PREDICTING STUDENTS' ACADEMIC PERFORMANCE: COMPARING ARTIFICIAL NEURAL NETWORK, DECISION TREE AND LINEAR REGRESSION

ZAIDAH IBRAHIM
Faculty of Information Technology and Quantitative Sciences
Universiti Teknologi MARA Malaysia
zaidah@tmsk.uitm.edu.my

DALIELA RUSLI
Faculty of Information Technology and Quantitative Sciences
Universiti Teknologi MARA Malaysia
daliela@hotmail.com

Predicting students' academic performance is critical for educational institutions because strategic programs can be planned in improving or maintaining students' performance during their period of studies in the institutions. The performance of the academic performance in this study is measured by their cumulative grade point average (CGPA) upon graduating. In this study, the students' demographic profile and the CGPA for the first semester of the undergraduate studies are used as the predictor variable for the students' academic performance in the under-graduate degree program. Three predictive models have been developed using SAS Enterprise Miner, that are, artificial neural network, decision tree and linear regression. The result of this study shows that all of the three models produce more than 80% accuracy. It also shows that artificial neural network outperforms the other two models.

Keywords: SAS Enterprise Miner

1. Introduction

The main product of universities is students. Upon graduation, the students may either continue their studies into the post-graduate program or become the manpower for the industry, government and private sectors. Thus, the students' performances are critical in ensuring the supply chain is fulfilled. Research in examining students' performance has been done in [1], [2] and [3] using statistical analysis. Artificial Neural Network (ANN) has been used to predict the students' success in [4] while a comparative study between ANN and statistical analysis for predicting final GPA of students has been in [5]. This study examines three models, that are, linear regression, decision tree (DT) and artificial neural network using SAS Enterprise Miner in predicting the final cumulative grade point average (CGPA) of the students upon graduation. Results arising from this study provide important reference materials for the planning of the future success of the students and faculty.

2. Model building and selection

206 students' data were obtained for this study. The correlation coefficient analysis is carried out in order to determine the relationship of the independent variables, that are, information technology application knowledge, previous school (boarding or non-boarding), programming knowledge and family financial status, with the dependent variable, which is the students' CGPA in the final semester. A correlation of 0.87654 is obtained which shows that the most significant input variable is the final CGPA.

Model selection is based on the square root of average squared error (RASE) for the validation data set.

3. Predictive Modeling Using Decision Tree

In this predictive modeling segment, the researcher examined 2-way split and 4-way split tree. The DT model is built on the significance level 0.1. Significance level 0.1 specifies the threshold p-value for the worth of the candidate splitting rule. For interval targets, PROBF (p-value of F test associated with the node variances) is used to specify the method of searching for and evaluating candidate splitting rules.

Table 1 summarize the variable name and label, the number of rules (or splits) in the tree that involve the variable (NRULES), the importance of the variable computed with the training data (IMPORTANCE), the importance of the variable computed with the validation data (VIMPORTANCE), and the ratio of VIMPORTANCE to IMPORTANCE. Variable CGPA is the only input variable that appears in the output window indicating that it is the most important variable.

Table 1: Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE	VIMPORTANCE	RATIO
1	CGPA	CGPA	5	1	1	1

Table 2 shows the tree leaf report. The tree leaf report shows that the tree has six leaves. The leaves in the table are in order from the largest number of training observations to the fewest training observations. Table 3 shows the RASE for different number of branch used.

Table 2: Tree Leaf Report

Node	Depth	Training Observations	Training Average	Validation Observations	Validation Average	Training Root ASE	Validation Root ASE
5	2	43	2.76	13	2.68	0.21501	0.20807
4	2	40	2.33	9	2.41	0.24822	0.14288
9	3	24	3.19	9	3.13	0.16742	0.12594
10	3	22	3.41	6	3.35	0.22327	0.17684
8	3	21	3.01	12	3.02	0.18231	0.20259
11	3	5	3.82	2	3.93	0.05276	0.11400

Table 3: RASE values for DT

Model	Maximum branch	Root average squared error (RASE)
1	2	0.1769 (max depth = 6)
2	4	0.1769 (max depth = 6)

As a conclusion, the RASE for both tree are similar, 0.1769. It is discovered that the tree diagram is similar for both 2-way split and 4-way split tree.

4. Predictive modeling using artificial neural network

Different ANN models are utilized in order to gauge the smallest RASE. Table 4 summarizes the findings of different ANN models. Randomization scale specifies the scale parameter for random numbers. Current estimation specifies whether the current estimates should be used as starting values for training. According to Table 4, the best ANN model is Model 1. Figure 1 is the graph of score ranking overlay for ANN model when predicting the final CGPA. Randomization scale specifies the scale parameter for random numbers. Current estimation specifies whether the

current estimates should be used as starting values for training. According to Table 4, the best ANN model is Model 1.

Table 4: RASE values for ANN.

Model	Number of hidden units	Current estimation = no	Current estimation= yes
1	5	0.1714, if randomization = 0.5; 0.1714	0.1918, if randomization = 0.5; 0.1918
2	10	0.1730, if randomization = 0.5 ; 0.1730	0.2275, if randomization= 0.5; 0.2275
3	15	0.1857; if randomization = 0.5 ; 0.1857	0.1766, randomization=0.5; 0.1780
4	20	0.1719; if randomization = 0.5; 0.1719	0.1766; if randomization= 0.5; 0.1715

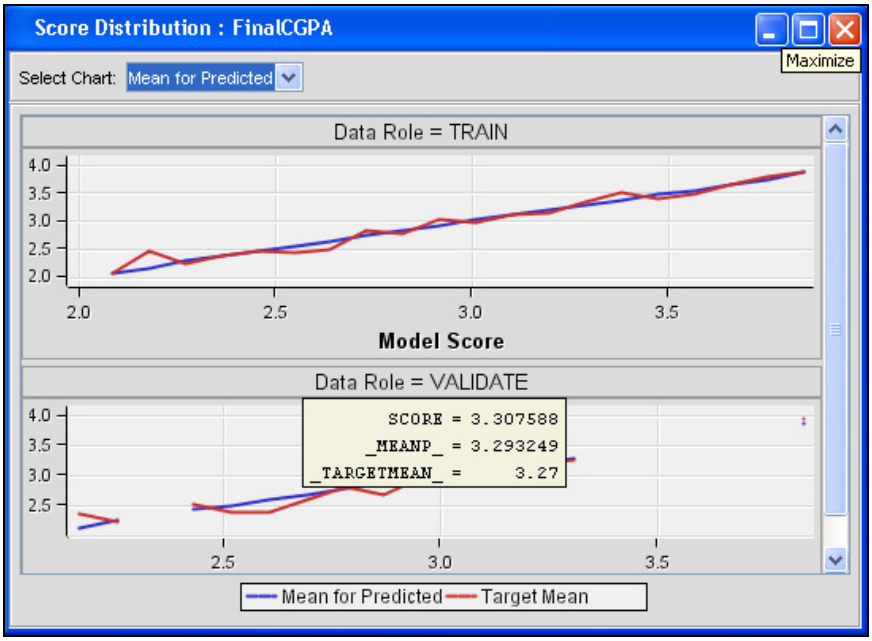

Figure 1: The score distribution plots for ANN

5. Predictive modeling using regression

Linear regression is used for interval target variable. Figure 2 shows the effects plot which displays a bar chart of the absolute values of the regression coefficients for each independent variable in the regression model. Positive and negative values are differentiated by different colors of the bars.

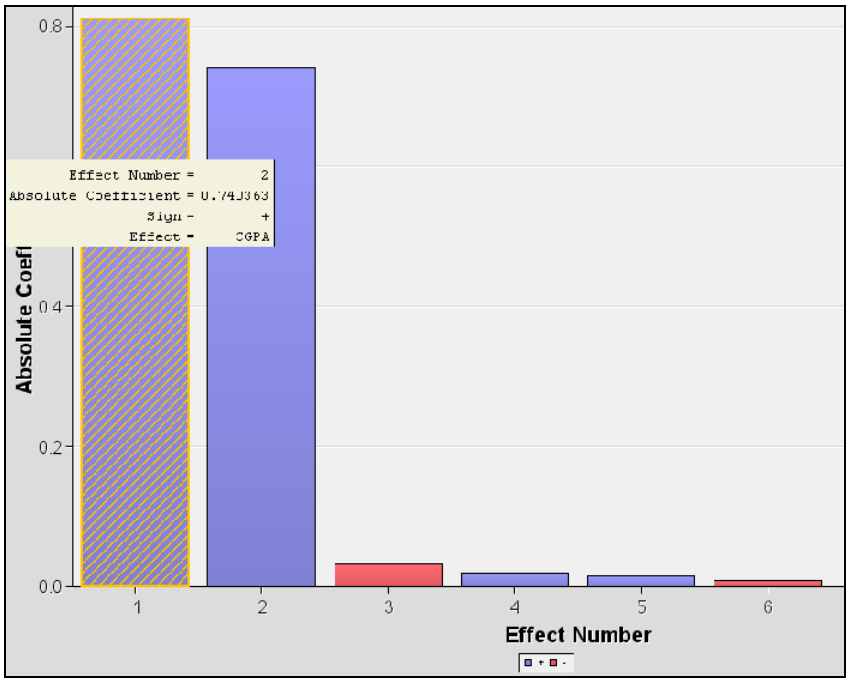

Figure 2: Effects plot

According to the effects plot the variable with the largest absolute coefficient is the intercept. Thus, the variable Intercept is the most important variable in this model. Table 5 shows list of variables entered accordingly. The RASE for validation data set is 0.1848. Figure 3 shows the score rankings overlay of the liner regression model.

Table 5: Effects plot

Variable	Absolute coefficient
1. Intercept	0.8103
2. CGPA	0.7404
3. bschool 1	-0.323
4. prior it	0.02
5. priorCP 1	0.0142
6. finance1	-0.0087

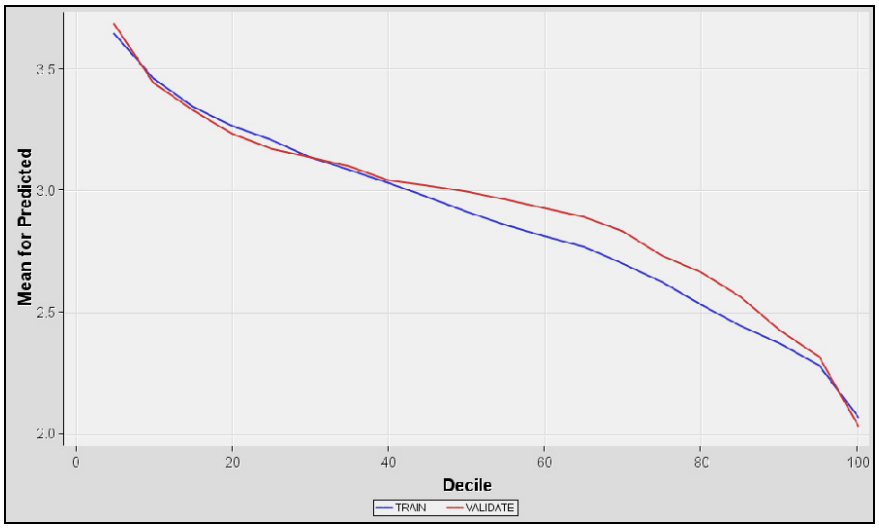

Figure 3: Score rankings overlay for linear regression model

6. Results and analysis

Table 6 shows the findings after model comparison node is added. Based on the validation data set, the smallest RASE is produced by ANN model which is 1.714, followed by DT model which is 0.1769 and linear regression which is 0.1848. Figure 4 shows the overlay score ranking plot

Table 6: RASE of ANN, REG, and DT.

RASE	ANN	REG	DT
Training	0.208	0.220	0.212
Validation	0.1714	0.1848	0.1769

7. Conclusion

Figure 4 compares assessment statistics for the various models across the various deciles computed the training and validation data sets. From the validation data plot, ANN model is better than the other two models in the first 10th deciles. In conclusion, the best model in predicting the final CGPA of the students upon graduation is ANN.

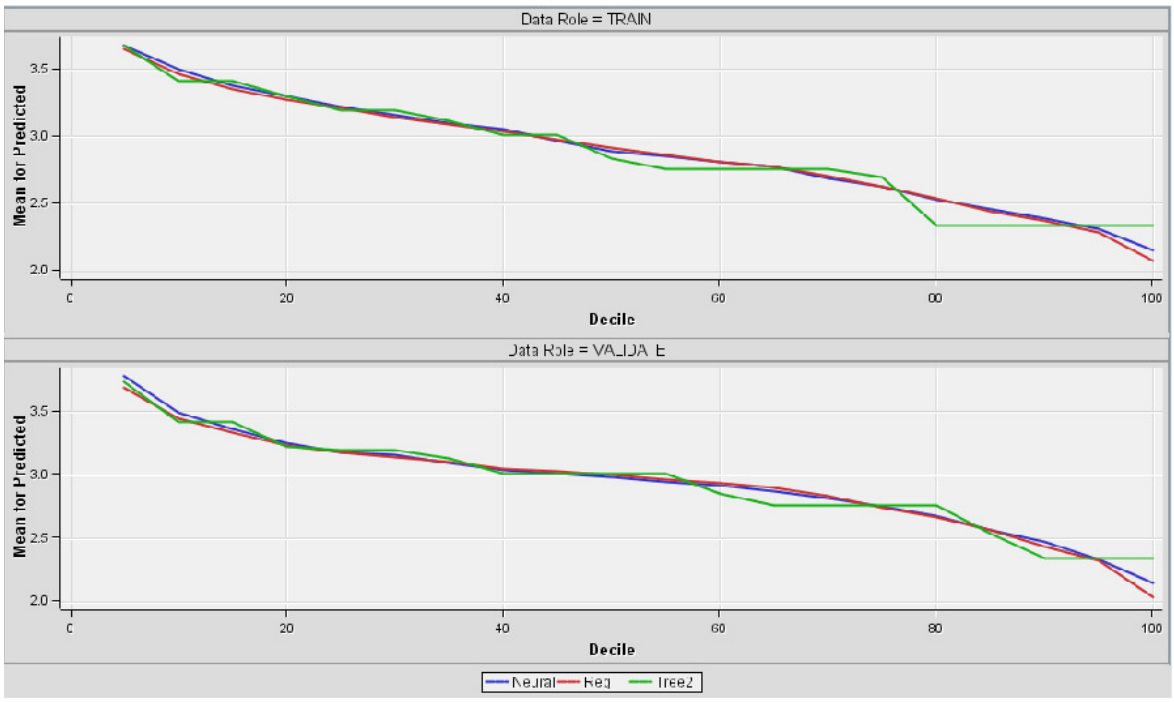

Figure 4: Score rankings overlay for model comparison

References

[1] Alfan, E. and Othman, M.N. Undergraduate Students' Performance : The Case of University of Malaya, Quality Assurance in Education, 2005, vol. 13, no. 4, pg. 329 – 343.

[2] Nelson, C. Van; Nelson, J. S. And Malone, B. G. Predicting Success of International Graduate Students in an American University, College and University, 2004, vol. 80, no. 1, pg. 19 – 27.

[3] Frantz, P. L. and Wilson, A. H. Student Performance in the Legal Environment Course: Determinants and Comparisons, Journal of Legal Studies Education, 2004, vol. 21, no, 2, pg. 225 – 239.

[4] Naik, B. and Ragotiaman, S. Using Neural Netwroks to Predict MBA Student Success, College Student Journal, 2004, vol. 38, no. 1, pg. 143 – 149.

[5] Anderson, J. L. Predicting Final CPA of Graduate School Students: Comparing Artificial Neural Networking and Simultaneous Multiple Regression, College and University, 2006, vol. 81, no. 4, pg. 19 – 29.