STUDENT PERFORMANCE PREDICTION

Project Summary

Madhu Shri Rajagopalan

Pallava Arasu Pari

Deepti Gupta

ANLY 500-90- O-2018/Summer

Harrisburg University of Science and Technology

I. Introduction

This is a recent real-world data collected from school reports and questionnaires which shows that in Europe, Portugal has high student failure rate. This failure rate is due to lack of student's success in two core classes – Mathematics and Portuguese. Our motivation to propose this project is Education. Education is one of the important building blocks of one's career. Learning appropriate moral values and proper nurturing at a tender age aids educational performance and help a student build his career. Our aim is to develop models to predict student's performance based on estimating the proportion of grades scored from their psychographic and lifestyle attributes.

II. Research Question:

Given the psychographic and lifestyle attributes of a students, can we predict their grades?

III. Dataset and Variables:

The original data used for this study is loaded from "Student Alcohol Consumption". Because it contained a detailed information on the lifestyle attributes of the students along with the grades they scored, we tried predicting grades in our models.

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets are merged into a single dataset and here is the description and class of the attributes:

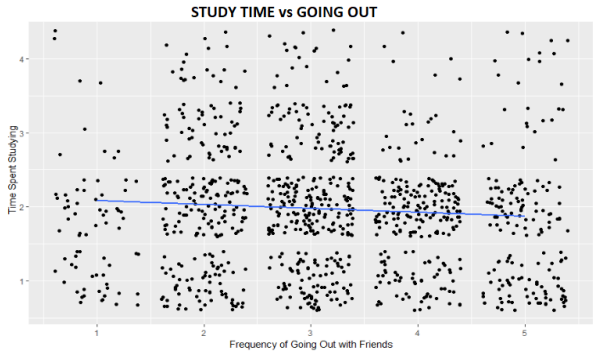
S.No.	Attribute	Description	Type
1	school	student's school ('GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)	binary
2	sex	student's sex ('F' - female or 'M' - male)	binary
3	age	student's age (from 15 to 22)	numeric
4	address	student's home address type ('U' - urban or 'R' - rural)	binary
5	famsize	family size ('LE3' - less or equal to 3 or 'GT3' - greater than 3)	binary
6	Pstatus	parent's cohabitation status ('T' - living together or 'A' - apart)	binary
7	Medu	mother's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)	numeric
8	Fedu	father's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)	numeric
9	Mjob	mother's job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')	nominal
10	Fjob	father's job ('teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')	nominal
11	reason	reason to choose this school (close to 'home', school 'reputation', 'course' preference or 'other')	nominal
12	guardian	student's guardian ('mother', 'father' or 'other')	nominal
13	traveltime	home to school travel time (1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)	numeric
14	studytime	weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)	numeric

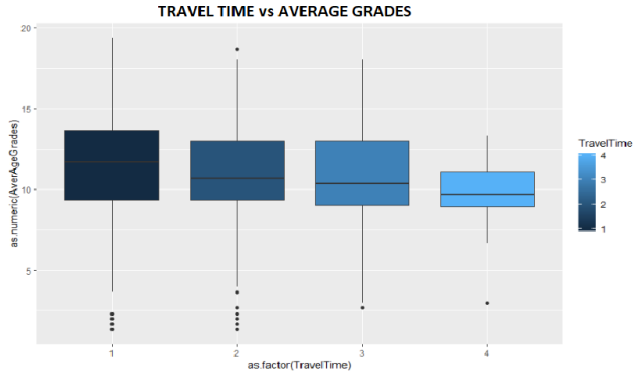
15	failures	number of past class failures (n if 1<=n<3, else 4)	numeric
16	schoolsup	extra educational support (yes or no)	binary
17	famsup	family educational support (yes or no)	binary
18	paid	extra paid classes within the course subject (Math or Portuguese) (yes or no)	binary
19	activities	extra-curricular activities (yes or no)	binary
20	nursery	attended nursery school (yes or no)	binary
21	higher	wants to take higher education (yes or no)	binary
22	internet	Internet access at home (yes or no)	binary
23	romantic	with a romantic relationship (yes or no)	binary
24	famrel	quality of family relationships (from 1 - very bad to 5 - excellent)	numeric
25	freetime	free time after school (from 1 - very low to 5 - very high)	numeric
26	goout	going out with friends (from 1 - very low to 5 - very high)	numeric
27	Dalc	workday alcohol consumption (from 1 - very low to 5 - very high)	numeric
28	Walc	weekend alcohol consumption (from 1 - very low to 5 - very high)	numeric
29	health	current health status (from 1 - very bad to 5 - very good)	numeric
30	absences	number of school absences (from 0 to 93)	numeric
31	G1	first period grade (from 0 to 20)	numeric
32	G2	second period grade (from 0 to 20)	numeric
33	G3	final grade (from 0 to 20, output target)	numeric

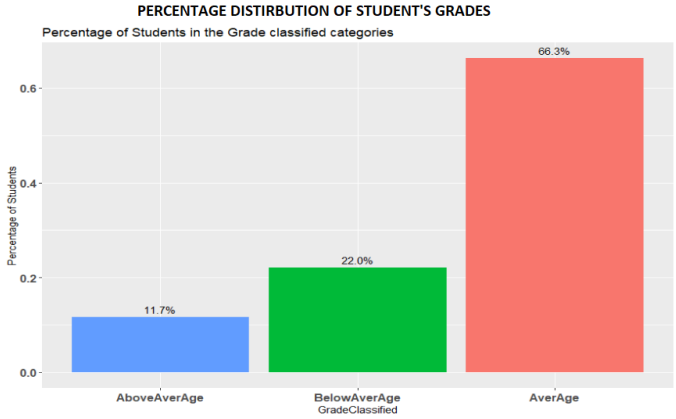
IV. Data Exploration

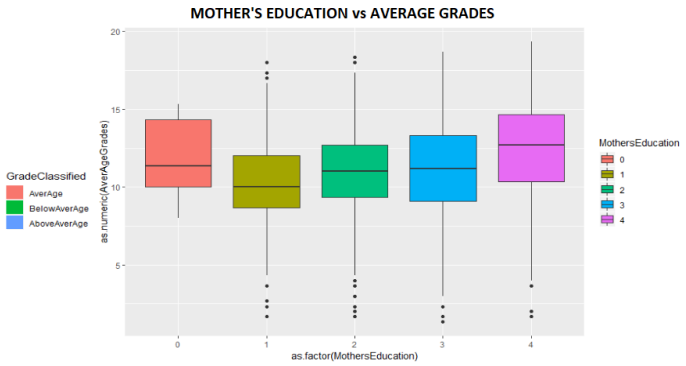
In this part, we sliced and diced, factorized the data to make it fit for a better data analysis. We tried studying patterns in the data by plotting several graphs and noted our observations to implement our models accordingly.

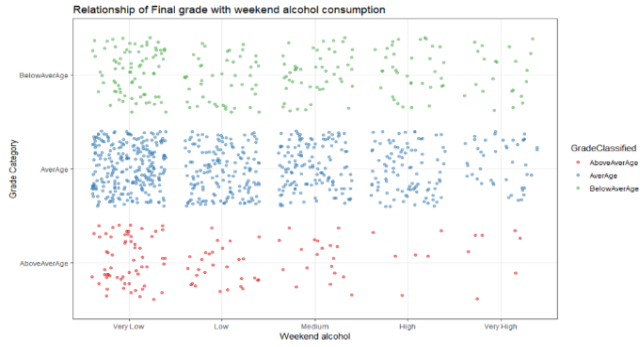
Following are few of the plots we found considerable.

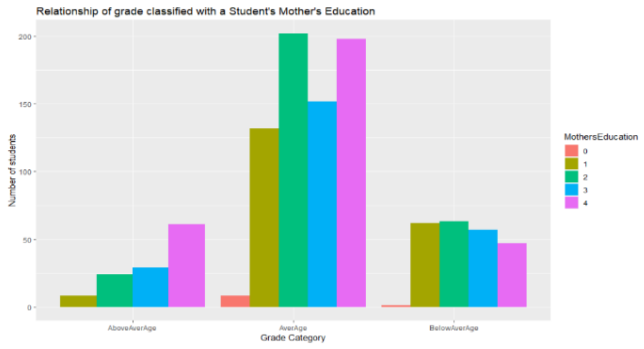

STUDY TIME vs GOING OUT
Time Spent Studying
Frequency of Going Out with Friends

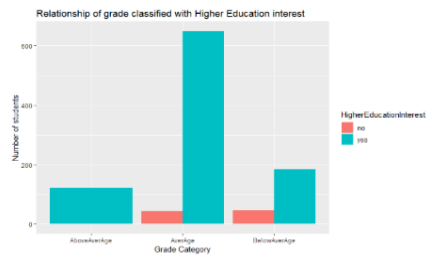

TRAVEL TIME vs AVERAGE GRADES
as.numeric(AverAgeGrades)
as.factor(TravelTime)
TravelTime

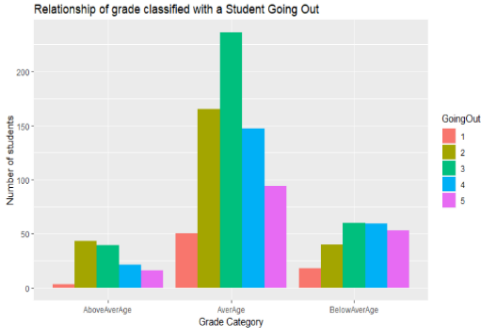

PERCENTAGE DISTIRBUTION OF STUDENT'S GRADES
Percentage of Students in the Grade classified categories
Percentage of Students
66.3%
22.0%
11.7%
AboveAverAge
BelowAverAge
AverAge
GradeClassified

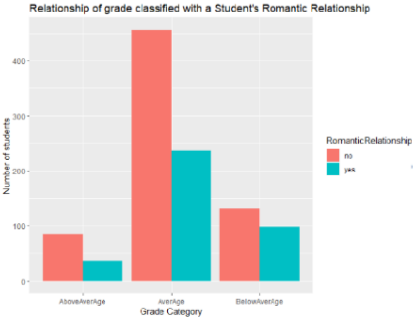

MOTHER'S EDUCATION vs AVERAGE GRADES
as.numeric(AverAgeGrades)
as.factor(MothersEducation)
GradeClassified
AverAge
BelowAverAge
AboveAverAge
MothersEducation


Relationship of Final grade with weekend alcohol consumption
Grade Category
Weekend alcohol
Very Low
Low
Medium
High
Very High
GradeClassified


Relationship of grade classified with a Student's Mother's Education
Number of students
Grade Category
MothersEducation


Relationship of grade classified with Higher Education interest
Number of students
Grade Category
HigherEducationInterest


Relationship of grade classified with a Student Going Out
Number of students
Grade Category
GoingOut


Relationship of grade classified with a Student's Romantic Relationship
Number of students
Grade Category
RomanticRelationship

During the data exploration process, the students are classified into Very Low, Low, Medium, High and Very High and the graphs are plotted. The first graph plotted is for number of students drinking on the weekday or weekend. We noted that the plot is positively skewed and the students who drank scored very low grades and only a few who did not consume alcohol scored very high grades.

The next influencer used is the failures affecting grades and we noted that as the number of failures increase, the grades decrease.

Then the mother's education is taken into account and based on the plot, we conclude that grades are better when the mother is educated.

The fourth influencer accounted is the travel time and with increasing travel time, the grades get bad.

Next factor looked at is how going out with friends would influence the grades and it is noted that more the students are out, lesser is the time spent in studying.

Besides all these major influencers, other random variables are also looked at such as the parent's cohabitation status and it proves that each variable available in the data has a little influence on the grades.

V. Methods:

We implemented 3 models – multiple linear regression, logistic regression and k-NN and came up with different results for all of them. The following are the details on our models and their predictions.

i. Multiple Linear Regression:

We started with multiple linear regression which was implemented in stepwise process. The initial model contained all the attributes against Average Grades and we got an $R^2 = 0.2758$. Based on the p-values greatest to lowest, we tried removing variables one by one from model 1 to 8 where finally we achieved a model where the p-values for all the attributes was below the significant value with adjusted-$R^2 = 0.2699$. In order to achieve better accuracy, we tried multinomial logistic regression as our next step.

ii. Multinomial Logistic Regression:

During the process of data analysis, the data is plotted and from all the variables identified, a new column is added titled as "grade classified" in which the final grades are classified into three factored levels namely above average, average and below average and these are classified based on grades scored below 10 and above 15. Using ggplot, we observed that 11.7% of the total students are above average, 22% at average and 66.3% below average. Several other factors that influence these grades are the romantic relationship, weekend alcohol, mother’s education, failures, going out, higher education, and student’s travel time.

Plotting these variables on the graph for all the grades scored, we notice that the students who answered yes to the survey for grades influenced by romantic relationship are higher at an average and lower at student's who are above average.

Relationship to the number of students who answered 1 to weekend alcohol consumption are higher at an average below average.

Students who answered mothers education being one of the key factor that influences grades are at above average which concludes that having either or both the parents educated would have an impact on the students grades.

Plotting the next factor, i.e. the students who think going out affects the grades. Most of the students selected between 2 and 3 points for this factor as having an impact.

Plotting the interest towards higher education, we notice that all the students who are at above average answered higher education to have the maximum influence on the grades of the students.

And lastly the travel time that takes students to reach schools show that the grades get lower with increasing travel time.

Modeling:

We proceed in a stepwise manner in Multinomial Logistic Regression too. In the 1st test scenario, we included all the influential variables and calculated the accuracy. And we got the data accuracy to be 92% which is incorrect and deviating. In the second test scenario, we reduced the model and kept certain attributes including Mother's education, Romantic relationship, Weekend alcohol, Going out, Higher education interest, Mother's education, Study time, Internet access and Health which gave us an accuracy of 67%. In test scenario 3, we added a few more variables to these and found the accuracy to be 68%. In test scenario 4, we included all the variables but excluded grades and found that the accuracy came to 68% just like test scenario 3 and that's where we stopped. Multinomial logistic regression gave us an accuracy of 68%. We now wanted to try k-NN model on our data.

k-Nearest Neighbor Model (KNN):

The last model we implemented is the KNN model. To prepare the data accordingly, first we needed to convert all the categorical variables into numeral type and then normalize it so that all the columns are scaled from 0-1 to make them even for analysis. We classified the grades into fail and pass classifier which will be our target variable. Fetched the sample of training and testing sets by randomly splitting the dataset in 80:20 ratio and giving an index value of 1 and 2 respectively. Applied k-NN by taking an initial k-value as 25 which is taken as a random square root value of the dataset. By randomly varying the value, we got the highest accuracy at k=21. Below is the cross-table matrix of our result-

	PredictedModel		
StudData_test$GradesClassified	Fail	Pass	Row Total
Fail	76	42	118
	0.644	0.356	0.541
	0.724	0.372	
	0.349	0.193	
Pass	29	71	100
	0.290	0.710	0.459
	0.276	0.628	
	0.133	0.326	
Column Total	105	113	218
	0.482	0.518	

It shows that when the model predicted a student fail, 76 out of 118 students were failed and when the model predicted a student pass, 71 out of 100 students were pass. Out of 218 total test records, 147 were predicted accurately giving us 70% accuracy.

k	Fail	Pass	Match	TOTAL	ACCURACY	
20	76	69	145	218	0.7005	
25	72	68	140	218	0.6763	
28	65	71	136	218	0.6570	
21	76	71	147	218	0.7053	****
22	72	69	141	218	0.6812	
19	74	69	143	218	0.6908	

Above is the analysis of the random k-values we tried, and it shows that the highest accuracy was achieved with k-value as 21.

VI. Appendices

StudentPerformanc ePrediction.Rmd

StudentPerformanc ePrediction_KnitRep

VII. Reference:

P. Cortez and A. Silva, pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7, [Online].Available: https://archive.ics.uci.edu/ml/datasets/Student+Performance