PREDICTION OF STUDENTS' FINAL RESULTS USING JMP PRO 12

Thai Cao, Juana Rodriguez
Oklahoma State University, Stillwater, OK 74075

Abstract

There are a host of factors that can affect school results of teenagers. By understanding those factors, schools and parents can provide adequately timely supports, which may help students improve their school grades, and have continued better performances in future. We have obtained data of 1,044 students in two Portugal public secondary schools ("GP" - Gabriel Pereira and "MS" - Mousinho da Silveira), from UCI Machine Learning Repository. The original dataset contains 33 input variables and 1 nominal target variable identified as "G" which contains the average final grade of students. In Portuguese middle school, a student passes a class if he gets 50% or more in the final result. For the purposes of our analysis, we have created a binary target out of the average final grades from the original data set. The new target variable is called "Final" and describes if the students have pass or failed the class. There are no missing values in the dataset; no multicollinearity was observed among input variables, and important input variables for model building were selected using chi-square test.
In this project, we use JMP 12 Pro to develop models (Regression, Decision Tree, and Neural Network) to predict if students pass or fail their classes, based on their personal characteristics, habits, family and demographic characteristics, and school performances. Such models will not only help predict students' results but also provide insights about how each specific factor impacts the performance of students.

Data Preparation and Initial Analysis

In order to have better understanding of the data, we have conducted some initial analysis, using ANOVA tests and distribution analysis of students' average final grade separated by different categorical variables:

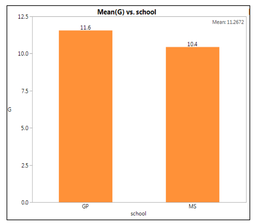

Figure 1. Average final grade "G" by school

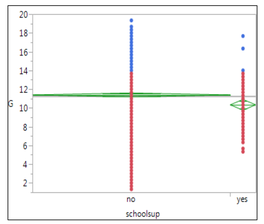

Figure 2. Average Final grade "G" by extra educational school support

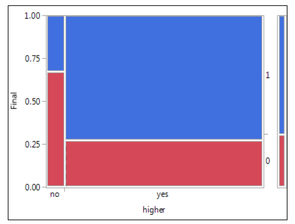

Figure 3. Contingency analysis of finale grade, "Final", by Higher (whether students want to take higher education)

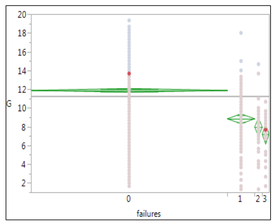

Figure 4. ANOVA analysis for average final grade "G" by number of past failures

Predictive Modeling

The data has been split into 60% training data and 40% validation data. The final model is chosen based on the validation metrics. For predictive modeling, we have built three different models taking "Final" as the target variable. We used the Model Comparison platform to compare the performance of the following models:

Measures of Fit for Final

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Fit Nominal Logistic		0.2005	0.3092	0.4933	0.4017	0.3221	0.2280	1044
Partition		0.1881	0.2923	0.501	0.4063	0.3318	0.2385	1044
Neural		0.1869	0.2905	0.5018	0.4064	0.3180	0.2328	1044

Figure 5: Measure fit of statistics from Model Comparison

1. Stepwise Logistic Regression model
2. Neural Network with one hidden layer and three activity functions
3. Decision tree with settings set for automatic splitting

The Model Comparison platform provides a report with the measures of fit of each of the three models. We used validation RMSE (Root Mean Square Error) and Misclassification Rate to select the best model.

Result

After comparing the models built, we found Stepwise Logistic Regression to be the best model given that it has the lowest RMSE and the lowest Misclassification Rate among all three models (Figure 5). Using a p-value threshold of 0.05, the stepwise process chose 8 significantly important variables (Figure 6).

Effect Summary

Source	LogWorth	PValue
failures{0-1&2&3}	15.846	0.00000
paid	4.138	0.00007
schoolsup	3.838	0.00015
school	3.560	0.00028
higher	3.069	0.00085
Fedu{0-1}	2.615	0.00242
studytime{1&2-3&4}	2.358	0.00438
freetime{1&2-3&4&5}	1.362	0.04349

Figure 6. Stepwise Variable Selection

Figure7: Odds Ratio

Odds Ratios

For Final odds of 0 versus 1
Tests and confidence intervals on odds ratios are likelihood ratio based.

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Fedu{0-1}	0.512631	0.332945	0.78851	1.9507224
studytime{1&2-3&4}	1.435415	1.116754	1.871104	0.6966628
failures{0-1&2&3}	0.368637	0.28555	0.470744	2.7126955
freetime{1&2-3&4&5}	0.776199	0.600101	0.992751	1.2883296

Odds Ratios for school

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
MS	GP	2.2635416	0.0003*	1.4588235	3.519415
GP	MS	0.4417856	0.0003*	0.2841381	0.6854839

Odds Ratios for schoolsup

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
yes	no	3.0269056	0.0001*	1.7174627	5.3355883
no	yes	0.3303704	0.0001*	0.1874208	0.5822543

Odds Ratios for paid

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
yes	no	2.5607127	<.0001*	1.6113685	4.0802361
no	yes	0.3905163	<.0001*	0.2450839	0.6205905

Odds Ratios for higher

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
yes	no	0.3135909	0.0009*	0.1562632	0.6199804
no	yes	3.1888677	0.0009*	1.6129543	6.3994594

The odds ratio (Figure 7) has indicated some interesting results:

- The odds of fail of students in MS School are 2.26 times greater than that of students in GP School.
- Those who want to pursuit higher education have greater chance of passing their classes.
- Those who had failed in past classes are 3.6 times more likely to fail than those who have never fail a class before.
- Those who study less than 5 hours a week are 1.43 times more likely to fail a class than those who study more than 5 hours.
- Students who have less free time after school are more likely to fail a class (Those, who have less spare time after school, might be more stressful than their peers, which may lead to bad result. However, we need to study more about students' after-school activities to get a better explanation)
- Those who had extra educational support, that is extra paid classes from their schools, have higher chance of failing their classes. (Those students might come from rich households, do not appreciate the support they have received, and put less effort on studying. But, we still need more research to get more specific evidences)

Conclusion

- Logistics Regression, Decision Tree, and Neural Network models were built to anticipate students' performances, based on their personal, family, and demographic information. Model comparison algorithm was utilized to compare 3 models, and Logistics Regression was selected as the winner.
- The predictive model has reasonable accuracy level (nearly 80%).
- By considering the projected results, and effects of different variables, schools and families can provide suitable and timely support to those who might fail their classes
- Our analysis was, however, limited by the small size of available data set. A better result could be achieved by increasing the number of observations.

References

- http://blogs.sas.com/content/jmp/2012/05/16/making-better-predictive-models-quickly-with-jmp/
- https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION

Acknowledgement

We would like to thank Dr. Goutam Chakraborty – Professor in Marketing Department and Founder of Marketing Analytics Certificate Program, Oklahoma State University for his generous support throughout this project.