Student Grade Prediction

Prathap Reddy Vasampalli

29/06/2022

## Summary

With the open education systems globally, there is the flexibility for students to study anything of their choice as content is readily and widely available. However, the students may become complacent with this freedom. In return, it makes it difficult for teachers and the teaching staff to predict the students’ performance in advance. For this problem, this project has attempted to develop a prediction system using machine learning algorithms in R to help predict the student’s grade in advance for proper study planning. The main objective of this project is to develop an algorithmic model that will provide students and teachers with the functionality of predicting an individual student’s grades using the final grade as the target variable. The project implements a linear regression model as a predictive model for the student’s grades. With the prediction of students’ grades, it will be easier for teachers to identify students who need assistance, which will help improve the students’ performance. The project is implemented in RStudio and developed using the R programming language.

## Introduction

The student’s performance in the present educational system seems to gradually worsen daily. However, to ensure the teachers and students can improve their performance and keep track of the student’s progress, it is a good practice to predict the student’s performance in advance. Therefore, this paper’s main focus is to discuss how this can be achieved. The paper implements a linear regression model that will help in student grade prediction to help students know their performance in advance and ensure they plan accordingly and improve on the areas of their weakness.

## Literature Review

In a study by Iqbal et al. (2017), the researcher implemented Collaborative Filtering, Restricted Boltzmann Machine, and Matrix Factorization techniques to predict the student’s grades for several courses. From their analysis, it was clear to them that the Restricted Boltzmann Machines was the best technique for the prediction compared to Matrix factorization and Collaborative filtering since it is the minimum Root Mean Squared value. In another study by Nabil et al. (2021), a predictive model for predicting students’ final grades was developed. From the eleven machine learning algorithms they had implemented in their study, the authors identified Decision Tree Classifier as the most suitable for students’ grades as it had the highest accuracy of 88% compared to the other algorithms developed. In the study by Abana (2019), the authors implemented three different models to predict students’ grade performance. The authors also implemented cross-validation to validate the performance of the predictive models. Of the three algorithms, Random Forest was the best as it had an accuracy of 75%, which was higher than that of the other algorithms. However, the authors proposed that more samples must be added to improve the accuracy of the predictive model (Abana, 2019).

## Data

The data has been retrieved from <https://archive.ics.uci.edu/ml/machine-learning-databases/00320/student.zip>. It consists of:

## school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason  
## 1 GP F 18 U GT3 A 4 4 at\_home teacher course  
## 2 GP F 17 U GT3 T 1 1 at\_home other course  
## 3 GP F 15 U LE3 T 1 1 at\_home other other  
## 4 GP F 15 U GT3 T 4 2 health services home  
## 5 GP F 16 U GT3 T 3 3 other other home  
## 6 GP M 16 U LE3 T 4 3 services other reputation  
## guardian traveltime studytime failures schoolsup famsup paid activities  
## 1 mother 2 2 0 yes no no no  
## 2 father 1 2 0 no yes no no  
## 3 mother 1 2 3 yes no yes no  
## 4 mother 1 3 0 no yes yes yes  
## 5 father 1 2 0 no yes yes no  
## 6 mother 1 2 0 no yes yes yes  
## nursery higher internet romantic famrel freetime goout Dalc Walc health  
## 1 yes yes no no 4 3 4 1 1 3  
## 2 no yes yes no 5 3 3 1 1 3  
## 3 yes yes yes no 4 3 2 2 3 3  
## 4 yes yes yes yes 3 2 2 1 1 5  
## 5 yes yes no no 4 3 2 1 2 5  
## 6 yes yes yes no 5 4 2 1 2 5  
## absences G1 G2 G3  
## 1 6 5 6 6  
## 2 4 5 5 6  
## 3 10 7 8 10  
## 4 2 15 14 15  
## 5 4 6 10 10  
## 6 10 15 15 15

View the structure of the data.

str(student\_perf)

## 'data.frame': 395 obs. of 33 variables:  
## $ school : chr "GP" "GP" "GP" "GP" ...  
## $ sex : chr "F" "F" "F" "F" ...  
## $ age : int 18 17 15 15 16 16 16 17 15 15 ...  
## $ address : chr "U" "U" "U" "U" ...  
## $ famsize : chr "GT3" "GT3" "LE3" "GT3" ...  
## $ Pstatus : chr "A" "T" "T" "T" ...  
## $ Medu : int 4 1 1 4 3 4 2 4 3 3 ...  
## $ Fedu : int 4 1 1 2 3 3 2 4 2 4 ...  
## $ Mjob : chr "at\_home" "at\_home" "at\_home" "health" ...  
## $ Fjob : chr "teacher" "other" "other" "services" ...  
## $ reason : chr "course" "course" "other" "home" ...  
## $ guardian : chr "mother" "father" "mother" "mother" ...  
## $ traveltime: int 2 1 1 1 1 1 1 2 1 1 ...  
## $ studytime : int 2 2 2 3 2 2 2 2 2 2 ...  
## $ failures : int 0 0 3 0 0 0 0 0 0 0 ...  
## $ schoolsup : chr "yes" "no" "yes" "no" ...  
## $ famsup : chr "no" "yes" "no" "yes" ...  
## $ paid : chr "no" "no" "yes" "yes" ...  
## $ activities: chr "no" "no" "no" "yes" ...  
## $ nursery : chr "yes" "no" "yes" "yes" ...  
## $ higher : chr "yes" "yes" "yes" "yes" ...  
## $ internet : chr "no" "yes" "yes" "yes" ...  
## $ romantic : chr "no" "no" "no" "yes" ...  
## $ famrel : int 4 5 4 3 4 5 4 4 4 5 ...  
## $ freetime : int 3 3 3 2 3 4 4 1 2 5 ...  
## $ goout : int 4 3 2 2 2 2 4 4 2 1 ...  
## $ Dalc : int 1 1 2 1 1 1 1 1 1 1 ...  
## $ Walc : int 1 1 3 1 2 2 1 1 1 1 ...  
## $ health : int 3 3 3 5 5 5 3 1 1 5 ...  
## $ absences : int 6 4 10 2 4 10 0 6 0 0 ...  
## $ G1 : int 5 5 7 15 6 15 12 6 16 14 ...  
## $ G2 : int 6 5 8 14 10 15 12 5 18 15 ...  
## $ G3 : int 6 6 10 15 10 15 11 6 19 15 ...

Check for any missing values in the dataset. There are no missing values in the dataset.

sapply(student\_perf,function(x) sum(is.na(x)))

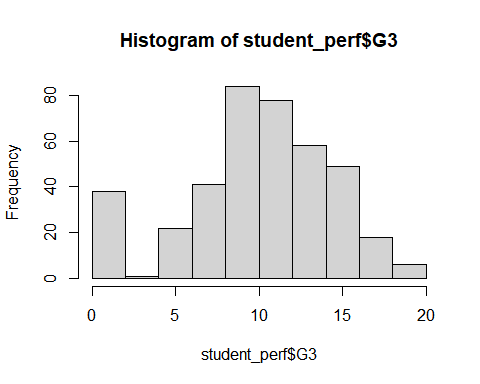
## school sex age address famsize Pstatus Medu   
## 0 0 0 0 0 0 0   
## Fedu Mjob Fjob reason guardian traveltime studytime   
## 0 0 0 0 0 0 0   
## failures schoolsup famsup paid activities nursery higher   
## 0 0 0 0 0 0 0   
## internet romantic famrel freetime goout Dalc Walc   
## 0 0 0 0 0 0 0   
## health absences G1 G2 G3   
## 0 0 0 0 0

normalize <- function(x){  
 return ( (x- min(x)) / ( max(x) - min(x)))  
}  
  
# normalizing data  
cols <- c('age','Medu','Fedu','traveltime','studytime','failures','famrel','freetime','goout','Dalc','Walc','health','absences','G1','G2')  
student\_perf[cols] <- lapply(student\_perf[cols], normalize)

## Methodology

Since the data is clean, we perform exploratory analysis and then build the linear regression model in this section. First, we need to understand the distribution of G3, Final grade, which is the target variable for this project by creating a histogram.

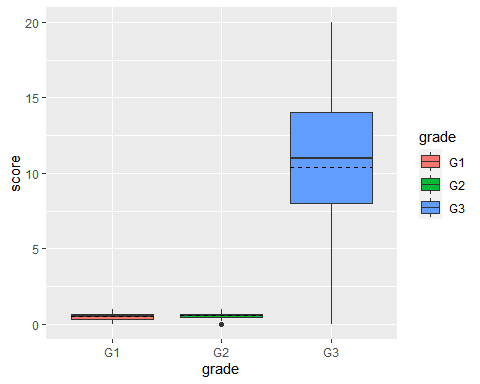
hist(student\_perf$G3)



Next, using a boxplot, we check the distribution of G1, first period exam, G2, second period exam, and G3, final exam.

Exams\_Grades <- gather(student\_perf,key = grade,value = score, G1,G2,G3)  
  
ggplot(Exams\_Grades, aes(x=grade, y=score, fill=grade)) + geom\_boxplot() +  
 stat\_summary(fun.y = mean, geom = "errorbar", aes(ymax = ..y.., ymin = ..y..),  
 width = .75, linetype = "dashed")

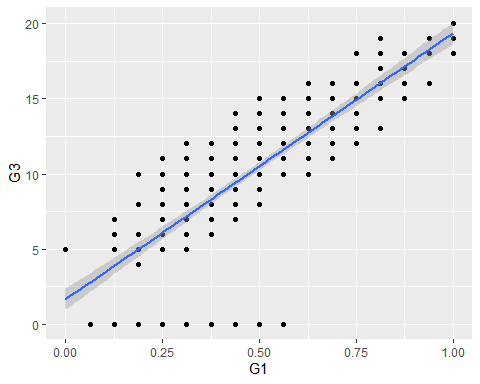
## Warning: `fun.y` is deprecated. Use `fun` instead.



Next, we visualize the relationship between the predictor variable G3 and G1 using a scatter plot.

ggplot(student\_perf,aes(x=G1,y=G3)) +  
 geom\_point() + geom\_smooth(method = 'lm')

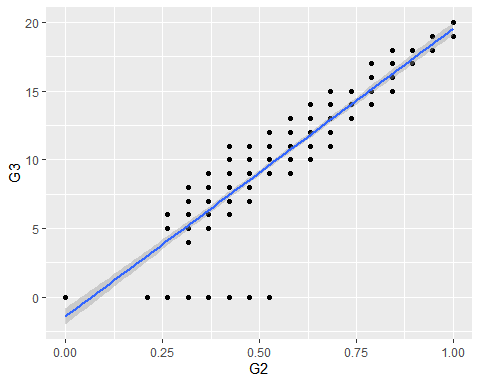
## `geom\_smooth()` using formula 'y ~ x'



Next, we visualize the relationship between the predictor variable G3 and G2 using a scatter plot.

ggplot(student\_perf,aes(x=G2,y=G3)) +  
 geom\_point() + geom\_smooth(method = 'lm')

## `geom\_smooth()` using formula 'y ~ x'



Analysis on the student performance and the aim to pursue higher education.

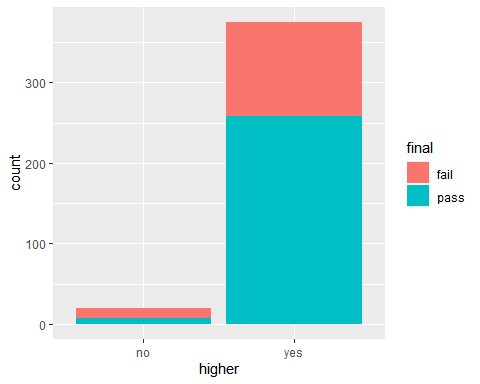
student\_perf$final <- factor(ifelse(student\_perf$G3 >= 10, 1, 0), labels = c("fail", "pass"))  
edu\_high <- dcast(unique(student\_perf), formula = higher ~ final, fun.aggregate = length)

## Using final as value column: use value.var to override.

edu\_high <- gather(edu\_high, type, count, fail,pass)  
edu\_high

## higher type count  
## 1 no fail 13  
## 2 yes fail 117  
## 3 no pass 7  
## 4 yes pass 258

ggplot(student\_perf, aes(x=higher, group=final,fill=final)) + geom\_bar()



Next, we split data to train and test in the ratio 80:20 respectively.

set.seed(123)  
  
train\_split <- createDataPartition(y = student\_perf$final, p = 0.8, list = FALSE)  
  
train <- student\_perf[train\_split,]  
test <- student\_perf[-train\_split,]  
  
  
  
# Check the dimension train and test sets  
dim(train)

## [1] 316 34

We need to find the most important predictor variables to build the model. Therefore we perform feature selection through stepwise regression. Final and G3 are the most important predictor variables for students’ grades.

options(warn = -1)  
# model with intercept only  
mod1 <- glm(final ~ 1 , data= train,family=binomial)   
  
  
# all predictor variables model  
mod2 <- glm(final ~ . , data= train,family=binomial)  
  
#step-wise algorithm  
Mod\_stepW <- step(mod1, scope = list(lower = mod1, upper = mod2), direction = "both", trace = 0, steps = 1000)   
  
# shortlisted variables  
formula(Mod\_stepW)

## final ~ G3

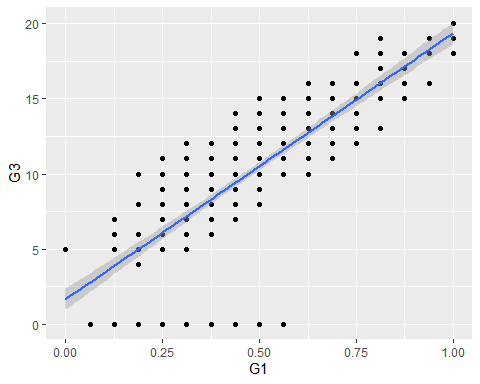
options(warn = -1)  
set.seed(1337)  
  
# 10-fold cross validation  
train\_ctrl <- trainControl(method="cv", number=10)  
  
# use glm to Train the model   
model <- train(formula(Mod\_stepW), data = train, method = "glm",trControl=train\_ctrl,family = binomial)  
  
# summary of the model  
summary(model)

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.162e-05 -2.110e-08 2.110e-08 2.110e-08 2.491e-05   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -411.21 73164.89 -0.006 0.996  
## G3 43.31 7674.40 0.006 0.995  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4.0040e+02 on 315 degrees of freedom  
## Residual deviance: 5.0054e-08 on 314 degrees of freedom  
## AIC: 4  
##   
## Number of Fisher Scoring iterations: 25

## Results

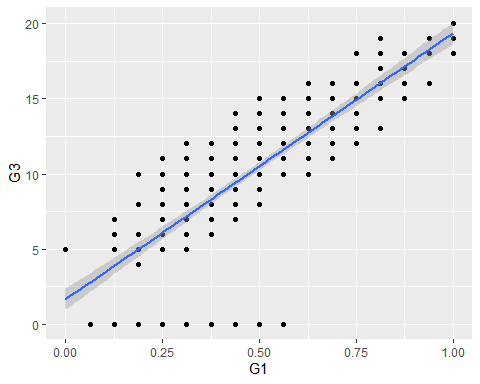
The scatter plot for the distribution of G3 and G2 illustrates that there is a strong linear relationship exists between variables G1 and G3. This is because, with an increase in the values for the variable G3, the variables for G1 also increase.

## `geom\_smooth()` using formula 'y ~ x'

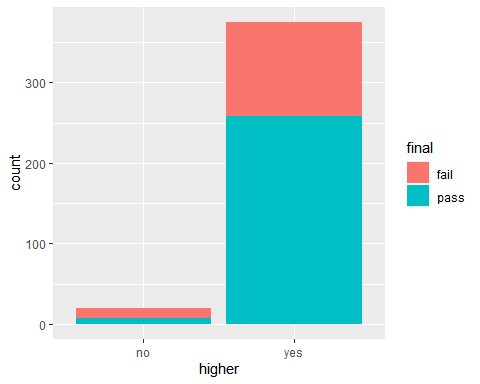


In addition, the scatter plot for the distribution of G3 and G2 illustrate that there is a strong linear relationship that exist between variables G2 and G3.This is because, with increase in the values for the variable G3 the variables for G2 also increases.

## `geom\_smooth()` using formula 'y ~ x'



The stacked barchart also illustrate that those students who plan to pursue higher education are more likely to pass unlike the students who do not have any plans of pursuing higher education.



The logistic regression model having been created, it has an accuracy of 92% of predicting the students grades performance.

## Confusion Matrix and Statistics  
##   
##   
## glm.probs fail pass  
## fail 26 0  
## pass 0 53  
##   
## Accuracy : 1   
## 95% CI : (0.9544, 1)  
## No Information Rate : 0.6709   
## P-Value [Acc > NIR] : 2.02e-14   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Sensitivity : 1.0000   
## Specificity : 1.0000   
## Pos Pred Value : 1.0000   
## Neg Pred Value : 1.0000   
## Prevalence : 0.6709   
## Detection Rate : 0.6709   
## Detection Prevalence : 0.6709   
## Balanced Accuracy : 1.0000   
##   
## 'Positive' Class : pass   
##

Lastly, final and grade are the main variables to consider in predicting grades for students.

##   
## Call: glm(formula = final ~ G3, family = binomial, data = train)  
##   
## Coefficients:  
## (Intercept) G3   
## -411.21 43.31   
##   
## Degrees of Freedom: 315 Total (i.e. Null); 314 Residual  
## Null Deviance: 400.4   
## Residual Deviance: 5.005e-08 AIC: 4

## Implications

This project only created one model that could be used to predict students’ grades performance. This might not give a real solution, and therefore for future works, and implementation of different machine learning models should be considered to identify the most suitable model for students’ grades prediction.

## Conclusion

This paper has achieved the objective of developing a logistic regression classifier to predict the students’ grades. The project involved importing data into RStudio, cleaning data by checking missing values and duplicates and normalizing the values of the columns to a similar scale.

## References

Abana, E. C. (2019). A decision tree approach for predicting student grades in Research Project using Weka. International Journal of Advanced Computer Science and Applications, 10(7).

Iqbal, Z., Qadir, J., Mian, A. N., & Kamiran, F. (2017). Machine learning based student grade prediction: A case study. arXiv preprint arXiv:1708.08744.

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students’ academic performance based on courses’ grades using deep neural networks. IEEE Access, 9, 140731-140746.