers. A layer of "input" units is connected to a layer of "hidden" units, which is connected to a layer of "output" units. Figure 1 illustrates the architecture of a MLP. With reference to West (2000), the propagation of the network in each layer is accomplished in following steps.

Step 1: A weighted sum is calculated at each neuron, that is the output value of each neuron in the proceeding network layer times the respective weight of the connection with that neuron.

Step 2: A transfer function *g(x)* is then applied to this weighted sum to determine the neurons output value.

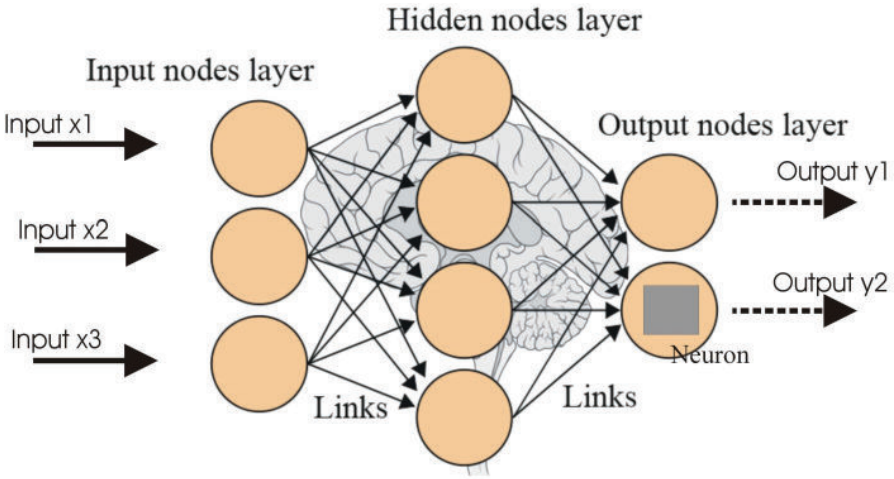
Step 3: The output value *Y*, for output neuron, *k*, can be expressed as a function of the input values and network weights, *w*,as follows:

$$Y(j) = \sum_{h=1}^{2} w_{kj} \left(g\left(\sum_{i=1}^{2} w_{ji} X^i \right) + w_{jb} \right) + w_{ib}, k = 1,2 \quad (13)$$

where *i* indexes the input neurons, *j* indexes the hidden layer neurons, and *b* indexes the respective bias values.

Three layer artificial neural network

Figure 1



[Source: Tadiou (2017)]

3.7 **Support Vector Machines (SVM)**

The main aim of SVM algorithm is to separate good credits ($y = 1$) from bad credits ($y = 0$) described with a *d* dimensional vector of characteristics

x. Inspired by Hardle et al. (2007), we use y = {-1,1} instead of the common y = {0,1} notation since it is more convenient in the following formal expressions. The SVM separates the two groups with the maximum distance (margin) between them. The score for x is computed as

$$f(x) = \sum_{i=1}^{n} \alpha_i \, y_i K(x, x_i) + b \tag{14}$$

In our classification problem, we use radial basis function or Gaussian kernel due to unknown relationships between the input variables. The Gaussian kernel on two samples x and x_i is defined to be:

$$K(x, x_i) = \exp\left(-\frac{||x - x_i||^2}{2\sigma^2}\right) \tag{15}$$

When the distance between x and x_i gets narrower, $K(x, x_i)$ becomes wider; therefore, the score *f(x)* is mainly defined by the observations that are close to x. The n factors α_i (Lagrange multipliers) are the free coefficients which are the solution of an SVM optimization problem and have higher magnitudes for the observations at the boundary between the classes which are most relevant for classification (Hardle et al., 2007).

3.8 Expected Maximum Profit (EMP) measure

To calculate the optimal cut-off value for the classification models, the Expected Maximum Profit (EMP) measure has been used. With reference to Verbraken et al. (2014)[1], the methodological framework begins with defining the average classification profit per borrower as;

$$P(t; b_0; c_1) = (b_0 * \pi_0 F_0(t)) - (c_1 * \pi_1 F_1(t)) \tag{16}$$

where $F_0(t)$ and $F_1(t)$ denote the true negative and false positive rates.

Optimizing the average profit which depends on the cut-off value t leads to the maximum profit measure:

$$MP = argmaxP(t; b_0; c_1) = P(T; b_0; c_1) \; \forall t \tag{17}$$

with T the optimal cut-off value under the given circumstances: $T = argmaxP(t; b_0; c_1) \; \forall t$

The optimal cut-off value T satisfies the first order condition;

1. Full derivation of the EMP measure can be found in Verbraken et al. (2014)

$$\frac{f_0 T}{f_1 T} = \frac{\pi_1}{\pi_0}\theta \tag{18}$$

where π_0 and π_1 denotes the prior probabilities of class 0 and 1, respectively. θ is the cost-benefit ratio, introduced for notational simplicity: $\theta = \frac{c_1}{b_0}$.

Parameter c_1 is the cost of incorrectly classifying a good applicant as a defaulter and is equal the return on investment (ROI) of the loan. Parameter b_0 is the benefit of correctly defining a defaulter; more precisely it is the fraction of the loan which can be lost due to default:

$$b_0 = \frac{LGD * EAD}{L} \tag{19}$$

L is the loan amount, LGD is the loss given default, and EAD is the exposure at default.

4. DATA

The methods are applied on real-world credit scoring data set. The data set is taken from a commercial Turkish bank and consists of 9915 granted loans of the retail loan portfolio for the period between August, 2014 and September, 2015, of whom 8401 were good credits and 1514 were bad credits implying that the default rate is around 15%.

For each customer in the data set, 80 characteristics and the class membership indicator is present. Even though the exact nature of the all characteristics cannot be disclosed for the confidentiality reasons, some relevant information is given in Appendix. 50 of the characteristics are quantitative and 30 are categorical. The characteristics are categorized as in Table A.1.

The summary statistics of the numerical variables are presented in Table A.2. We observe that all variables show non-normality behaviour as indicated with the significant Jarque-Bera test statistics. The mean and standard deviations of the subsets of defaulted and non-defaulted loans are also given in Table A.2 for the comparison purposes. The Kruskal-Wallis test of independence is conducted in order to test the differences between the defaulted and non-defaulted samples as by doing this we can have some initial information about the significance of the variables in explaining the default characteristics. The significant attributes are higlighted in bold-italic characters in the Table A.2. The statistical properties of the categorical variables are given in Table A.3.

5. MODEL DEVELOPMENT

5.1 Data preprocessing

Before moving into model building process, missing data points observed in the data set have been imputed with using the KNN algorithm[1]. When developing models we include all the explanatory variables summarized in Table A.1. Generally, using a stepwise procedure should be the preferred mode of action in choosing the explanatory variables that have significant predictive powers over the response variable. However, developing and testing a handful of models with a stepwise procedure is computationally expensive. Therefore, in order to determine the variables that have significant explanatory power, we resort to Chi-square test for the categorical variables. We also eliminate the continuous variables that are highly correlated. The inclusion of highly correlated explanatory variables may cause problems in practice; therefore we applied Principal Component Analysis (PCA) to select the significant factors among continuous variables.

In model development process, we use 70% of the data set (6936 observation), randomly selected from the full data set and referred to as the training set. The outstanding 30% (2976 observation) is used for model evaluation purposes and is referred to as the out-of-sample test set. By employing a PCA algorithm on 50 quantitative characteristics, we obtained 50 new statistical factors. The variance decomposition of the factors is presented in Table A.4. The explanatory power for each factor is distributed as follow: factor 1 – 16.72%; factor 2 - 12.64%, factor 3 – 9.27 % and so on. The proportion of the total variation explained by the first 27 factors is just above 95%. Therefore, we included these 27 statistical factors for the model implementation process.

The results of the Chi-square test results are presented in Table A.5. According to Chi-square test results only eight of the 30 categorical variables are found to be strong enough on explaining the default status of a loan. These variables are X51, X64, X70, X71, X73, X77, X79, and X80. The exact nature of the selected attributes can be seen in Table A.1. Moreover, we replaced the categorical variables with their corresponding weight of evidence (WOE) values for the sake of computation. Therefore, we totally obtained 35 numerical statistical factors for the model implementation procedure. The statistical properties of the final data used in model implementation are presented in Table A.6.

5.2 Model implementation in R

RStudio and R packages have been used in the estimation processes. Table 1,below, lists the classification methods and corresponding R packages.

1. We used the R package VIM (Kowarik and Temple, 2016) for the KNN imputation process.

Implemented models and R packages

Table 1

Model	R Package
Discriminant Analysis (Linear [LDA]; Quadratic [QDA])	MASS
Generalized Linear Models (Logit; Probit; Poisson)	stats
Generalized Additive Model (GAM)	mgcv
K Nearest Neighbours (KNN)	DMwR2
Classification and Regression Trees (CART)	evtree
Artificial Neural Networks (ANN)	nnet

5.2.1 **Discriminant analysis (MASS)**

The lda and qda functions from the **MASS** library (Venables and Ripley, 2002) have been used to implement the Discriminant analysis as follows;

```
lda_mod <- lda(Y~.,training_data)
qda_mod <- qda(Y~.,training_data)
```

5.2.2 GLM (stats)

For the estimation of the Logit, Probit and Poisson regression methods following glm functions from the **stats** library (R Core Team, 2016) have been utilized.

```
logitMod <- glm(Y ~ .,family=binomial, data = training_data)
probitMod <- glm(Y ~ ., family=binomial(link="probit"), data = training_data)
poissonMod <- glm(Y ~ ., family=poisson, data = training_data)
```

5.2.3 **GAM (mgcv)**

The gam function from the **mgcv** library (Wood, 2016) has been used for the estimation of the Logit GAM regression for the loan defaults.

```
mgcv::gam(Y~s(PC1)+s(PC2)+s(PC3)+s(PC4)+s(PC5)+s(PC6)+s(PC7)+s(PC8)+s(PC9)+s(
PC10)+
s(PC11)+s(PC12)+s(PC13)+s(PC14)+s(PC15)+s(PC16)+s(PC17)+s(PC18)+s(PC19)+
                  s(PC21)+s(PC22)+s(PC23)+s(PC24)+s(PC25)+s(PC26)+s(PC27)+
                  WOEX64 + WOEX70 + WOEX71 + WOEX73 + WOEX77 + WOEX79 + WOEX80
                ,family = binomial,data=training_data)
```

5.2.4 KNN (DMwR2)

The kNN function from the **DMwR2** library (Torgo, 2016) is applied to classify good and bad loans according to KNN algorithm as;

```
knn.mod <- DMwR2::kNN(Y ~ .,training_data,test_data,stand=FALSE,k=k)
```

5.2.5 CART (evtree)

The **evtree** (Grubinger et al., 2014) package has been used in predicting the loan status with a regression tree method.

```
weights <- array(1, nrow(training_data))
weights[training_data$Y == 0] <- 5
mod.tree <- evtree(as.factor(Y) ~ ., data = training_data, weights = weights)
```

5.2.6 ANN (nnet)

The following R script which runs the nnet function from the **nnet** package (Venables and Ripley, 2002) is used for the ANN model.

```
nnet_mod <- nnet(as.factor(Y) ~ .,
                 data=training_data,
                 size=2, skip=TRUE, MaxNWts=10000, trace=FALSE, maxit=100)
```

5.2.7 SVM (kernlab)

SVM model has been employed by using the ksvm function from the kernlab library (Karatzoglou et al., 2004)

```
svm_mod <- ksvm(as.factor(Y) ~ .,
                data=training_data,
                kernel="rbfdot",
                prob.model=TRUE)
```

6. MODEL VALIDATION

In order to test the efficacy of the models, the paper subjected the results to a hold -out sample (2976 credits). The validation result is displayed in Table 2 below. At this model validation stage, the predictive ability of the models are compared through accuracy (percentage of instances classified correctly); the sensitivity rate (percentage of correctly classified good credits); the specificity rate (percentage of correctly classified bad credits) and average classification error (arithmetic mean of the incorrectly classified instances).

Out-of-sample model performance

Table 2

	Accuracy	Sensitivity	Specificity	Avg. Classification Error	Net Economic Profit	Economic ROI
CART	84%	87%	68%	23%	5,912,598 TL	20%
GAM	85%	87%	75%	19%	7,065,970 TL	24%
kNN	89%	99%	31%	34%	6,557,733 TL	22%
LDA	88%	96%	43%	30%	6,557,695 TL	22%
Logit	85%	87%	77%	18%	7,131,303 TL	24%
ANN	86%	91%	62%	24%	7,178,851 TL	25%
Poisson	84%	86%	72%	21%	6,783,377 TL	23%
Probit	86%	88%	73%	19%	7,100,732 TL	24%
QDA	62%	56%	96%	24%	3,678,791 TL	13%
SVM	86%	87%	76%	18%	7,108,703 TL	24%

According to the results presented in Table 2, Logit regression and SVM models seems to outperform other models in terms of lowest average classification rates. KNN and QDA models produced highest sensitivity and specificity values, respectively; however they are not satisfactory models in terms of average classification error statistic as they also yield lowest specificity and sensitivity values. The economic profit and return on investment (ROI) values are also presented for the model evaluation purposes. ANN model has the highest net economic profit and ROI values.

7. CONCLUSION

The paper aimed at developing a credit scoring and risk assessment model by applying a profit-based classification measure with using several statistical and machine learning techniques. In relation to "profit-loss" trade-off, we found that the GLM, GAM,SVM and ANN models have a good chance of reducing risk of loss and high chance of increasing profits because they produced a high sensitivity rate and also accuracy out -of-sample thereby yielding a low type I error rate.

Our findings are in line with the previous studies in the field (Lessmann, 2015) as higher accuracy and lower misclassification errors do not necessarily yield more profitable scorecards. While kNN and SVM models have highest (lowest) accuracy (misclassification costs); but ANN model outperforms other models by providing highest profit rate.

We consider banks and financial institutions should use profit-based scoring methods instead of the traditional credit scoring models since even a minor improvement in predictive accuracy and profit rate are of critical importance. As stated by West (2000), even a meagre 1% improvement in accuracy would reduce losses in a large credit portfolio and save millions of dollars.

REFERENCES

1. **Abdou, Hussein A.,** and **John Pointon.**, 2011, "Credit scoring, statistical techniques and evaluation criteria: a review of the literature." *Intelligent Systems in Accounting, Finance and Management* 18.2-3 (2011): 59-88.
2. **Altman, Edward I.**, 1968, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." *The journal of finance* 23.4 (1968): 589-609.
3. **Andreeva, Galina, Jake Ansell,** and **Jonathan Crook**., 2007, "Modelling profitability using survival combination scores." European Journal of Operational Research 183.3 (2007): 1537-1549.
4. **Barrios, Luis Javier Sánchez**, **Galina Andreeva**, and **Jake Ansell**., 2014, "Monetary and relative scorecards to assess profits in consumer revolving credit." Journal of the Operational Research Society 65.3 (2014): 443-453.
5. **Berg, Daniel**., 2007, "Bankruptcy prediction by generalized additive models." *Applied Stochastic Models in Business and Industry* 23.2 (2007): 129-143.
6. **Bravo, C.**, **vanden Broucke**, **S. and Verbraken**, **T.,** 2017, "EMP: Expected Maximum Profit Classification Performance Measure", R package version 2.0.2.

7. **Finlay, Steven**., 2008, "Towards profitability: A utility approach to the credit scoring problem." Journal of the Operational Research Society 59.7 (2008): 921-931.
8. **Finlay, Steven** "Credit scoring for profitability objectives." European Journal of Operational Research 202.2 (2010): 528-537.
9. **Fisher, Ronald A**., 1936, "The use of multiple measurements in taxonomic problems." *Annals of human genetics* 7.2 (1936): 179-188.
10. **Grubinger, T., Zeileis, A**. and **Pfeiffer, K.-P.**, 2014, "evtree: Evolutionary Learning of Globally Optimal Classification and Regression Trees in R," *Journal of Statistical Software* (61:1), 2014, pp. 1-29.
11. **Hand, David J**., and **William E. Henley.,** 1997, "Statistical classification methods in consumer credit scoring: a review." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 160.3 (1997): 523-541.
12. **Härdle, Wolfgang K.**, **Rouslan A. Moro**, and **Dorothea Schäfer**., 2007, "Estimating probabilities of default with support vector machines." Working Paper, Humboldt-Universitat zu Berlin.
13. **Hastie, Trevor,** and **Robert Tibshirani**., 1987, "Generalized additive models: some applications." *Journal of the American Statistical Association* 82.398 (1987): 371-386.
14. **Karatzoglou, A., Smola, A., Hornik, K.** and **Zeileis, A**., 2004, "kernlab – An S4 Package for Kernel Methods in R," *Journal of Statistical Software* (11:9), 2004, pp. 1-20.
15. **Kocenda, Evzen**, and **Martin Vojtek**., 2011, "Default predictors in retail credit scoring: Evidence from Czech banking data." *Emerging Markets Finance and Trade* 47.6 (2011): 80-98.
16. **Kowarik, A.** and **Templ, M.**, 2011, "Imputation with the R Package VIM," *Journal of Statistical Software* (74:7), 2016, pp. 1-16.
17. **Lessmann, Stefan**, **et al**., 2015, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research." *European Journal of Operational Research* 247.1 (2015): 124-136.
18. **Liu, Wensui,** and **Jimmy Cela**., 2007, "Improving credit scoring by generalized additive model." *SAS global forum*.
19. **Louzada, Francisco**, **Anderson Ara**, and **Guilherme B. Fernandes**., 2016, "Classification methods applied to credit scoring: Systematic review and overall comparison." Surveys in Operations Research and Management Science.
20. **McCullagh, Peter**, and **James A. Nelder**., 1989, "Generalized Linear Models, no. 37 in Monograph on Statistics and Applied Probability."
21. **R Core Team**, 2016, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria.
22. **Serrano-Cinca, Carlos**, and **Begoña Gutiérrez-Nieto**., 2016, "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending." Decision Support Systems 89 (2016): 113-122.
23. **Stewart, Rob T.**, 2011, "A profit-based scoring system in consumer credit: making acquisition decisions for credit cards." Journal of the Operational Research Society 62.9 (2011): 1719-1725.
24. **So, Mee Chi, et al.**, 2014, "Using a transactor/revolver scorecard to make credit and pricing decisions." Decision Support Systems 59 (2014): 143-151.
25. **Tadiou, K.** "Artificial Neural Networks - The Future of Human Evolution" *Available at: http://futurehumanevolution.com/artificial-intelligence-future-human-evolution/artificial-neural-networks* [Accessed 5 Sept. 2017].
26. **Torgo, L.,** 2016, *Data Mining with R, learning with case studies, 2nd edition*, Chapman and Hall/CRC.
27. **Venables, W. N.** and **Ripley, B. D**., 2002, *Modern Applied Statistics with S*, Springer, New York.

28. **Verbraken, Thomas, Cristián Bravo, Richard Weber**, and **Bart Baesens,** 2014, "Development and application of consumer credit scoring models using profit-based classification measures." *European Journal of Operational Research* 238.2 (2014): 505-513.
29. **Vojtek, Martin**, and **Evzen Kocenda**., 2006, "Credit-scoring methods." *Czech Journal of Economics and Finance (Finance a uver)* 56.3-4 (2006): 152-167.
30. **West, David.**, 2000, "Neural network credit scoring models." *Computers & Operations Research* 27.11 (2000): 1131-1152.
31. **Wood, Simon N**., 2016 *Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.*

APPENDIX

Explanatory characteristics

Table A.1

Variable	Type	Description
X1	Numerical	Credit Amount (Turkish Lira)
X2	Numerical	Market Interest Rate (%)
X3	Numerical	Loan Term (Month)
X4	Numerical	Monthly Payment Amount
X5	Numerical	Basis Interest Rate (%)
X6	Numerical	Customer Interest Rate (%)
X7	Numerical	Current Interest Rate (%)
X8	Numerical	Total Monthly Cost
X9	Numerical	Default Interest Rate (%)
X10	Numerical	Default Basis Interest Rate (%)
X11	Numerical	Effective Cost Rate (Yearly)
X12	Numerical	Deposit Interest Rate (%)
X13	Numerical	Loan Granting Cost
X14	Numerical	Number of Defaulted Payments
X15	Numerical	Credit Bureau Total Monthly Payment
X16	Numerical	Duration of Current Employment
X17	Numerical	Total Duration of Employment
X18	Numerical	Monthly Expenses of the Applicant
X19	Numerical	Loan Costs (Yearly)
X20	Numerical	Current Term
X21	Numerical	*Descripiton of the variable cannot be disclosed*
X22	Numerical	*Descripiton of the variable cannot be disclosed*
X23	Numerical	*Descripiton of the variable cannot be disclosed*

X24	Numerical	*Descripiton of the variable cannot be disclosed*
X25	Numerical	*Descripiton of the variable cannot be disclosed*
X26	Numerical	*Descripiton of the variable cannot be disclosed*
X27	Numerical	*Descripiton of the variable cannot be disclosed*
X28	Numerical	*Descripiton of the variable cannot be disclosed*
X29	Numerical	*Descripiton of the variable cannot be disclosed*
X30	Numerical	*Descripiton of the variable cannot be disclosed*
X31	Numerical	*Descripiton of the variable cannot be disclosed*
X32	Numerical	*Descripiton of the variable cannot be disclosed*
X33	Numerical	Age of the Applicant
X34	Numerical	Child Number
X35	Numerical	House Member Count
X36	Numerical	House Look after Count
X37	Numerical	Duration of Current Residence
X38	Numerical	Total Income
X39	Numerical	Bank Score of the Applicant
X40	Numerical	Turkish Credit Bureau (KKB) Score of the Applicant
X41	Numerical	Number of Liquidated Loans of the Applicant
X42	Numerical	Total Amount of Late Payments (Other Banks)
X43	Numerical	Number of KKB Inquiries for the Applicant(Last 46 month)
X44	Numerical	Number of KKB Inquiries for the Applicant(Last 23 month)
X45	Numerical	Number of KKB Inquiries for the Applicant(Last 1 month)
X46	Numerical	Total Amount of Late Payments
X47	Numerical	Total Number of Late Payments (Other Banks)
X48	Numerical	Number of Charged-off loans
X49	Numerical	Number of Active Loans of the Applicant
X50	Numerical	Number of Active Credit Cards of the Applicant
X51	Categorical	New Customer X51A: No X51B: Yes
X52	Categorical	Home address information X52A: No X52B: Yes
X53	Categorical	Work address information X53A: No X53B: Yes
X54	Categorical	Home phone information X54A: No X54B: Yes

X55	Categorical	Work phone information X55A: No X55B: Yes
X56	Categorical	Mobile phone information X56A: No X56B: Yes
X57	Categorical	E-mail X57A: No X57B: Yes
X58	Categorical	Own a bank account X58A: No X58B: Yes
X59	Categorical	Own a real estate X59A: No X59B: Yes
X60	Categorical	Own stock market assets X60A: No X60B: Yes
X61	Categorical	Own a credit card X61A: No X61B: Yes
X62	Categorical	Known customer X62A: No X62B: Yes
X63	Categorical	Collateral X63A: No X63B: Yes
X64	Categorical	Own a salary account X64A: No X64B: Yes
X65	Categorical	Own a pension account X65A: No X66B: Yes
X66	Categorical	Own an investment account X66A: No X66B: Yes
X67	Categorical	Secured loan X67A: No X67B: Yes
X68	Categorical	Cash collateralized X68A: No X68B: Yes
X69	Categorical	Refinancing purpose X69A: No X69B: Yes
X70	Categorical	Own a credit deposit account X70A: No X70B: Yes
X71	Categorical	Education status X71A: No school X71B: Primary education X71C: Secondary education X71D: High school X71E: Vocational college X71F: Bachelor degree X71G: Master degree X71H: PhD

X72	Categorical	House type X72A: Rent X72B: Own X72C: Social estate X72D: Belong to a family member
X73	Categorical	Payment plan type X73A: Fixed X73B: Flexible
X74	Categorical	Gender X74A: Male X74B: Female
X75	Categorical	Marital status X75A: Married X75B: Divorced X75C: Widowed X75D: Single X75E: No Information
X76	Categorical	Housing country X76A: Turkey X76B: Foreign
X77	Categorical	Occupation type X77A: Self-employed X77B: Retired X77C: Government X77D: Housewife X77E: Unemployed X77F: Retired but working X77G: Retired (self-employee) X77H: Student X77I: Private sector
X78	Categorical	Bank score code X78A: Bad X78B: Good
X79	Categorical	Credit Bureau score code X79A: Medium X79B: Good
X80	Categorical	Product Code There are total 10 categories in this attribute but details are not available

Summary statistics of the numerical variables

Table A.2

		Total Loans				Defaults		Non-Defaults			
	NAs	Mean	Stdev	Jarque-Bera	p.val	Mean	Stdev	Mean	Stdev	Kruskal-Wallis	p.val
X1	4	10058.51	13371	1.32E+06	0.00	11773.95	11582.99	9749.93	13646.41	***121.25***	***0.00***
X2	0	1.75	0.20	2.76E+04	0.00	1.79	0.13	1.74	0.21	***20.33***	***0.00***
X3	0	23.92	12.06	5.60E+02	0.00	31.07	11.31	22.63	11.73	***557.16***	***0.00***
X4	4	603.65	1935	2.40E+08	0.00	459.19	408.50	629.70	2094.59	0.26	0.61
X5	0	1.75	0.20	2.76E+04	0.00	1.79	0.13	1.74	0.21	***20.33***	***0.00***
X6	0	1.12	0.36	1.52E+04	0.00	1.30	0.24	1.08	0.36	***576.96***	***0.00***
X7	0	1.12	0.36	1.52E+04	0.00	1.30	0.24	1.08	0.36	***576.96***	***0.00***
X8	0	0.96	0.05	2.36E+03	0.00	0.97	0.05	0.96	0.05	***72.70***	***0.00***
X9	0	2.27	0.26	2.76E+04	0.00	2.32	0.17	2.26	0.27	***20.33***	***0.00***
X10	0	1.75	0.20	2.76E+04	0.00	1.79	0.13	1.74	0.21	***20.33***	***0.00***
X11	0	16.24	4.27	7.97E+05	0.00	16.53	3.22	16.18	4.43	***80.76***	***0.00***
X12	0	1.12	0.36	1.52E+04	0.00	1.30	0.24	1.08	0.36	***576.96***	***0.00***
X13	871	92.54	91.30	3.52E+03	0.00	87.51	90.46	93.48	91.44	***17.70***	***0.00***
X14	0	1.67	2.55	7.81E+03	0.00	5.31	2.32	1.02	1.97	***3324.***	***0.00***
X15	3	44162.37	2.8E+4	1.01E+10	0.00	3781.22	8383.27	51442.32	3107262.00	2.90	0.09
X16	0	59.97	75.50	1.20E+04	0.00	54.45	69.45	60.97	76.50	2.77	0.10
X17	0	146.95	104.9	7.63E+02	0.00	153.93	104.84	145.69	104.93	***9.52***	***0.00***
X18	4	1582.05	3898	3.67E+07	0.00	2171.14	4923.39	1475.84	3674.08	***6.77***	***0.01***
X19	0	16.24	4.27	7.97E+05	0.00	16.53	3.22	16.18	4.43	***80.76***	***0.00***
X20	0	23.92	12.06	5.60E+02	0.00	31.07	11.31	22.63	11.73	***557.16***	***0.00***
X21	3	60.32	7.23	1.65E+03	0.00	60.79	7.90	60.24	7.10	1.67	0.20
X22	912	62.80	29.87	6.33E+02	0.00	65.15	30.87	62.34	29.65	***11.93***	***0.00***
X23	3	60.76	7.47	1.45E+03	0.00	61.11	8.02	60.69	7.36	0.71	0.40
X24	812	60.92	27.67	3.98E+02	0.00	62.94	28.52	60.54	27.49	***10.82***	***0.00***
X25	3	60.77	7.48	1.44E+03	0.00	61.22	8.08	60.69	7.36	1.71	0.19
X26	795	60.95	27.45	3.75E+02	0.00	63.07	27.99	60.54	27.32	***13.78***	***0.00***
X27	3	60.32	7.23	1.65E+03	0.00	60.79	7.90	60.24	7.10	1.67	0.20
X28	885	50.59	19.28	2.26E+01	0.00	51.64	20.02	50.38	19.13	***4.53***	***0.03***
X29	3	60.76	7.47	1.45E+03	0.00	61.11	8.02	60.69	7.36	0.71	0.40
X30	685	49.82	18.31	4.43E+01	0.00	51.07	19.49	49.58	18.06	***5.93***	***0.01***
X31	3	60.77	7.48	1.44E+03	0.00	61.22	8.08	60.69	7.36	1.71	0.19
X32	652	49.99	18.26	4.74E+01	0.00	51.55	19.42	49.68	18.01	***10.69***	***0.00***
X33	0	40.58	11.90	5.06E+02	0.00	40.70	11.57	40.55	11.96	0.99	0.32
X34	0	0.28	0.77	1.17E+06	0.00	0.19	0.61	0.30	0.80	***25.39***	***0.00***

X35	0	1.13	1.52	1.64E+03	0.00	1.00	1.45	1.15	1.54	***11.89***	***0.00***
X36	0	0.50	1.01	1.51E+04	0.00	0.43	0.96	0.51	1.02	***13.54***	***0.00***
X37	0	105.67	94.28	1.67E+04	0.00	102.21	89.51	106.29	95.11	0.27	0.61
X38	0	3001.95	5468	1.40E+07	0.00	3776.04	7320.07	2862.44	5050.67	***7.30***	***0.01***
X39	0	766.40	115	5.74E+02	0.00	672.57	111.78	783.31	107.19	***1059***	***0.00***
X40	0	1157.85	310.2	7.69E+03	0.00	909.75	373.90	1202.56	274.46	***943.94***	***0.00***
X41	0	0.04	0.33	2.23E+07	0.00	0.13	0.66	0.03	0.23	***147.91***	***0.00***
X42	0	498.57	1308	8.80E+09	0.00	2635.34	33397.09	113.49	597.38	***526.08***	***0.00***
X43	0	1.91	2.58	1.40E+05	0.00	3.17	3.70	1.68	2.25	***290.79***	***0.00***
X44	0	1.42	2.03	1.11E+05	0.00	2.49	3.01	1.23	1.72	***315.95***	***0.00***
X45	0	1.31	2.12	1.78E+05	0.00	2.76	3.45	1.05	1.64	***541.13***	***0.00***
X46	0	0.15	0.36	7.21E+03	0.00	0.33	0.47	0.12	0.33	***434.12***	***0.00***
X47	0	0.37	1.26	5.43E+05	0.00	1.36	2.60	0.19	0.67	***532.20***	***0.00***
X48	0	0.29	0.99	1.83E+06	0.00	0.48	1.36	0.26	0.90	***50.90***	***0.00***
X49	0	0.49	0.98	7.21E+06	0.00	0.45	0.67	0.50	1.03	0.19	0.66
X50	0	0.26	0.46	3.39E+03	0.00	0.00	0.00	0.31	0.49	***608.78***	***0.00***

Summary statistics of the categorical variables

Table A.3

		Total Loans		Defaults		Non-defaults	
Variable	Category	Number	%	Number	%	Number	%
X51	A	7407	74.70%	967	63.87%	6438	76.63%
	B	2508	25.30%	547	36.13%	1961	23.34%
X52	A	2981	30.07%	508	33.55%	2473	29.44%
	B	6934	69.93%	1006	66.45%	5926	70.54%
X53	A	1115	11.25%	106	7.00%	1009	12.01%
	B	8800	88.75%	1408	93.00%	7390	87.97%
X54	A	6943	70.03%	1196	79.00%	5746	68.40%
	B	2972	29.97%	318	21.00%	2653	31.58%
X55	A	3908	39.42%	521	34.41%	3387	40.32%
	B	6007	60.58%	993	65.59%	5012	59.66%
X56	A	77	0.78%	21	1.39%	56	0.67%
	B	9838	99.22%	1493	98.61%	8343	99.31%
X57	A	6009	60.61%	1036	68.43%	4973	59.20%
	B	3906	39.39%	478	31.57%	3426	40.78%
X58	A	9769	98.53%	1506	99.47%	8261	98.33%
	B	146	1.47%	8	0.53%	138	1.64%
X59	A	9824	99.08%	1511	99.80%	8311	98.93%
	B	91	0.92%	3	0.20%	88	1.05%
X60	A	9888	99.73%	1512	99.87%	8374	99.68%
	B	27	0.27%	2	0.13%	25	0.30%
X61	A	9850	99.34%	1510	99.74%	8338	99.25%
	B	65	0.66%	4	0.26%	61	0.73%
X62	A	1142	11.52%	246	16.25%	896	10.67%
	B	8769	88.44%	1268	83.75%	7499	89.26%
X63	A	570	5.75%	45	2.97%	524	6.24%
	B	9345	94.25%	1469	97.03%	7875	93.74%
X64	A	8346	84.18%	1472	97.23%	6872	81.80%
	B	1569	15.82%	42	2.77%	1527	18.18%
X65	A	9571	96.53%	1507	99.54%	8062	95.96%
	B	344	3.47%	7	0.46%	337	4.01%
X66	A	9838	99.22%	1513	99.93%	8323	99.07%
	B	77	0.78%	1	0.07%	76	0.90%
X67	A	9746	98.30%	1472	97.23%	8272	98.46%
	B	169	1.70%	42	2.77%	127	1.51%
X68	A	9631	97.14%	1514	100.00%	8115	96.60%
	B	284	2.86%	0	0.00%	284	3.38%
X69	A	7885	79.53%	1213	80.12%	6671	79.41%
	B	2030	20.47%	301	19.88%	1728	20.57%
X70	A	7385	74.48%	1445	95.44%	5938	70.68%

	B	2530	25.52%	69	4.56%	2461	29.29%
X71	A	22	0.22%	3	0.20%	19	0.23%
	B	3297	33.25%	313	20.67%	2984	35.52%
	C	5004	50.47%	930	61.43%	4074	48.49%
	D	939	9.47%	164	10.83%	775	9.23%
	E	257	2.59%	49	3.24%	208	2.48%
	F	201	2.03%	37	2.44%	164	1.95%
	G	7	0.07%	2	0.13%	5	0.06%
	H	187	1.89%	16	1.06%	171	2.04%
X72	A	772	7.79%	112	7.40%	660	7.86%
	B	4369	44.06%	746	49.27%	3623	43.13%
	C	333	3.36%	44	2.91%	289	3.44%
	D	4440	44.78%	612	40.42%	3828	45.57%
X73	A	9333	94.13%	1137	75.10%	8196	97.56%
	B	582	5.87%	377	24.90%	205	2.44%
X74	A	2158	21.77%	259	17.11%	1899	22.60%
	B	7757	78.23%	1255	82.89%	6502	77.40%
X75	A	7153	72.14%	1070	70.67%	6083	72.41%
	B	584	5.89%	136	8.98%	448	5.33%
	C	155	1.56%	28	1.85%	127	1.51%
	D	2020	20.37%	280	18.49%	1740	20.71%
	E	3	0.03%	0	0.00%	3	0.04%
X76	A	9914	99.99%	1514	100.00%	8400	99.99%
	B	1	0.01%	0	0.00%	1	0.01%
X77	A	2553	25.75%	264	17.44%	2289	27.25%
	B	420	4.24%	119	7.86%	301	3.58%
	C	1304	13.15%	170	11.23%	1134	13.50%
	D	42	0.42%	1	0.07%	41	0.49%
	E	5	0.05%	0	0.00%	5	0.06%
	F	15	0.15%	3	0.20%	12	0.14%
	G	16	0.16%	0	0.00%	16	0.19%
	H	8	0.08%	1	0.07%	7	0.08%
	I	5550	55.98%	955	63.08%	4595	54.70%
X78	A	1	0.01%	0	0.00%	1	0.01%
	B	9914	99.99%	1514	100.00%	8400	99.99%
X79	A	7432	74.96%	747	49.34%	6685	79.57%
	B	2483	25.04%	767	50.66%	1716	20.43%
X80	A	889	8.97%	111	7.33%	778	9.26%
	B	910	9.18%	188	12.42%	722	8.59%
	C	656	6.62%	67	4.43%	589	7.01%
	D	1257	12.68%	232	15.32%	1025	12.20%
	E	754	7.60%	102	6.74%	652	7.76%

F	701	7.07%	109	7.20%	592	7.05%
G	810	8.17%	180	11.89%	630	7.50%
H	1685	16.99%	248	16.38%	1437	17.11%
I	1251	12.62%	174	11.49%	1077	12.82%
J	1002	10.11%	103	6.80%	899	10.70%

Variance decomposition of the Principal Components

Table A.4

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	2.8915	16.72%	16.72%
PC2	2.5142	12.64%	29.36%
PC3	2.1523	9.27%	38.63%
PC4	1.5797	4.99%	43.62%
PC5	1.5195	4.62%	48.24%
PC6	1.5125	4.58%	52.81%
PC7	1.3931	3.88%	56.70%
PC8	1.3540	3.67%	60.36%
PC9	1.2712	3.23%	63.59%
PC10	1.2277	3.01%	66.61%
PC11	1.1985	2.87%	69.48%
PC12	1.1316	2.56%	72.04%
PC13	1.0767	2.32%	74.36%
PC14	1.0318	2.13%	76.49%
PC15	0.9981	1.99%	78.48%
PC16	0.9861	1.95%	80.43%
PC17	0.9395	1.77%	82.19%
PC18	0.9119	1.66%	83.86%
PC19	0.8945	1.60%	85.46%
PC20	0.8769	1.54%	86.99%
PC21	0.8707	1.52%	88.51%
PC22	0.8403	1.41%	89.92%
PC23	0.8017	1.29%	91.21%
PC24	0.7393	1.09%	92.30%
PC25	0.7370	1.09%	93.39%
PC26	0.7235	1.05%	94.43%
PC27	**0.6912**	**0.96%**	**95.39%**
PC28	0.6617	0.88%	96.26%
PC29	0.6240	0.78%	97.04%
PC30	0.6174	0.76%	97.80%
PC31	0.5149	0.53%	98.33%
PC32	0.4945	0.49%	98.82%
PC33	0.4840	0.47%	99.29%
PC34	0.4253	0.36%	99.65%

PC35	0.3786	0.29%	99.94%
PC36	0.1649	0.05%	100.00%
PC37	0.0487	0.01%	100.00%
PC38	0.0000	0.00%	100.00%
PC39	0.0000	0.00%	100.00%
PC40	0.0000	0.00%	100.00%
PC41	0.0000	0.00%	100.00%
PC42	0.0000	0.00%	100.00%
PC43	0.0000	0.00%	100.00%
PC44	0.0000	0.00%	100.00%
PC45	0.0000	0.00%	100.00%
PC46	0.0000	0.00%	100.00%
PC47	0.0000	0.00%	100.00%
PC48	0.0000	0.00%	100.00%
PC49	0.0000	0.00%	100.00%
PC50	0.0000	0.00%	100.00%

Chi-square test statistics

Tablo A.5

Variable	Test statistic	Variable	Test statistic
X51	***110.9985***	X66	11.7075
X52	10.3387	X67	12.2017
X53	32.2493	X68	52.6908
X54	68.5081	X69	0.3858
X55	18.7284	X70	***413.0104***
X56	8.6413	X71	***142.9307***
X57	45.8005	X72	20.0135
X58	10.9779	X73	***1171.2603***
X59	10.1762	X74	22.7681
X60	1.2936	X75	34.1257
X61	4.2024	X76	0.1802
X62	39.1432	X77	***133.9226***
X63	25.4242	X78	0.1802
X64	***228.4648***	X79	***624.6961***
X65	48.2463	X81	***102.1314***

Properties of the statistical factors used in the model estimation

Table A.6.

	Minimum	Maximum	Mean	Stdev
PC1	-11.5108	20.5094	-0.0142	2.9097
PC2	-5.8733	16.0644	0.0093	2.5207
PC3	-11.0334	6.6961	-0.0057	2.1626
PC4	-7.6264	9.7532	0.0121	1.5822
PC5	-8.4784	11.4624	-0.0093	1.5354
PC6	-6.1846	28.4442	0.0057	1.5310
PC7	-6.7391	113.2358	0.0022	1.5709
PC8	-45.7332	7.1008	-0.0022	1.3925
PC9	-28.1234	3.4248	0.0019	1.2878
PC10	-7.7701	6.0489	-0.0096	1.2202
PC11	-9.3706	9.0706	0.0065	1.2040
PC12	-14.0071	21.7067	0.0093	1.1435
PC13	-23.1220	16.9375	0.0132	1.0780
PC14	-11.0420	33.3106	0.0072	1.0206
PC15	-66.2455	2.1827	-0.0048	1.1717
PC16	-32.8103	16.8630	-0.0021	0.9881
PC17	-18.8240	20.2310	0.0135	0.9749
PC18	-8.0208	2.9128	-0.0034	0.9245
PC19	-8.5349	8.3472	-0.0011	0.9071
PC20	-5.9403	3.9868	0.0062	0.8840
PC21	-10.6112	7.7189	-0.0031	0.8634
PC22	-6.0948	3.7956	0.0051	0.8348
PC23	-4.8944	6.6012	-0.0052	0.8138
PC24	-16.4745	13.9135	-0.0016	0.7530
PC25	-5.8648	4.3383	0.0103	0.7430
PC26	-4.9758	8.9677	-0.0026	0.6786
PC27	-6.0000	4.7103	-0.0037	0.6910
WOEX51	-0.4368	0.1825	0.0252	0.2696
WOEX64	-0.1725	1.8798	0.1588	0.7551
WOEX70	-0.3000	1.8606	0.2549	0.9440
WOEX71	-0.7966	0.6561	0.0477	0.3705
WOEX73	-2.3228	0.2617	0.1119	0.6038
WOEX77	-1.6998	2.0138	0.0468	0.3469
WOEX79	-0.9083	0.4780	0.1315	0.6002
WOEX81	-0.4608	0.4601	0.0282	0.2830
Y	0.0000	1.0000	0.8474	0.3596