**Predicting Student Performance with Decision Tree**

**Abstract**

This project aims to find out the factors that lie beyond student performance; it uses decision tree method to examine which factor affects one’s performance in what way. The datasets with thirty-three attributes have been used to give the readers a holistic view of which attribute should be considered mainly while examining student performance.

**Introduction**

When the new school year begins, people in the education system – students, parents, and teachers – operate their own student performance evaluation system in their head. While students and their parents try to evaluate their peers to survive the vicious competition, teachers assess the students to determine how much effort should be put on every single individual. This project is aimed to help the people in the education system analyze students’ performances easily, by giving them the hint of which factor should be considered mainly for one’s performance.

**Problem Statement**

The main objective of this research is to find out the factor beyond student performance. Commonly, the factor that affects one’s performance in school was thought of as one’s diligence in school: one’s attendance and time that one puts into studying, and peer relation that one has in his or her school. There has been a lot of studies conducted regarding student performance; however, it is questionable that there have been only few studies conducted that took the holistic approach upon students’ personal factors. Thus, for this research, I am going examine which attribute affects student performance the most among students’ various personal features, which will work as a great asset in predicting one’s school performance as a whole. Based on this notion, the dataset with students’ personal factors from demographic to social and study related features was chosen for analysis. Also, the research questions were set up to work as a guideline for the research.

RQ1: Among various attributes from one’s sex to parents’ jobs and relationship status, which factor does affect the student performance the most?

RQ2: Do the factors that affect student performance differ depending on the subjects?

RQ3: Are the final trees generalizable?

**Proposed Method**

To take the holistic approach among various factors regarding one’s personal life, I have chosen to create decision tree through data mining. Decision tree method is one of the classification methods that is frequently used in data mining; it is used to predict chance event outcomes with the training set based on various attributes.

**Dataset Introduction**

Two datasets were used to answer the research questions. These datasets were provided by the “Student Performance Data Set”, retrieved from the UCI Machine Learning Repository. The data were originally collected for the research that was conducted in Portugal, in the purpose of predicting the students’ grade in their final years in two distinct subjects: math and Portuguese language (Cortez & Silva, 2008).

Both datasets consist of thirty-three attributes. The first thirty attributes offer information about student’s personal life; demographic, social, and school-related features. The last three attributes represent the students’ annual score, ranging from 0 to 20.

**Detailed Steps**

The steps that I have taken for decision tree creation are as follows:

**1) Data preparation**

Before building a decision tree, I preprocessed the data in the form to fit my requirements. I checked if there were any missing values in the dataset, and added two new columns to help to build decision trees. First, as the research aimed to find the factors that would affect student’s performance in general, I added a column that reflects the one’s average score throughout their three years in secondary education. Second, as the original data was offering the students’ numeric scores as the means to examine their future performance, I set up the grading scale arbitrarily, in order to carry out classification rather than prediction. Meanwhile, I deleted one of the columns that are not relevant to the research question to avoid confusion.

|  |  |
| --- | --- |
| Original Score | Performance level Set Up for the Research |
| 17~20 | Excellent |
| 14~16 | Good |
| 11~13 | Fair |
| 8~10 | Poor |
| 0~7 | Fail |

[Figure 1] Performance level set up for the research in accordance with original scores

**2) Data mining**

The package chosen to build a draft decision tree was the tree package. After building the first tree, I used cross-validation method for tree-pruning to avoid overfitting issues. Cross-validation is a method that randomly partitions the dataset into training and testing dataset; then it iteratively tests the training sample with testing sample for a certain amount of times that user has assigned to the method. After cross-validation, I created another set of trees that were pruned, according to the deviance value that I have discovered in cross-validation.

**3) Model Evaluation**

Lastly, I have evaluated the models I have built with the prediction function in R. To perform this, I have created a new variable and assigned the value that I have calculated with the predict function in R. Then I used the *confusionMatrix* function which I have retrieved from the *caret* library. *ConfusionMatrix* function offers one with confusion matrix and overall statistics to help one evaluate the model one has made earlier.

**Experimental Setup and Results**

Total two sets of trees were obtained using the cross-validation method described above. Although I had to choose the tree with the least deviance according to the cross-validation method, I decided to compare the accuracy of the several trees and finalize one tree for each subject, since the tree with the least deviance only showed the results with only two grading scales out of five. While comparing the accuracy of the several trees, I have found that there’s a certain rule among the accuracy; that accuracies increase following the size of the tree and if it reaches the highest point, it decreases. According to the rule I have found, I have finalized one tree for each subject; the trees are as follows.

**1) Decision tree for mathematics**

1) root 280 849.400 Fair ( 0.06429 0.17857 0.28214 0.19643 0.27857 )

2) failures < 1.5 256 774.700 Fair ( 0.07031 0.15234 0.30859 0.20703 0.26172 )

4) schoolsup: no 222 677.200 Fair ( 0.08108 0.14414 0.30631 0.23423 0.23423 )

8) absences < 0.5 61 189.800 Fail ( 0.11475 0.29508 0.19672 0.24590 0.14754 ) \*

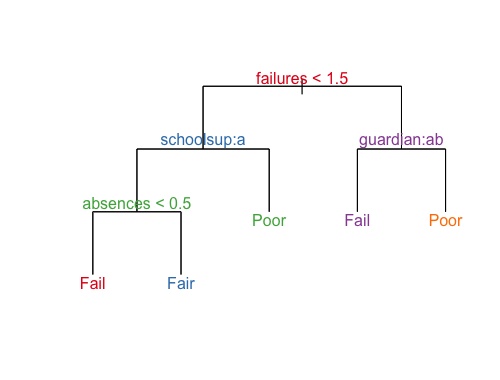
9) absences > 0.5 161 468.100 Fair ( 0.06832 0.08696 0.34783 0.22981 0.26708 ) \*

5) schoolsup: yes 34 78.550 Poor ( 0.00000 0.20588 0.32353 0.02941 0.44118 ) \*

3) failures > 1.5 24 44.270 Fail ( 0.00000 0.45833 0.00000 0.08333 0.45833 )

6) guardian: father,mother 16 19.870 Fail ( 0.00000 0.68750 0.00000 0.00000 0.31250 ) \*

7) guardian: other 8 8.997 Poor ( 0.00000 0.00000 0.00000 0.25000 0.75000 ) \*



[1] failures: number of past class failures (numeric: n if 1<=n<3, else 4)

[2] schoolsup: extra educational support (binary: yes or no)

[3] absences: number of school absences (numeric: from 0 to 93)

[4] guardian: student's guardian (nominal: "mother", "father" or "other")

**2) Decision tree for Portuguese language**

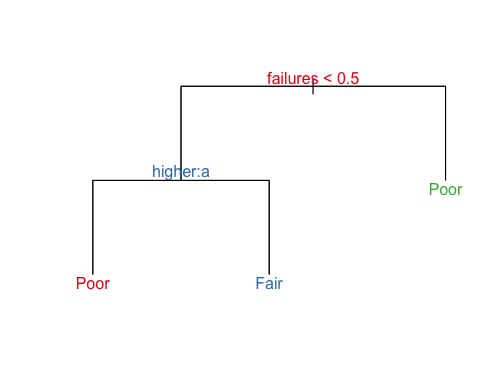
1) root 460 1238.00 Fair ( 0.04130 0.05870 0.40217 0.19783 0.30000 )

2) failures < 0.5 390 997.80 Fair ( 0.04872 0.02564 0.45641 0.22821 0.24103 )

4) higher: no 29 47.38 Poor ( 0.00000 0.10345 0.20690 0.00000 0.68966 ) \*

5) higher: yes 361 905.90 Fair ( 0.05263 0.01939 0.47645 0.24654 0.20499 ) \*

3) failures > 0.5 70 135.40 Poor ( 0.00000 0.24286 0.10000 0.02857 0.62857 ) \*



[1] failures: number of past class failures (numeric: n if 1<=n<3, else 4)

[2] higher: wants to take higher education (binary: yes or no)

**1) Confusion Matrix and Statistics for Mathematics**

Prediction Excellent Fail Fair Good Poor

Excellent 0 0 0 0 0

Fail 2 17 7 7 6

Fair 2 4 25 8 22

Good 0 0 0 0 0

Poor 0 3 2 1 9

Overall Statistics

Accuracy : 0.4435

95% CI : (0.3509, 0.5391)

No Information Rate : 0.3217

P-Value [Acc > NIR] : 0.004186

Kappa : 0.2381

Mcnemar's Test P-Value : NA

**2) Confusion Matrix and Statistics for Portuguese language**

Prediction Excellent Fail Fair Good Poor

Excellent 0 0 0 0 0

Fail 0 0 0 0 0

Fair 9 5 70 40 28

Good 0 0 0 0 0

Poor 0 9 7 0 21

Overall Statistics

Accuracy : 0.4815

95% CI : (0.4084, 0.5552)

No Information Rate : 0.4074

P-Value [Acc > NIR] : 0.02342

Kappa : 0.1658

Mcnemar's Test P-Value : NA

**Discussion**

As shown in the trees above, the three research questions that has been stated earlier were answered.

**RQ1) Among various attributes from one’s sex to parents’ jobs and relationship status, which factor does affect the student performance the most?**

The attribute *number of past class failures* resulted as the root node of the decision tree, in other words, it works as a top determinant in predicting how one would do in the class. However, the following attributes that determine student performance differed depending on subjects. In case of mathematics, extra educational support and who the student’s guardian is mattered following the number of the past failures. The number of absences mattered affected one’s performance too; it was notable that how this attribute affected student performance was the opposite from the common notion that students who go to the school every day are diligent so that they will get good grades. According to the decision tree, if one failed in the class less than 1.5 times and hadn’t had any extra educational support, and if they didn’t miss a single class from school, they were more likely to fail in mathematics.

Meanwhile, for the subject Portuguese language, student’s willingness for higher education followed the number of their past class failures. If one hasn’t gone through any class failures and is willing to look forward to higher education, they were likely to get fair grades, which is around 55%~70% out of 100%.

**RQ2: Do the factors that affect student performance differ depending on the subjects?**

As analyzed above, the overall factors that affect one’s performance differed by the school subjects; however, it is notable that both trees have the same root node. Although it is difficult to generalize to other subjects since it is only an analysis between mathematics and Portuguese language, it is a common sense that the student who didn’t have class failures would affect one’s future school performance; therefore, it would be reasonable to assume that the attribute *number of one’s past class failures* is generalizable.

**RQ3: Are the final trees generalizable?**

According to the confusion matrix and statistics retrieved from the *confusionMatrix* function, both trees are neither generalizable nor usable, even for the dataset the model had its base upon. The accuracies of both decision trees do not exceed 50%, and the accuracy of the model with training dataset has at most 12% difference compared to the model without training dataset. In conclusion, the difficulty lies in the generalization of the tree models built for the research.

**Limitations and future suggestions**

Despite the meaningful attempt to detect the general factor that would affect student performance, the research had some limitations. First, there wasn’t any specific performance level set upon the student performance that I had to set up the performance level arbitrarily. Second, the accuracy of the trees was too low to apply to not only upon all students or subjects but also upon the dataset where the models had their bases on. There should be further research conducted to solve this problem in order to build a better model; the reasons may be various, from choices of the attributes to the pruning methods. Third, as the data was collected among two Portuguese schools, it is expected that the trees built in the research will have difficulties to be generalized. Further data collection is suggested in terms of regions and school subjects, in order to create a model that would have a higher possibility to be applied to more general students. Lastly, none of the models have presented the way to seek students with either ‘Excellent’ or ‘Good’ marks, whereas the models had the best accuracy on predicting the students with ‘Fair’ marks. The reason for the skewness of the results may be that there was only one data mining method used; for further research, the use of association rule learning method or clustering method should be strongly encouraged to seek better rules or relationship among the datasets.

**Summary**

To sum up, the research aimed to seek the factor that affects the student performance. To fulfill the objective of the research, the decision tree method was used among the dataset with various factors regarding students’ demographic, social, and school-related features. The decision trees have shown that there are certain factors that affect one’s performance at school depending on individual subjects; however, it was difficult to assume that there is a common factor that can be applied to the whole student performance. On the other hand, the fact that both trees had the same root node should be highlighted, opening the opportunities for the future research. On the whole, although there were some limitations discovered, it must be said that the research opened up the possibilities for future research in student performance.

**References**

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.