

STUDENT DROP OUT FACTOR ANALYSIS AND TREND PREDICTION USING DECISION TREE

Jeeranan Chareonrat*

Received: November 05, 2015; Revised: February 05, 2016; Accepted: March 24, 2016

Abstract

Issues relating to increases in student drop-out rates are becoming a top priority in many educational institutions. This paper aims to identify and explore the factors influencing this growing phenomenon focusing on a university in provincial Thailand. Research conducted between 2010 and 2014 targeted Management Science students attending Sakon Nakhon Rajabhat University. Survey database on 14 attributes of 4,163 current students. Data analysis was undertaken using algorithm J48 Data Mining techniques with a decision- tree classification and Weka's 10-fold cross validation program. The findings of the research indicated that the four most significant factors that induced student drop-out were low GPA results, studying loans, earlier educational attainment, and parents' monthly incomes. Further analysis indicated that in the 2010-2011 year low GPA attainment was the most significant factor, and added with studying loans in 2012 to 2013 then plused parents' incomes in 2014. This suggests a trend in line with the Classification Rule that may predict drop-out rates in the current year 2015.

Keywords: Data mining, Classification, Prediction, Student drop out

Introduction

It is widely known that Information Technology (IT) which brings about efficient working process and decision making has been skyrocketed developing these days. It plays a vital role in most organization for manipulating and collecting data of huge databases. Educational institutions, for example, have stored many

aspects of information including, students' data, and lectures' websites and e-learning administration. Despite, obtaining a large amount of data, they seldom made use of those data for other benefits especially for prediction analysis. Data mining is the process of discovering interesting patterns and knowledge

Department of Business Computer, Faculty of Management Science, Sakon Nakhon Rajabhat University, Sakon Nakhon 47000, Thailand. Tel. 0-4297-0028; Fax. 0-4297-0028; E-mail: jeeranan.ch@snru.ac.th

** Corresponding author*

from large amount of data (Han *et al.*, 2011). Gulati (2015) predicted student' drop out by using data mining technique, Yukselturk *et al.* (2014) predicted students' drop out of the On-line program using K-Nearest Neighbour(K-NN), Decision Tree(DT), Naive Bayes (NB) and Neural Network (NN). Omkar and Parag (2015) use Data mining J48, Random Forest, Rep Tree and BF Tree of Decision Tree and JRip rule.

The Management Science faculty has also been facing with students' drop out which is considered to be an important problem in the educational system as it directly affects organization's budget management. Some data from the Education promotion Department revealed that the drop out rate was 20.68 percent in between year 2010-2014 which considered "high". This study aims to form a model that can predict factors affecting annual students' drop out by using data mining technique. This suggests a trend that contributes administrators to help prevent students from dropping out.

Materials and Methods

This research was conducted according to Cross-Industry Standard Process for Data Mining (CRIPS-DM) (Chapman *et al.*, 2000). The process of study was as followings :

1. Business Understanding: By gathering student data from the educational Promotion department and studying related research, there were 14 significant attributes. The study was started by Attribute-Class relationship analyzing. Attributes that suitable to analyze the drop out rate shown in Table 1.

2. Data Understanding: Researcher selected 4163 datasets of the management science students stored in the department of academic promotion's databases system during the year 2010-2014 then classified by year as shown in Table 2.

3. Data preparation: Steps in preparing data to use with the WEKA program (Bouckaert *et al.*, 2013) shown below.

3.1) Data Cleansing: Initial data with

Table 1. Student related Variables

Variable	Description	Possible Values
Sex	gender	{F,M}
Province	habitation	{Sakon, Nakon, Mukda, Out_sanok}
Occup_farther	Occupation of father	{Occ_far0, Occ_far1, Occ_far2, Occ_far3, cc_far4, Occ_far5, Occ_far6, Occ_far7, Occ_far8, Occ_far9}
Revenue_far	Incomes of father	{Rev_far1, Rev_far2, Rev_far3, Rev_far4, Rev_far5}
occup_mother	Occupation of mother	{Occ_mom0, Occ_mom1, Occ_mom2, Occ_mom3, Occ_mom4, Occ_mom5, Occ_mom6, Occ_mom7, Occ_mom8, Occ_mom9}
Revenue_mom	Incomes of mother	{Rev_mom1, Rev_mom2, Rev_mom3, Rev_mom4, Rev_mom5}
Parent_status	Status of parents	{Par_st1, Par_st2, Par_st3, Par_st4, Par_st5, Par_s61, Par_st7, Par_st8, Par_st9, Par_st10}
GPA_school	Average GPA from secondary school	{Weak, Medium, Good, Best, Excellent}
Old_Edu	Educational attainment from secondary school	{Old_Edu1, Old_Edu2, Old_Edu3, Old_Edu4}
curriculum	Learning program	{Curriculum2, Curriculum4}
Major	Subject oriented	{Major1, Major2, Major3, Major4, Major5, Major6, Major7, Major8, Major9, Major10}
GPA	Average GPA	{Weak, Medium, Good, Best, Excellent}
Loan	Studying loans	{Yes, No}
Drop Out	Learning abandon	Class
		{Yes, No}

missing value, error or noisy data were dropped at first step of cleaning up data. This has left 4163 data sets out of the total 4366.

3.2) Data Adjusting: Raw data consisted of both numerical and alphabetic aspects so they were needed to adjust into a common from that can be analyzed.

4. Modeling and selecting the right technique: Classification Data mining technique was used to form a model and used Decision Tree algorithm J48 to predict trends then used WEKA's 10-fold Cross Validation to specify results test form.

5. Evaluation:

5.1) K-fold cross validation (Hastie *et al.*, 2008)

divide datasets into K equal parts
use K-1 parts to form train set
use the rest datasets to be a test set
repeat the process until every dataset brought to test

5.2) Accuracy (Mohammed and Wagner Meira, 2014)

The accuracy of a data can be calculated into percentage by using formula as following

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$TP Rate = \frac{TP}{(TP+FN)} \quad (2)$$

$$TN Rate = \frac{TN}{(TN+FP)} \quad (3)$$

$$FP Rate = \frac{FP}{(FP+TN)} \quad (4)$$

$$FN Rate = \frac{FN}{(TP+FN)} \quad (5)$$

where as : TP is a True Positive value
TN is a True Negative value
FP is a False Positive value
FN is a False Negative value

6. Development: Bringing a qualified model to modify and use as a predicting tool for the year 2015.

Results and Discussion

The researcher brought datasets of each year to from a model using classification data mining technique with algorithm J48 Decision Tree to predict the data and The WEKA 3.7.5 10-fold Cross Validation to specify result testing form. This model was employed to seek factors affecting students drop-out and to predict changing trends of those factors in each year. Figures 1 to 5 identified results and Classification Rule and Table3 expressed True values.

Table 2. datasets by year

Academic years	Data sets
2010	715
2011	902
2012	874
2013	775
2014	897
Total	4,163

Table 3. accuracy comparison by year

Year	Classifier J48				
	2010	2011	2012	2013	2014
Accuracy	92.59%	96.23%	94.39%	96.90%	95.42%
TP Rate	0.998	0.994	0.984	0.976	0.967
FP Rate	0.366	0.136	0.198	0.052	0.135
TN Rate	0.634	0.864	0.802	0.748	0.868
FN Rate	0.002	0.006	0.016	0.024	0.033

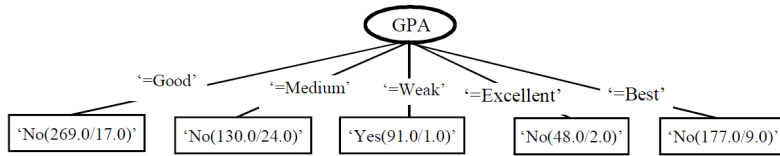


Figure 1. Rule model produced by J48 Decision Tree Year 2010

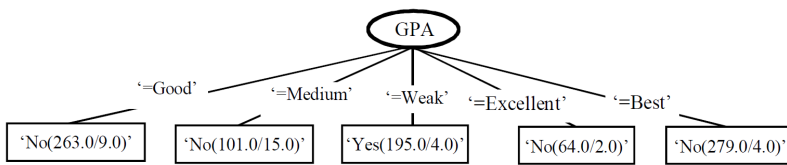


Figure 2. Rule model produced by J48 Decision Tree Year 2011

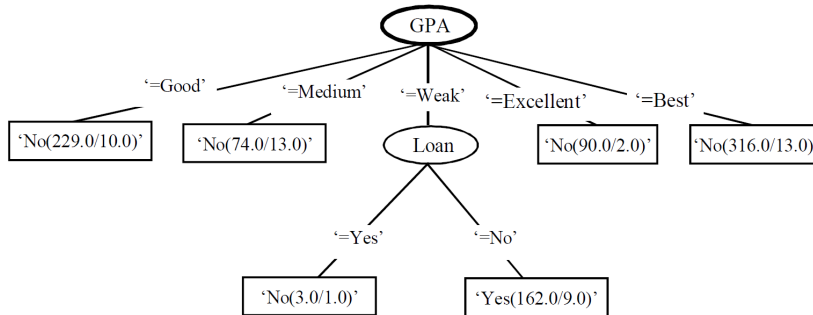


Figure 3. Rule model produced by J48 Decision Tree Year 2012

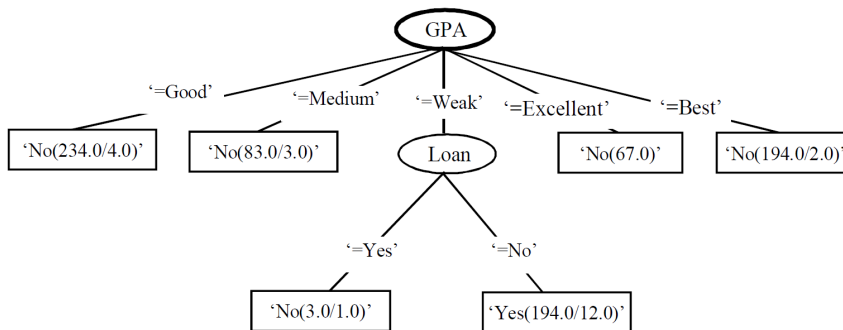


Figure 4. Rule model produced by J48 Decision Tree Year 2013

According to Figure 1 Rules of Decision from Decision Tree in year 2010 were:

IF GPA = Weak THEN student Drop Out

IF GPA = Medium THEN student not Drop Out

IF GPA = Good THEN student not Drop Out

IF GPA = Best THEN student not Drop Out

IF GPA = Excellent THEN student not Drop Out

The important Decision Rule in year 2010 indicated the average GPA of less than 2.00 as the affecting factor to students' drop out.

According to Figure 2 Rules of Decision from Decision Tree in year 2011 were:

IF GPA = Weak THEN student Drop Out

IF GPA = Medium THEN student not Drop Out

IF GPA = Good THEN student not Drop Out

IF GPA = Best THEN student not Drop Out

IF GPA = Excellent THEN student not Drop Out

The important Decision Rule in year 2011 indicated the average GPA of less than 2.00 as the affecting factor to students' drop out.

According to Figure 3 Rules of Decision from Decision Tree in year 2012 were:

IF GPA = Weak AND Loan = No THEN student Drop Out

IF GPA = Weak AND Loan = Yes THEN student not Drop Out

IF GPA = Medium THEN student not Drop Out

IF GPA = Good THEN student not Drop Out

IF GPA = Best THEN student not Drop Out

IF GPA = Excellent THEN student not Drop Out

The important Decision Rule in year 2012 indicated the average GPA of less than 2.00 and not being allowed to receive student loans as factors affecting students' drop out.

According to Figure 4 Rules of Decision from Decision Tree in year 2013 were:

IF GPA = Weak AND Loan = No THEN student Drop Out

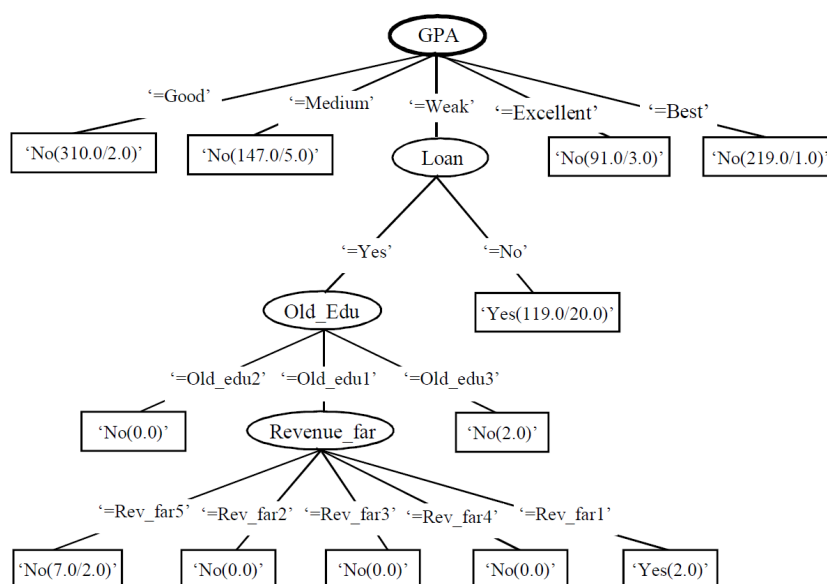


Figure 5. Rule model produced by J48 Decision Tree Year 2014

IF GPA = Weak AND Loan = Yes THEN student not Drop Out	revenue_far =Rev_far5 THEN student not Drop Out
IF GPA = Medium THEN student not Drop Out	IF GPA = Weak AND Loan = Yes AND Old_Edu = Old_Edu2 THEN
IF GPA = Good THEN student not Drop Out	student not Drop Out
IF GPA = Best THEN student not Drop Out	IF GPA = Weak AND Loan = Yes AND Old_Edu = Old_Edu3 THEN
IF GPA = Excellent THEN student not Drop Out	student not Drop Out
The important Decision Rule in year 2013 indicated the average GPA of less than 2.00 and not being allowed to receive student loans as factors affecting students' drop out.	IF GPA = Medium THEN student not Drop Out
According to Figure 5 Rules of Decision from Decision Tree in year 2014 were:	IF GPA = Good THEN student not Drop Out
IF GPA = Weak AND Loan = No THEN student Drop Out	IF GPA = Best THEN student not Drop Out
IF GPA = Weak AND Loan = Yes AND Old_Edu = Old_Edu1 AND	IF GPA = Excellent THEN student not Drop Out
revenue_far =Rev_far1 THEN student Drop Out	
IF GPA = Weak AND Loan = Yes AND Old_Edu = Old_Edu1 AND	The important Decision Rule in year 2014 indicated the average GPA of less than 2.00, being allowed to receive student loans, high school graduation, and father's monthly incomes of less than 12,500 bath as factors affecting students' drop out.
revenue_far =Rev_far2 THEN student not Drop Out	According to Table 3 the predicting model obtained an accuracy value of more than 90 percent (92.57% - 96.90%). Therefore, the year 2013 obtained the highest accuracy value of 96.90% Figure 1 to 5 can be concluded factors affecting students' drop out by year shown in Table 4.
IF GPA = Weak AND Loan = Yes AND Old_Edu = Old_Edu1 AND	The Table 4 showed the average GPA of less than 2.00 displayed the mutual factor though out 5 years that went in accordance with Thinsungnoen <i>et al.</i> (2012) and the team's study. Study loans was the factor added in the year 2012 and 2013 and the 2 new factors i.e. students' background educational attainment and father's monthly incomes of less than 12500 Bht. were found in the year 2015 in line with the study of Pahannarat <i>et al.</i> (2009) and
revenue_far =Rev_far3 THEN student not Drop Out	
IF GPA = Weak AND Loan = Yes AND Old_Edu = Old_Edu1 AND	
revenue_far =Rev_far4 THEN student not Drop Out	
IF GPA = Weak AND Loan = Yes AND Old_Edu = Old_Edu1 AND	

Table 4. the affecting factors by year

Year	Affecting factors
2010	The average GPA of less than 2.00
2011	The average GPA of less than 2.00
2012	The average GPA of less than 2.00, not being allowed to receive student loans
2013	The average GPA of less than 2.00, not being allowed to receive student loans
2014	The average GPA of less than 2.00, being allowed to receive student loans, high school graduation, and father's monthly incomes of less than 12,500 Baht.

team which had found that students from poor families trended to leave classroom to make ends meet for families.

Conclusions

The results of this study can be concluded as follow

1. There were 4 important factors affecting students' drop-out including the average GPA, student loans, earlier educational attainment and father's monthly incomes. Administrators and advisors can use these factors to plan and encourage students to be able to finish their studying.

2. The changing trends of factors from year 2010 to 2011 was the average GPA, from year 2012 to 2013 were the average GPA and student loans and for the student loans and plused with earlier educational attainment, and father's monthly incomes in 2014.

This suggests a classification rule to be developed for the year 2015 prediction.

Acknowledgments

This research was funded by the Faculty of Management Science, Sakon Nakhon Rajabhat University in fiscal year 2015. The author acknowledges the Office of Academic Promotion and Registrarion, Sakon Nakhon Rajabhat University for providing the data used in this research. The author also acknowledges Assoc. Prof. Dr. Kittisak Kerdprasop and Assoc. Prof. Dr. Nittaya Kredprosob from the Suranaree University of Technology for their invaluable guidance and consulting about the research on data mining techniques. The author would like to express the appreciation and gratitude for all supports and assistances.

References

- Bouckaert, R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., and Scuse, D. (2013). WEKA Manual for Version 3-7-8, University of Waikato, Hamilton, New Zealand, 327p.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide, Technical report, SPSS inc, USA, 78p.
- Gulati, H. (2015). IEEE Conference Publications; March 11-13, 2015; New Delhi, India, p. 713-716.
- Han, J., Kamber, M., and Pei, J. (2011). Data Mining Concepts and Techniques. 3rd ed. Elsevier, USA, 703p.
- Hastie, T., Tibshirani, R., and Friedman, J. (2008). The Elements of Statistical Learning Data Mining, Inference, and Prediction. 2nd ed. Springer, 739p.
- Mohammed, J., and Wagner Meira, JR. (2014). Data Mining and Analysis Fundamental Concepts and Algorithms. Cambridge University Press, NY, USA, 593p.
- Omkar, S., and Parag, M. (2015). Predicting Dropout Students Using Data-Mining Techniques. IJR J., 2(1):365-375.
- Pahannarat, N., Wangrangsimaikul, K., and Pipatpen M. (2009). The Problems of Withdrawal of the Youths Receiving Scholarship from World Vision Foundation of Thailand in Songkhla Province. PNU J., 1(3):130-144.
- Thinsungnoen, T., Kulnawin, K., and Thinsungnoen, M. (2012). Payap University Research Symposium 2012; February 17, 2012; Chiang Mai, Thailand, p. 34-42.
- Yukselturk, E., Ozekes, S., and Türel, Y. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. EURODL J., 17(1):118-133.

