**PREDICTION OF STUDENT**

**PERFORMANCE**

**MAY 10,2017**

**AARTHI SHUNMUGAM**

**STEVENS INSTITUTE OF TECHNOLOGY**

**TABLE OF CONTENTS**

**MOTIVATION OF THE PROJECT……………………………………………………………………………………………3**

**PROJECT IDEA…………………………………………………………………………………………………………………….3**

**DATASET…………………………………………………………………………………………………………………………….3**

**SOURCE………………………………………………………………………………………………………………….3**

**ATTRIBUTES…………………………………………………………………………………………………………….4**

**EXPLORATORY DATA ANALYSIS………………………………………………………………………………………….5**

**DATA CLAENING…………………………………………………………………………………………………….5**

**DATA PREPARATION……………………………………………………………………………………………….5**

**METHODOLOGY OR APPROACH USED TO FIND THE BEST PREDICTOR……………………………….6**

**BOXPLOT……………………………………………………………………………………………………………….7**

**LINEAR MODEL - MULTIPLE REGRESSION……………………………………………………………….8**

**REGRESSION TREE ANALYSIS………………………………………………………………………………….9**

**PREDICTIVE PERFORMANCE ANALYSIS………………………………………………………………………………10**

**RANDOM FOREST…………………………………………………………………………………………………………….10**

**CROSS VALIDATION TO IDENTIFY THE BEST PREDICTING MODEL………………………………………11**

**KEY FEATURES FOR PREDICTION………………………………………………………………………………………12**

**PARTIAL DEPENDENCE OF THE FEATURES…………………………………………………………………………13**

**RESULTS AND FINDINGS………………………………………………………………………………………………….14**

**REFERENCES…………………………………………………………………………………………………………………....14**

**MOTIVATION OF THE PROJECT:**

The motivation of the project was inspired from the paper work done by research students in University of Minho on the Title “**USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE**”. The goal is to analyze the 30 attributes from the student performance dataset and to find the factors or attributes, which is the best predictor of student grades.

**PROPOSED IDEA:**

The aim of this project is to identify the features that are important predictors of Student Grades, particularly interested in examining whether Alcohol Consumption has any predictive power over student grades. The prediction can be done by estimating Data Mining Models such as Linear Model, Regression Tree and Random Forest.

**DATASET:**

The data were obtained from a survey of students in secondary education of two Portuguese schools and it was collected by using school reports and questionnaires. It contains a lot of interesting social, gender and study information about students. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).

**SOURCE:**

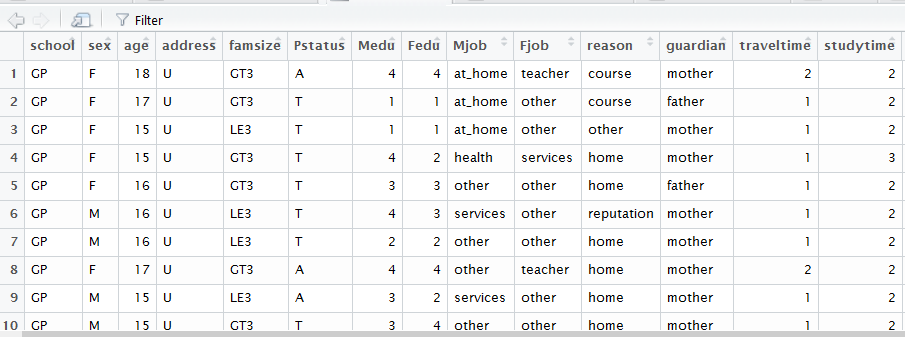
<https://archive.ics.uci.edu/ml/datasets/Student+Performance>

The dataset was taken from UCI Machine Learning Repository from the year 2014.

**ATTRIBUTES:**

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:

* school, sex , student's age, address, family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
* Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
* Medu - mother's education, Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
* Mjob - mother's job, Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')



**Fig 1: DATASET TABLE**

* Dalc - workday alcohol consumption, Walc - weekend alcohol consumption.
* reason to choose this school, guardian
* Travel time, study time, weekly study time, failures, extra-curricular activities
* higher - wants to take higher education (binary: yes or no)
* paid - extra paid classes within the course subject, internet - Internet access at home.
* schoolsup - extra educational support, famsup - family educational support.
* free time, goout, health, absences, nursery, romantic, famrel - quality of family relationships.

These grades are related with the course subject, Math or Portuguese:

* G1 - first period grade (numeric: from 0 to 20)
* G2 - second period grade (numeric: from 0 to 20)
* G3 - final grade (numeric: from 0 to 20, output target)

**EXPLORATORY DATA ANALYSIS:**

The following steps are followed to perform the process of Data Cleaning and Data Preparation.

**DATA PREPARATION:**

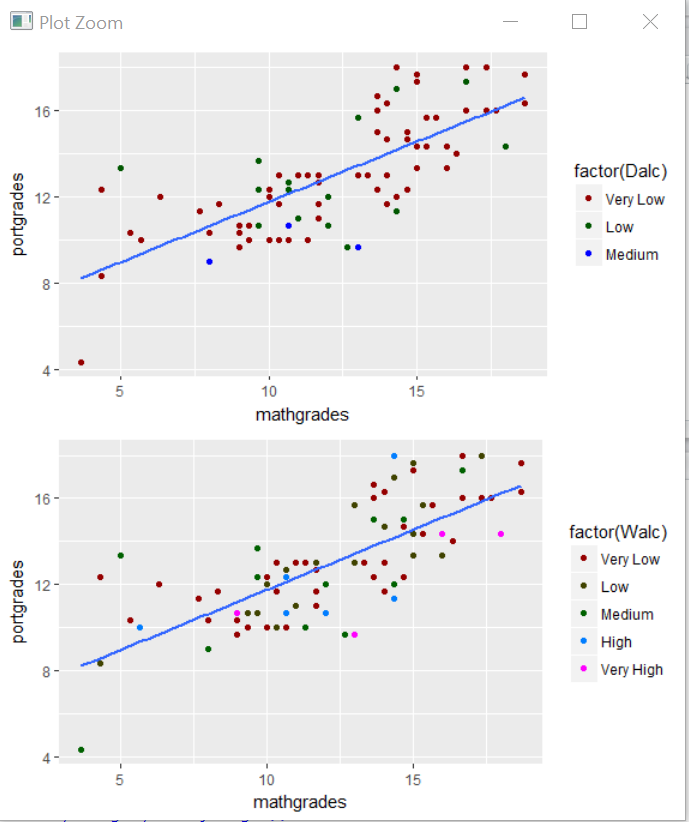
* First step is to conduct Exploratory Data Analysis to find whether the two csv files (Math and Portuguese files) can be combined. This can be done by using “merge” function.
* Average of Math and Portuguese grades are calculated and added as a column.

**DATA CLEANING:**

* Some of the attributes like ‘paid’ which is more related to course and not to specific to students are eliminated. Also, the attributes that are repeated are eliminated from the data.
* As a result, there are 85 students who belong to both the tables.

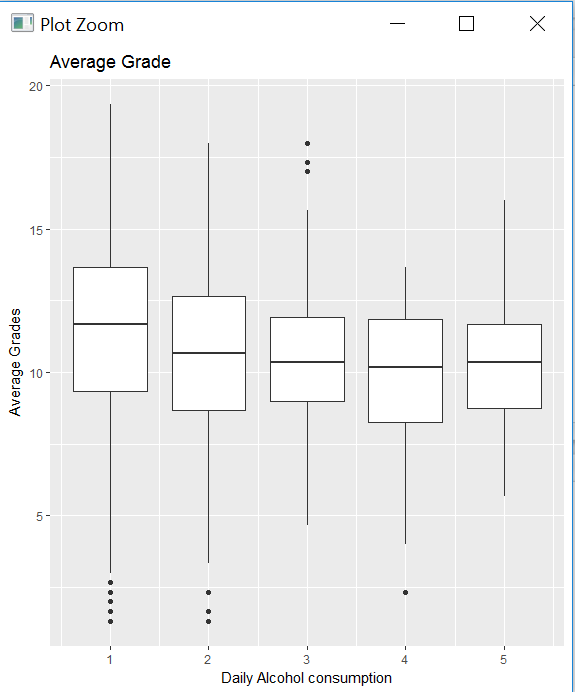
**METHODOLOGIES OR APPROACH USED TO FIND THE BEST PREDICTOR:**

Alcohol consumption on weekdays (Dalc) and weekend (Walc) are considered as factors and using “ggplot2” ad “grid” package, a scattered plot is drawn between Math grades and Portuguese grades as x and y axes respectively. The values from 1 to 5 is mapped to “Very Low”, “Low”, “Medium”, “High”, “Very High”.



**Fig 2: Scatter plot indicating the levels of Alcohol consumption for Portuguese Vs Math Grades**

**BOXPLOT:**

****

**FIG 3: BOX PLOT – AVERAGE GRADES Vs DAILY ALCOHOL CONSUMPTION.**

* The figure shows the boxplot of average subject grades grouped by the levels of daily alcohol consumption.
* The median average grade is visually higher among those students who had very low levels of daily alcohol consumption.

The median average grade is visually higher among those students who had very low levels of daily alcohol consumption. However, the median grade of the students with medium, high, and very high levels of daily alcohol consumption doesn't seem to be very different.

To predict the average grades using all other variables, the following approach is done.

🡪run a multiple linear regression

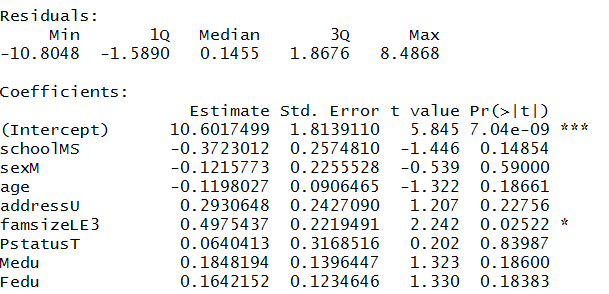
🡪build a regression tree of average grades on all other variables.

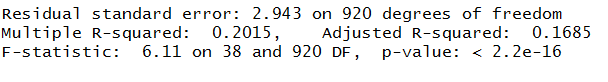
Variable "failures” is closely related to my target variable, avggrades. Since past failures and avggrades represent the same general student aptitude (thus it is rather a target rather than a feature), the variable "failures" is removed from the dataset.

**LINEAR MODEL:**

**MULTIPLE REGRESSION:**

While running the multiple regression model, the following results are obtained.



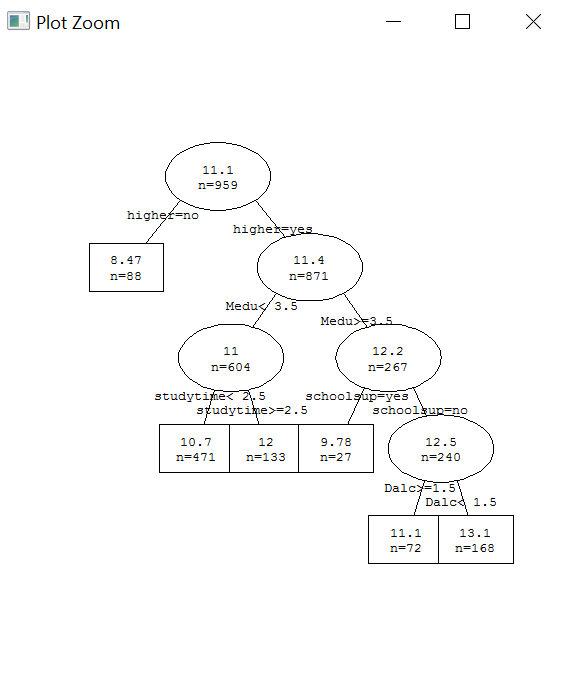


**Fig 4: MULTIPLE REGRESSION ANALYSIS RESULT.**

* Adjusted R-squared in the above regression is only 0.17, which is quite low. It implies that only 17% of the variation in the average grades is explained by the variation in everything else. The variables that have statistically significant impact on average grade are studytime, schoolsup, paid, and higher.

**Regression Tree Analysis:**

The following is the tree plot obtained from running the Regression Tree Analysis.

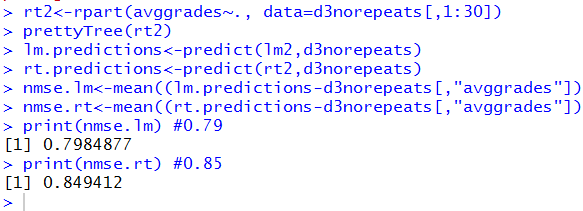


**Fig 5: Regression tree**

* The overwhelming majority of surveyed students would like to pursue higher education and their average grade (11.4/20) is significantly higher than the average grade of those who don't (8.47/20). Done using “rpart”.
* Mother's education is another important feature. Students whose mothers had at least secondary education had significantly higher grade (12.2) than the students whose mothers do not (their average grade was 11).

**Predictive performance analysis:**

To analyze the predictive performance of each model, Normalized Mean Square Error is calculated for both Linear Model and the Regression Tree. The following result is obtained.

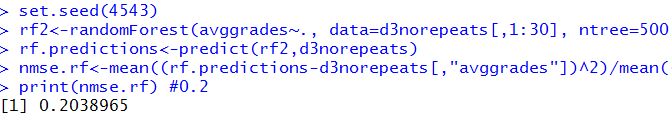


**Fig 6: Normal Mean Square Error for Linear Model and Regression Tree**

* By analyzing the predictive performance of the linear and the regression tree models, it is clear linear model performs better than the regression tree.
* This is because the NMSE (normalized mean square error of linear is 0.79 which is less than the regression tree which is 0.84).

**Random Forest:**

The performance of Linear Model and Regression Tree is not that satisfactory. Hence the predictive performance of Random Forest Model is analyzed. The Normal Mean Square Error of the Random Forest is as follows.

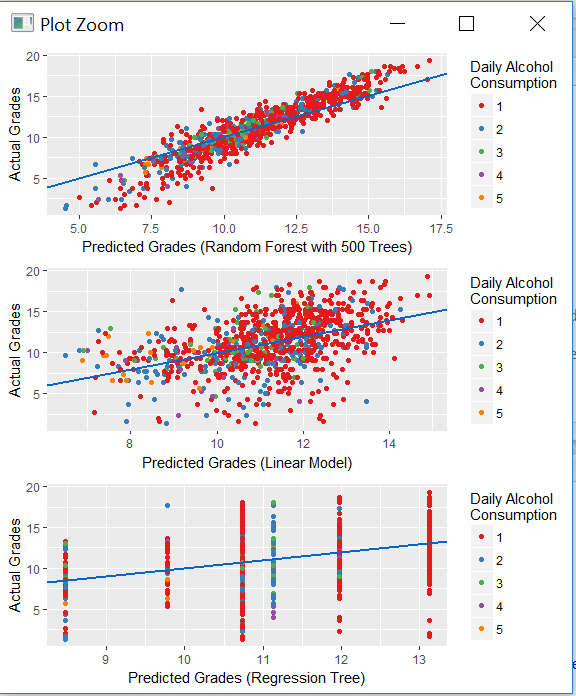


**Fig 7: Normal Mean Square Error for Random Forest Model**

NMSE of the random forest implementation is 0.2 and it is much, much lower than that of the linear and regression tree models**.**

**Cross Validation to identify the best predicting model:**

In the below graphs, horizontal axes represent predicted grades while the vertical axes represent true grades.

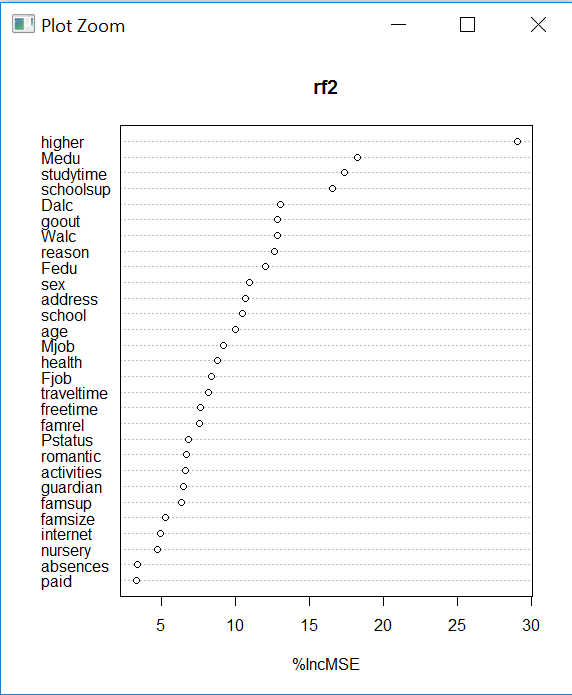
****

**Fig 8: Error Scatter Plots for Linear, Regression and Random Forest model.**

* Even though, random forest seems to systematically under predict the grades of low grade earners and overpredict the grades of high grade earners, overall, random forest seems to be a much better predictor of average grades than either the linear regression or regression tree model. If the model is accurate in predicting actual grades then predicted grades must be equal to actual grades and thus the scatter points should line up along the 45degree (blue) line.
* The NMSEs and error plots indicate, neither of the two models (linear and regression tree) seems to do a decent job in predicting student average grades.

**Key feature for prediction:**

Using the Random Forest Model, the important feature which is a good predictor for student grades is identified. The following is the plot, which displays the percentage of mean square error of all the attributes.

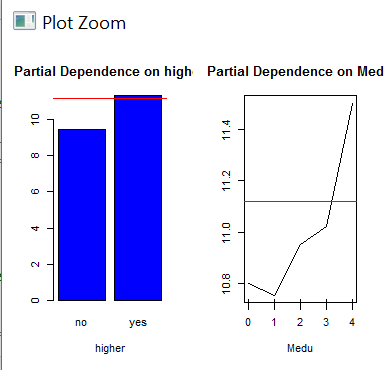


**Fig 9: Percentage of Mean Square Error for all the attributes using Random Forest Model**

* These results imply that both weekday and weekend alcohol consumption are important predictors of student average grades. Removing either of these two variables will increase the MSE of predictions by between 10-20%.
* Some features that would be conventionally thought as important did not end up in the top ten list (variables such as Pstatus, famsupport, famrel, & absences are among those) while some other variables (such as higher and Medu) turned out to be very important.

**Partial Dependence of the features:**

The following plot shows the partial dependence of the two factors the students who wants to pursue higher education and the students whose mother are educated.

****

**Fig 10: Partial Dependence plot of the features “high” and “medu”**

The above plot shows a partial dependence plot for each feature in the dataset (ranked by importance). Partial dependence plots give a graphical depiction of the marginal effect of a feature on the response. In this case, partial dependencies are produced by the best performing Random Forest model with 500 trees. As the plots indicate, shifting from "very low" levels to just "low" levels of alcohol consumption on weekdays, will reduce the expected average grade from 11.22 to 10.88. Similarly, moving from "very low" levels to just "low" levels of alcohol consumption on weekends will reduce the average predicted grade from 11.19 to 11.17.

**Results and Findings:**

The top 10 most important variables that impact student average grades are:

* **higher**- wants to take higher education (binary: yes or no)
* **Medu** - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 - secondary education or 4 – higher education)
* **studytime** - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
* **schoolsup** - extra educational support (binary: yes or no)
* **Dalc** - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
* **goout** - going out with friends (numeric: from 1 - very low to 5 - very high)
* **Walc** - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
* **reason** - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
* **Fedu** - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
* **sex** - student's sex (binary: 'F' - female or 'M' - male)

Through the above implications, results and findings the features which are more important in predicting the student grades can be identified. By analyzing the above plots, willingness to pursue higher education increases average predicted grade from 9.42 to 11.25. Hence, it is observed that the students who are interested in pursuing higher education are scoring good grades than others. Also, increase in mother's education from none to more than secondary education increases predicted average grade from 10.8 to 11.5. Thus, the students whose mothers are well educated able to obtain better grades than other students in the class.

**References:**

“USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE”- Paulo Cortez and Alice Silva -Dep. Information Systems/Algoritmi -R&D Centre -University of Minho.

Kotsiantis S.; Pierrakeas C.; and Pintelas P. “Predicting Students’ Performance in Distance Learning Using Machine Learning Techniques”. Applied Artifi- cial Intelligence (AAI), 18, no. 5, 411–426.

Luan J., “Data Mining and Its Applications in Higher Education. New Directions for Institutional Research”, 113, 17–36. Ma Y.;

Liu B.; Wong C.; Yu P.; and Lee S. “Targeting the right students using data mining”. In Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 457–464.