Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 106 (2013) 667 – 676

Procedia
Social and Behavioral Sciences

4th International Conference on New Horizons in Education

Comparing the predictive and classification performances of logistic regression and neural networks: a case study on timss 2011

Oykum Esra Askin[a*], Fulya Gokalp[b]

[a]Yildiz Technical University, Faculty of Arts and Science 34220, "Istanbul", Turkey
[b]Yildiz Technical University, Faculty of Arts and Science 34220, "Istanbul", Turkey

Abstract

Investigating effective factors on students' achievement has wide application area in educational studies. Specially, Trends in International Mathematics and Science Study (TIMSS) allows researchers to determine correlates of mathematics and science achievement for different countries. In this study, the predictive and classification performances of logistic regression and neural networks are compared to identify the impact levels of variables on students' mathematics achievement in Turkey. Age, gender and scales created by TIMSS team for 8th grade students (students like learning, value learning, confident in math, engaged in math, bullied at school, home educational resources), are selected as predictive variables. Model fitting statistics show that two methods give similar results in prediction and classification. In addition to model results, students' confidence is found as the most effective factor to improve mathematics achievement.

Selection and peer-review under responsibility of The Association of Science, Education and Technology-TASET, Sakarya Universitesi, Turkey.

Keywords: TIMSS 2011, neural networks, logistic regression

1. INTRODUCTION

Education policy makers desire to get reliable information about main factors effecting students' achievement. In this sense, obtained information could be used for monitoring current problems and identifying effective solutions to improve the quality of education system. One of the common ways to acquire the mentioned information is using a reliable database such as Trends in International Mathematics and Science Study (TIMSS). TIMSS is the largest international comparative study which provides useful information about educational achievement and learning contexts for policy makers, educators, researchers, and practitioners (Martin, Mullis, Beaton, Gonzalez, Smith, & Kelly, 1997). There is an extensive literature uses TIMSS data to determine factors

* Corresponding author. Tel.: +90 507 956 77 29; fax: +90 212 383 44 21
E-mail address: oykumesra@gmail.com

Selection and peer-review under responsibility of The Association of Science, Education and Technology-TASET, Sakarya Universitesi, Turkey.
doi:10.1016/j.sbspro.2013.12.076

impact on students' achievement. While some studies investigate whether school characteristics such as class/school size, resources and the curriculum are significant determinants of achievement (Papanastasiou, (2000); Pong and Pallas (2001); Schreiber (2002); Webster and Fisher (2003); Woessmann (2003)), some studies focus on teachers' characteristics such as education level, teaching experience, qualification and hours spent on planning lessons ((Akiba, LeTendre and Scribner (2007); Dodeen, Abdelfattah, Shumrani and Hilal (2012); Hanushek and Luque (2003); Jürges and Schneider (2004); Mills and Holloway (2013); Woessmann (2003)). Student characteristics such as age, gender, family background (e.g., parental education level, income, SES), number of books at home, and attitudes are another important factors concerned with achievement (Aypay, Erdoğan and Sözer (2007); Neuschmidt, Barth and Hastedt (2008); Woessmann (2004; 2005); Yayan and Berberoğlu (2004)).

Several statistical methods have been applied in order to find significant relationship between factors mentioned above and students' achievement. With the rapid development of information technologies, a great number of techniques within data mining are attracting educational researchers (see. Romero and Ventura (2007)). However, still there is a lack of studies used data mining tools for TIMSS data. In this paper, two well-known data mining methods; logistic regression (LR) and neural networks (NNs), are used. In addition to the usage of LR and NNs to identify the impact levels of several factors on students' achievement, this study also attempts to fill another gap by evaluating the prediction and classification performance of LR against to NNs. In this vein, motivations of this paper can be described as followings: (i) to determine the significance of factors on achievement by using TIMSS 2011 survey data of Turkish 8^{th} grade students (ii) to compare the prediction and classification effectiveness of two methods based on different model fitting statistics.

2. Methodology

2.1. Study methods

LR and NNs which are two popular data mining methods aim to describe variable(s) in relation to the other(s) by looking for rules of classification or prediction based on the data (Giudici & Figini, 2009) and they have been applied in many fields such as health, finance, agriculture and engineering. In addition to LR and NNs methods, several different methodologies such as decision trees, cluster analysis and ensemble models are performed on the relevant data and the results of outcomes are compared according to their predictive/classifier accuracy. In this way, the question of which approach is outperformed by another(s) can be reported. In recent years, such studies are rapidly increasing in educational studies. To give an example for Turkish literature, the predictive powers of four different data mining methods on secondary education placement test results are investigated and the ranked-importance of factors on prediction models are determined by Şen, Uçar and Delen (2012). As it mentioned before, this study deals with TIMSS 2011 Turkey data and two popular methods, namely LR and NNs are used. A short brief of these methods is given in the following sections.

2.1. 1. Logistic regression

Logistic regression is a variation of the ordinary least squares (OLS) regression, but unlike OLS, this method allows to use two or more categorical variables as a dependent variable. This study focus on binary LR method when the dependent variable is dichotomous.

$\pi(x)$ is the probability of outcome of interest and it represents the conditional mean of Ygiven X ($\pi(x) = E[Y \backslash X = x]$). The residuals do not follow a normal distribution because Y gets only two possible values. Tthe conditional mean falls within the [0,1] interval. The change in the conditional mean per unit with the change in X becomes progressively smaller as the conditional mean gets closer to 0 or 1. Therefore, the shape of the fitted regression line is likely to be S-shaped. Despite different distribution functions are proposed by researchers, the most common used in the literature is logistic distribution function. The form of the simplest logistic regression model is as follows (Hosmer & Lemeshow, 2000);

$$\pi(x) = \frac{e^{x'\beta}}{1-e^{x'\beta}} = \frac{e^{x'\beta} / e^{x'\beta}}{(1-e^{x'\beta}) / e^{x'\beta}} = \frac{1}{e^{x'\beta}} \tag{1}$$

where $\beta' = (\beta_0, \beta_1, \dots, \beta_k)$ denotes the vector of parameters and k is the number of independent variables used in the model. In order to convert the model from nonlinear to linear form, a logit transformation of $\pi(x)$ is applied.

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = x'\beta \tag{2}$$

The left side of Equation (2) is also named as a logit and it can be expressed by logit (Y) or logit $\pi(x)$. The logit is the natural logarithm of odds which is derived by taking the ratio of probability of success to probability of failure of an event. As it can be seen clearly, the odds ratio of an event take values in the interval $[0, +\infty]$ whereas the logit take the values in the interval $[-\infty, +\infty]$. The logit may be continuous and it is linear in its parameters (Hosmer & Lemeshow (2000); Alpar (2011)).

In LR, the goal is to predict the logit of dichotomous outcome Y from the categorical or the continuous independent variable X . The unknown regression coefficients are estimated by using maximum likelihood estimation (MLE) method which maximises the likelihood function. The log likelihood function is given below (Hosmer & Lemeshow, 2000);

$$L(\beta) = \ln\left[l(\beta)\right] = \sum_{i=1}^{n}\left\{y_i \ln\left[\pi(x_i)\right] + (1-y_i)\ln\left[1-\pi(x_i)\right]\right\} \tag{3}$$

where $l(\beta)$ denotes the likelihood function of the parameters. By taking first derivates of $L(\beta)$ with respect to β and then equalizing the likelihood function to zero, we obtain the maximum likelihood estimator of $\beta's$. The value of maximum likelihood estimate, denoted as $\hat{\beta}$, shows how much logit changes with one unit change in X. From the Equation (2), it can be easily interpreted that the odds ratio is equal to $\exp(\beta)$ and for the value of 1, no relationship between X and logit (Y) is interpreted.

Three different tests based on the likelihood function are used in order to test the significancy of estimated β coefficients. To test the significancy of overall LR model, the likelihood ratio test is performed. The test statistic (G) is the difference between two deviance values: deviance for the model with only constant (the null model) and the deviance for the model with constant and variable(s). It follows a Chi Square distribution and the formula of the test statistics can be expressed as $G = D(null\ model) - D(model\ with\ variables)$ or $G = -2\ln\{likelihood\ (the\ null\ model)/likelihood\ (model\ with\ variables)\}$. If the null hypothesis: $h_0: \beta_0 = \beta_1 = \cdots = \beta_k$ is rejected at chosen significance level (usually at $\alpha = 0.05$ significance level), it is concluded that variable(s) added to model increase(s) the model fit. The Wald Test is performed in order to test whether any individual parameter (slope coefficient) equals to zero or not. Let $\hat{\beta}$ denotes the maximum likelihood estimate of unknown parameter β and $SE(\hat{\beta})$ denotes the standard error of $\hat{\beta}$. The Wald statistics follows a Chi Square distribution with one degrees of freedom and can be obtained as follows: $W = \hat{\beta}^2 / SE(\hat{\beta})^2$. The last one is Lagrange Multiplier (Score) test which is based on the distribution theory of the derivates of the log likelihood and limited software packages includes this test due to complicated matrix calculations (Hosmer & Lemeshow, 2000).

2.1.2. Neural networks

Neural networks are widely used computational models and the main idea is to mimic the process of human brain. Algorithms of NNs are used as an alternative of standard statistical techniques in prediction and classification due to having following benefits: (1) Like human brain, NNs have the ability of learning and generalizing from past experience. According to Sivanandam, Sumathi and Deepa (2006), the ability of learning allows NNs to adjust themselves to the dynamic and changing environment. (2) NNs are easily overcome with complex non-linear relationship between inputs (independent variables) and outputs (dependent variables), and there is no need to make priori assumptions about the mathematical forms between inputs and outputs. (3) NNs are robust to noisy, missing and incomplete data. (4) Possible interactions among inputs can be easily determined by NNs. However, Linoff and Berry (2011) mentioned about some weakness of NNs versus other statistical techniques. Because of black box nature, NNs can produce well established model but cannot explain how to do it. For example unlike logistic regression, there is no way to get an accurate set of rules from NNs. Also LR enables to make confidence intervals and to develop hypothesis tests in order to make decisions through statistical theory. Non availability of such statistical procedures is the other drawback of NNs. Lastly, when training sets are inaccurate for the large amounts and chaotic data, an over fitting problem may occur (Sivanandam, Sumathi, & Deepa, 2006).

As it mentioned before, NNs have the ability of learning and learning can be done by training. The back propagation algorithm is the most popular supervised training method and it performs learning on a multilayer perceptron (Han & Kamber, 2006).

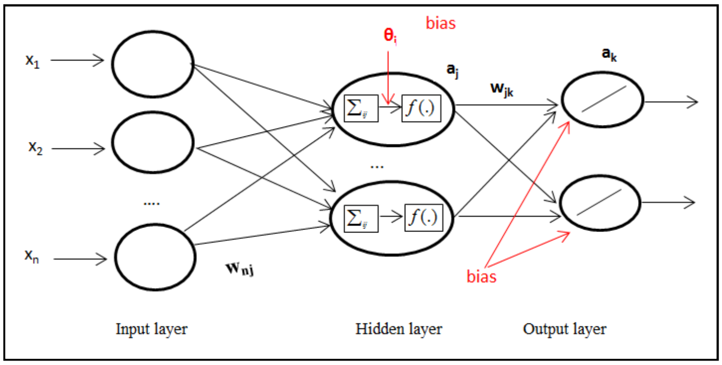

Fig. 1. An architecture of a simple MLP with one hidden layers

The Multilayer Perceptron (MLP) consists of three components: an input layer, one or more hidden layers and an output layer. Figure 1 shows architecture of two layers MLP. Neurons which are the smallest unit of a NN are represented by circles. The weight value carries information from one neuron to another. Steps of backpropagation algorithm used in training of MLP are described below (Han & Kamber, 2006);

- Step 1: (Initialize the network weights) Weights are randomly assigned as small numbers.
- Step 2: (Propagate the inputs forward) Each input layer unit j carries its information to all hidden units. For an input j, its output a_j is equal to its input value I_j. Then the net input of unit j according to the previous layer i is computed as $I_j = \sum_i w_{ij} a_i + \theta_j$ where I_j is the linear combination of inputs $x_i: i = 1, \ldots, n$; w_{ij} is the weight value, a_i is the output of unit i from the previous layer and θ_j is the bias of the unit j (sometimes bias takes the value 1). Next, each hidden unit j computes the output value a_j with respect to the type of activation function which is shown as $f(\cdot)$. The most popular activation function used in the hidden layer is the sigmoid or logistics which is a continuous and differentiable nonlinear function. With the help of the sigmoidal activation function, the output of unit j can be calculated as $a_j = 1/(1 + e^{-I_j})$.

- Step 3: (Backpropagate errors) The error is propagated backward by updating the weights and biases to reflect the error of the network's prediction. The output layer error is obtained by $Err_j = a_j(1 - a_j)(T_j - a_j)$ where T_j is the known desired value of the output. Then, error of the hidden layer j is calculated as $Err_j = a_j(1 - a_j) \sum_k Err_k w_{jk}$ where Err_k is the error of unit k.
- Step 4: (Updated the weights and biases) For each w_{ij}, weight increment is calculated by $\Delta w_{ij} = (l) Err_j a_j$ where l is the learning rate which takes value between 0 and 1. Then each weight is updated for each bias θ_j in the network with the formulation: $w_{ij} = w_{ij} + \Delta w_{ij}$. Similarly, each bias θ_j is updated by $\Delta\theta_j = (l) Err_j$ and then the new value of bias is calculated as $\theta_j = \theta_j + \Delta\theta_j$. Training stops until a pre-specified number of epochs have expired and the sufficiently small overall error is satisfied (repeat steps 2-4). Training in network generally continues until it reaches a certain error rate such as 95% . Too much training may cause any over fitting problem. Additionally, before applying step 2, input variables are usually normalized to 0-1 scale. Thus, changing the scale helps speed up the learning phase.

2.2. Data and variables

TIMSS is an international assessment of mathematics and science at the 4th and 8th grade students from different nations and it is conducted every four years by the International Association for the Evaluation of Educational Achievement (IEA) since 1995. In 2011, more than 600,000 students from 63 countries participated to this international study and 45 of them administered the eighth grade assessment (Mullis, Martin, Foy, & Arora, 2012). Data used in this study includes 6928 Turkish students' information from student questionnaire and some missing or inaccurate values exist. To handle this problem, 192 students' data are excluded from the analysis.

Table 1. Independent variables used in study

Variable name	Type of data	Codes of the statements
Age	Ratio	ITBIRTHD
Gender	Nominal	ITSEX
Students like learning math	3 point likert scale	BSBM14A- BSBM14B, BSBM14C, BSBM14D, BSBM14E.
Students confident in math	3 point likert scale	BSBM16A, BSBM16B, BSBM16C, BSBM16D, BSBM16E, BSBM16F, BSBM16G, BSBM16H, BSBM16I
Students value math	3 point likert scale	BSBM16J, BSBM16K, BSBM16L, BSBM16M, BSBM16N, BSBM14F
Students engaged in math	3 point likert scale	BSBM15A, BSBM15B, BSBM15C, BSBM15D, BSBM15E
Students bullied at school	3 point likert scale	BSBG13A, BSBG13B, BSBG13C, BSBG13D, BSBG13E, BSBG13F
Home educational resources	3 point likert scale	BSBG04, BSDGEDUP, DSDG055

Before applying LR and NNs methods to determine the significance of variables on students' mathematics achievement and to compare the prediction/classification performances, dependent and independent variables are determined. In this study, first plausible value of the mathematics test is chosen as a dependent variable and scores for each student are encoded as binary values of 1 or 0 (if the score is over the average 500, y=1; otherwise, y=0). In order to find factors affecting mathematics achievement, the results of eight student survey questions are selected as potentially influencing variables. These independent variables are given in Table 1. According to Table 1, six of them except age and gender are created based on the students' responses to different statements by TIMSS team. Statements used to create the relevant scale are given in the last column of the table with their codes. For example, "*Students like learning mathematics*" is a 3 point likert scale (1=like learning math: 2=somewhat like learning math: 3=don't like learning math) and it measures students' feeling about mathematics based on the following 4 point likert statements: "I enjoy learning mathematics" (BSBM14A), "I wish I did not have to study mathematics" (BSBM14B), "Mathematics is boring" (BSBM14C), "I learn many interesting things in mathematics" (BSBM14D) and "I like mathematics" (BSBM14E). "*Students confident in mathematics*" scale aims to investigate students' beliefs about their abilities in mathematics and it is a 3 point likert scale (1=confident: 2=somewhat confident: 3=not confident). "I usually do well in mathematics", "I learn things quickly in mathematics" and "My teacher tells me I am good at mathematics" and six more statements are used in order to formed this scale (for further details, see Martin and Mullis (2012)).

3. RESULTS

3.1. Descriptive statistics

Purified data set from missing and inaccurate values consists of 6,736 Turkish 8th grade students' information. SAS enterprise miner 5.2 software is used to find descriptive statistics, construct models and compare models accuracy. Table 2 shows some descriptive statistics for independent variables (except age) with respect to dependent. It is seen that 2112 students get the mathematics score above 500, whereas 4624 students get the score below 500. When examining the created scales, "*home educational resources*" has the highest mean for both groups. Also, while the mean of "*students confident in math*" scale is equal to 1.879 for students whose scores are above 500, the mean of this scale is 2.584 for students belongs to other group. The mean which is close to 3 refers that many students replied the frequency of their feelings about confident as "not confident". Hence, it can be said that scale of "*student confident in math*" may be an effective factor describing the achievement.

Table 2. Descriptive statistics of independent variables

Independent variable	Above 500			Below 500		
	Min-Max	Mean	Std.dev	Min-Max	Mean	Std.dev
Age	12-16	14.085	0.421	12-16	14.0205	0.595
Students like learning math	1-3	1.652	0.737	1-3	2.087	0.729
Students confident in math	1-3	1.879	0.762	1-3	2.584	0.561
Students value math	1-3	1.512	0.647	1-3	1.771	0.728
Students engaged in math	1-3	1.674	0.611	1-3	1.926	0.609
Students bullied at school	1-3	1.518	0670	1-3	1.696	0.760
Home educational resources	1-3	2.131	0.608	1-3	2.631	0.503
Total	n=2,112			n=4,624		

3.2. LR and NN results

The binary LR analysis is performed in order to find which variables are significantly differentiated in two groups of students. Variable selection procedure is determined through forward stepwise selection. The Newton-Raphson Ridge Optimization algorithm, which is default in SAS, is used to calculate maximum likelihood parameter estimates. After 4 iterations, the process is terminated. Table 3 shows the likelihood ratio test results which is used as the measure of goodness of fit of the model. According to the test results, estimated model fits the data well and the set of predictors have a significant effect on the logarithm of odds of a dependent variable at 95% confidence level (p-value<0.05). Another chi-square goodness of fit tests is Hosmer and Lemeshow and the p-value of test statistic is equal to 0.203. The null hypothesis of "no significant difference between observed values and predicted values" is rejected at 5% significance level, so the model fits the data well.

Table 3. Likelihood ratio test statistics

-2 Log Likelihood		Likelihood ratio		
The null model	Model with variables	Chi-Square	DF	p-value
3352.077	2525.401	826.6757	14	0.0001

Table 4 shows the results of maximum likelihood estimation and analysis of variable effect (in the last column). The last ordered category is chosen as the reference category for each six created scale. According to the analysis of variable effects, four predictors including "*age*", "*home educational resources*", "*students bullied at school*" and "*students confident in math*" are statistically significant at 95% confidence level. In addition to these four predictors, "*the students value mathematics*" scale statistically effects the mathematics achievement if the confidence level is taken as 90%. The parameter estimate of $\hat{\beta}_1$ and $\hat{\beta}_6$is negative and the p-values of Wald statistics with one degrees of freedom are statistically significant ($p<0.05$). Thus, it is said that these variables are statistically significant and make negative contribution to the model. On the other hand, $\hat{\beta}_5$, $\hat{\beta}_{11}$, $\hat{\beta}_{12}$ and $\hat{\beta}_{13}$ significantly increase the model fit, positively. In other words, "confidence in math" , "almost never bullied at school", "about monthly bullied at school" and "many educational resources" help to be in successful group for a student.

Table 4. LR results

Scale	Variable name	$\hat{\beta}$	SE($\hat{\beta}$)	Wald ChiSq	p-value ChiSq	Odds	p-value (effect)
	constant ($\hat{\beta}_0$)	5.556	1.3794	16.23	0.0001	258.951	---
Age	age ($\hat{\beta}_1$)	-0.395	0.0974	16.5	0.0001	0.673	0.0001
Gender	gender ($\hat{\beta}_2$)	-0.022	0.0504	0.2	0.6515	(1 vs 2) 0.955	0.6515
Students like learning math	like learning ($\hat{\beta}_3$)	0.218	0.0892	6.00	0.0143	(1 vs 3) 1.347	0.018
	somewhat like learning ($\hat{\beta}_4$)	-0.139	0.0693	4.05	0.0441	(2 vs 3) 0.941	
Students confident in math	confident ($\hat{\beta}_5$)	1.652	0.112	217.9	0.0001	(1 vs 3) 21.32	0.0001
	somewhat confident ($\hat{\beta}_6$)	-0.246	0.0731	11.34	0.0008	(2 vs 3) 3.191	
Students value math	value ($\hat{\beta}_7$)	-0.213	0.0835	6.53	0.0106	(1 vs 3) 0.754	0.008
	somewhat value ($\hat{\beta}_8$)	0.1444	0.074	3.81	0.051	(2 vs 3) 1.078	
Students engaged in math	engaged ($\hat{\beta}_9$)	-0.008	0.0931	0.01	0.9237	(1 vs 3) 1.048	0.6750
	somewhat engaged ($\hat{\beta}_{10}$)	0.064	0.0731	0.78	0.3762	(2 vs 3) 1.128	
Students bullied at school	almost never ($\hat{\beta}_{11}$)	0.243	0.0721	11.39	0.0007	(1 vs 3) 2.028	0.0001
	about monthly ($\hat{\beta}_{12}$)	0.221	0.0781	7.99	0.0047	(2 vs 3) 1.983	
Home educational resources	many resources ($\hat{\beta}_{13}$)	0.939	0.1733	29.34	0.0001	(1 vs 3) 7.462	0.0001
	some resources ($\hat{\beta}_{14}$)	0.132	0.0999	1.75	0.1862	(2 vs 3) 3.330	

Additionally, the value of odds ratio should be greater than 1 in order to mention about significant relationship between the relevant predictor and the dependent variable. For example, the odds ratio estimate for $\hat{\beta}_5$ versus $\hat{\beta}_0$ (or 1 vs. 3) is equal to 21.32. It can be interpreted as the odds of being in successful group are 21.32 times higher for a student who has confidence in math than a student who has not confidence in math. Furthermore, this parameter estimate has the highest value of odds ratio and it can be said that students' confidence in math is the most effective factor in being the successful group. The other predictors can be interpreted as the same way.

Finally, the NNs model is constructed with the same variables used for LR model. The MLP model is performed with a network which has one hidden layer that consists of three neurons. The process is terminated after 38 iterations when the convergence criteria is satisfied and the objective function is found 0.4585. The examination of the values of independent variables' normalized importance shows that of "*students confidence in math*", "*home educational resources*", "*age*" and "*students bullied at school*" are statistically significant predictors. The most important predictors among these four variables is the scale of "students confidence in mathematics".

3.2. Model comparisons

While assessing the prediction and the classification accuracy of LR and NNs methods, the holdout method is used to partition the data into two independent sets: the training and the testing data set. 40% of the data is taken as the training set in order to build the initial model. The rest is split into two as the validation (30% of data) and the test data (30% of data). The validation data set is used to adjust the initial model to make it more general and less tied to the idiosyncrasies of the training data set. The test data is also named as a holdout sample and it is used for final assessment of the model accuracy (Linoff & Berry, 2011). In the SAS model comparison tool, misclassification rate is preferred as the model selection criteria.

Classification tables (confusion matrixes) which are used to evaluate the models' predictive abilities are shown in Table 5 and 6. Through these tables, the accuracy, the sensitivity (true positive rate) and the specificity (true negative rate) of each classifier method are also calculated. Calculation methods can be found in Han and Kamber (2006).

Table 5. Classification accuracy of the LR model

		Train data set predicted		Percent Correct	Validation data set Predicted		Percent Correct
		y=0	y=1		y=0	y=1	
observed	y=0	1668	182	61.89%	1235	152	89.04%
	y=1	390	455	31.35%	289	345	45.58%
overall				78.78%	1524	497	78.18%

Table 6. Classification accuracy of the NN model

		Train data set predicted		Percent Correct	Validation data set Predicted		Percent Correct
		y=0	y=1		y=0	y=1	
observed	y=0	1688	162	62,63%	1254	133	90,41%
	y=1	404	441	31,35%	314	320	49,53%
overall				79%	1524	497	77,88%

According to the tables, the overall percent corrects of train and validation data sets for both LR and NN methods are very close, that means that two methods give similar results for their predictive ability. While the classification accuracy of LR method is 78.5%, NN method has 78.6% accuracy rate. The sensitivities of models are calculated as 0.7 and 0.72, respectively. The specificity of LR is equal to 0.89 and the specificity of NN is equal to 0.91.

When other measures of model comparison such as misclassification rate, MSE, gain, gini coefficient and ROC index are taken into account, it is seen that both methods have similar classification and prediction ability. Table 7 indicates fitting statistics for the test data. It can be concluded that given statistics exhibit the same pattern across LR and NN models.

Table 7. Some fitting statistics for the comparisons of models

Method	Misclassification rate	MSE	Gain	Gini coeff.	ROC index
LR	0.21	0.15	181.2	0.62	0.81
NN	0.21	0.15	187.2	0.61	0.80

The classification similarities between two models are visualized in the ROC curves given below (Figure 2). ROC curves for the train, the validate and the test data look highly similar.

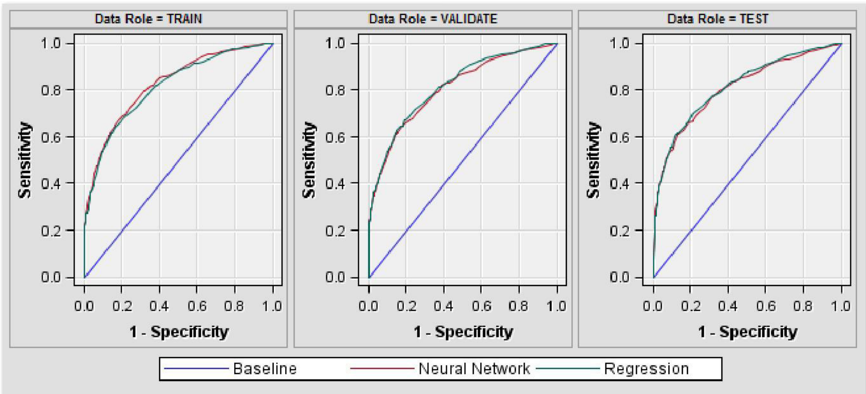

Fig. 2. ROC curves for LR and NN models

4. Conclusion

In this study, 6736 Turkish 8th grade students' information from TIMSS 2011 database is used for two main purposes. Firstly, it is aimed to determine the significance of selected factors on mathematics achievement. Then, two popular data mining methods, namely binary LR and NNs are performed and the prediction and the classification effectiveness of these two methods based on different model fitting statistics are implemented.

Both LR and NN results show that "*students confidence in math*" scale is the most important predictor on mathematics achievement. Also a student who "almost never" or about monthly bullied at school" and who has "many educational resources" is included in successful group. On the other hand, when the prediction accuracy of models are compared, it is concluded that two methods give similar overall percent corrects. When some measures of model comparison such as misclassification rate, MSE, gain, gini coefficient and ROC index are examined, it is seen that predictive and the classification accuracy of LR versus NN has similar performance for the data used in this study. One reason of this similarity could be that nonlinearity does not exist between the dependent variable and the selected six independents. Because of the black box nature, unlike LR, NN cannot explain how it produces the model. LR enables researchers to calculate confidence intervals and to develop hypothesis tests in order to make decisions based on the statistical theory. Thus, LR appears to be the preferred model choice for this study.

References

Akiba, M., LeTendre, G., & Scribner, J. (2007). Teacher quality, oppotunity gap, and national achievement in 46 countries. *Educational Research , 36*, 369-387.

Alpar, R. (2011). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler* (3. Baskı b.). Ankara: Detay Yayıncılık.

Aypay, A., Erdoğan, M., & Sözer, M. (2007). Variation among schools on classroom practices in science based on TIMSS-1999 in Turkey. *Studies in Educational Evaluation , 44* (10), 1417-1435.

Dodeen, H., Abdelfattah, F., Shumrani, S., & Hilal, A. M. (2012). The effects of teachers' qualifications, practices and perceptions on student achievement in TIMSS mathematics: A conparison of two countries. *International Journal of Testing , 12* (1), 61-77.

Giudici, P., & Figini, S. (2009). *Applied Data Mining for Business and Industry* (Second Edition b.). 2009 John Wiley & Sons, Ltd.

Han, J., & Kamber, M. (2006). *Data Mining:Concepts and Techniques* (Second Edition b.). Morgan Kaufmann .

Hanushek, E., & Luque, J. (2003). Efficiency and equity in schools around the world. *Economics of Education Review , 22* (5), 481-502.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second Edition b.). Hoboken: John Wiley and Sons, Inc.

Jürges, H., & Schneider, K. (2004). International differences in student achievement: An economic perspective. *German Economic Review , 5* (3), 357-380.

Linoff, G., & Berry, M. (2011). *Data Mining Techniques: For Marketing, Sales and Customer Relationship Management* (Third Edition b.). NY: Wiley.

Martin, M., & Mullis, I. (2012). Methods and procedures in TIMSS and PIRLS 2011. Chestnut Hill, MA: TIMSS&PIRLS International Study Center, Boston College.

Martin, M., Mullis, I., Beaton, A., Gonzalez, E., Smith, T., & Kelly, D. (1997). *Science achievement in the primary school years: IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Boston College.

Mills, J., & Holloway, E. (2013). The development of statistical literacy skills in eighth grade: exploring the TIMSS data to evaluate student achievement and teacher characteristics in the United States. *Educational Research and Evaluation: An International Journal on Theory and Practice , 19* (4), 323-345.

Mullis, I., Martin, M., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics.* Chestnut Hill, MA, USA: TIMSS & PIRLS International Study Center.

Neuschmidt, O., Barth, J., & Hastedt, D. (2008). Trends in gender differences in mathematics and science (TIMSS 1995-2003). *Studies in Educational Evaluation , 34* (2), 56-72.

Papanastasiou, C. (2000). Internal and external factors affecting achievement in mathematics: Some findings from TIMSS. *Studies in Educational Evaluation , 26* (1), 1-7.

Pong, S., & Pallas, A. (2001). Class size and eight-grade math achievement in the United States and Abroad. *Educational Evaluation and Policy Analysis , 23* (3), 251-273.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert System with Applications , 33*, 135-146.

Schreiber, J. B. (2002). Institutional and student factors and their influence on advanced mathematics achievement. *The Journal of Educational Research , 95* (5), 274-286.

Sivanandam, S., Sumathi, S., & Deepa, S. (2006). *Introduction to Neural Networks Using Matlab 6.0.* New Delhi: Tata McGraw-Hill Publishing Company.

Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications , 39*, 9468-9476.

Webster, B., & Fisher, D. (2003). School environment and students outcomes in mathematics. *Learning Environments Research , 6*, 309-326.

Woessmann, L. (2005). Educational production in East Asia: The impact of family background and schooling policies on student performance. *German Economic Review , 6* (3), 331-353.

Woessmann, L. (2004). *How equal are educational opportunities?Family background and student achievement in Europe and the United States.* University of Munich, Ifo Institute for Economic Research. CESifo Working Paper No.1162.

Woessmann, L. (2003). Schooling resources, educational institutions and student performance: The international evidence. *Oxford Bulletin of Economics and Statistics , 65* (2), 117-170.

Yayan, B., & Berberoğlu, G. (2004). A re-analysis of the TIMSS 1999 Mathematics Assessment data of the turkish students. *Studies in Educational Evaluation , 30* (1).