Predicting Students Final Grade in School, by Exploring Three Different Machine Learning Models.

ABSTRACT

We notice that the educational level of the Portuguese population has a tremendous improvement in this past years, specifically, in the field or courses like Mathematics and a language course like the Portuguese language. In this study, we aim to predict the students final grade score in a secondary education by using some machine learning techniques. 3 models used for this analysis are (1) The Regression model, (2) Decision Tree and (3) Random Forest. It is a known fact that students final score is often influenced by his or her past grades. However, an explanatory analysis has shown that there are also other relevant factors that contributes to students final grade. (e.g. attendencs, family size, parent education, students relationship life style e.t.c). The choice of model is because we are interested in predicting the student overall grade using some of the features in the data set. And since it's a regression problem, the aforementioned models are good for such task.

INTRODUCTION

Education as we know to be the the act or process of imparting or acquiring general knowledge, developing the powers of reasoning and judgment has been one of the key factors for achieving a long term economic progress. In this past years, the education of the Portuguese has greatly improved. In this work, we will analyze recent real-world data from two Portuguese secondary schools. ('GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira). Since the objective of study was to predict student final grade and also to identify some other key factors/features that influences students success/failure in Mathematics. From the survey that were carried out some information retrieved were demographic, social and school related features (e.g. student's age, parents education level, travel time e.t.c). The two core courses involved were Mathematics and Portuguese language course. The structure of the data is as follows: The Mathematics course data had 396 rows with 32 columns while the Portuguese language course consist of 650 rows with 32 columns. Each row defines the student while the columns are the attributes. We have both the categorical columns like school, sex, famsize, Psatus, Medu, Fedu et.c). while some continuous variables are age, G1, G2, G3, absence, Dalc, Walc e.t.c. Since the aim is to predict students final grade, my responds variable would be G3 (final grade), while the remaining features are the explanatory variables.

DATA PROCESSING

The data was tested for missing variables using the descriptive statistic and it was observed that there was no missing variable in the dataset. The categorical variables were converted to factors, using as.factor() function in r. After which data was split into training and testing (75% and 25%) respectively, this is because, after training the different models, in order to measure success we need to evaluate or trained model on a text set. This would enable us compare results among the 3 models, and also know the performances of each model on new data set.

METHODOLOGY

In analysis, the benefits of visualization cannot be over emphasized. For this reason, we carried out some visualization mainly to answer these 2 questions (1) what are the age distribution of the students in both Mathematics and the Portuguese language course. (2) Do Girls Perform Better in School? The answers can be seen in the result section (EDA). Furthermore, when building models in Machine Learning, it is advisable to start with a baseline/ or a simple model. Hence, the first model used was a linear regression with all the variables. From the result, we noticed that most of the variables were not significant, hence, we decided to reduce the number of features by using only the numerical variables, and that give rice to model 1, a residual plot as well as the normal Q-Q plot was plotted to visualize and see how the model fit the data. The second model we looked at was the Decision Tree, when the **decision trees was built** many of the branches where reflecting noise or outliers in the training data. Hence, a technique called pruning was introduced. **Tree pruning** methods address this problem as well as overfitting (a situation where the model cannot generalize well on new instances). The last model explored was the Random Forest. This method Random is also called random decision forests, It's an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the mean prediction of the individual trees. After all these 3 models were explored, we then make predictions using the set aside test set. The correlation between the predictions made and the original test set were estimated. Also, the metric use for evaluation is the RMSE, To select the best model, we compere the RMSE and the Correlation of each of the models, the model with the least RMSE and highest correlation of (predicted vs actual) is chosen to be the best model.

RESULTS

Exploratory Data Analysis

Table 1: Age distribution across the two course.

Courses / Ages	15	16	17	18	19	20	21	22
Mathematics course	82	104	98	82	24	3	1	1
Portuguese language course	112	177	179	140	32	6	2	1

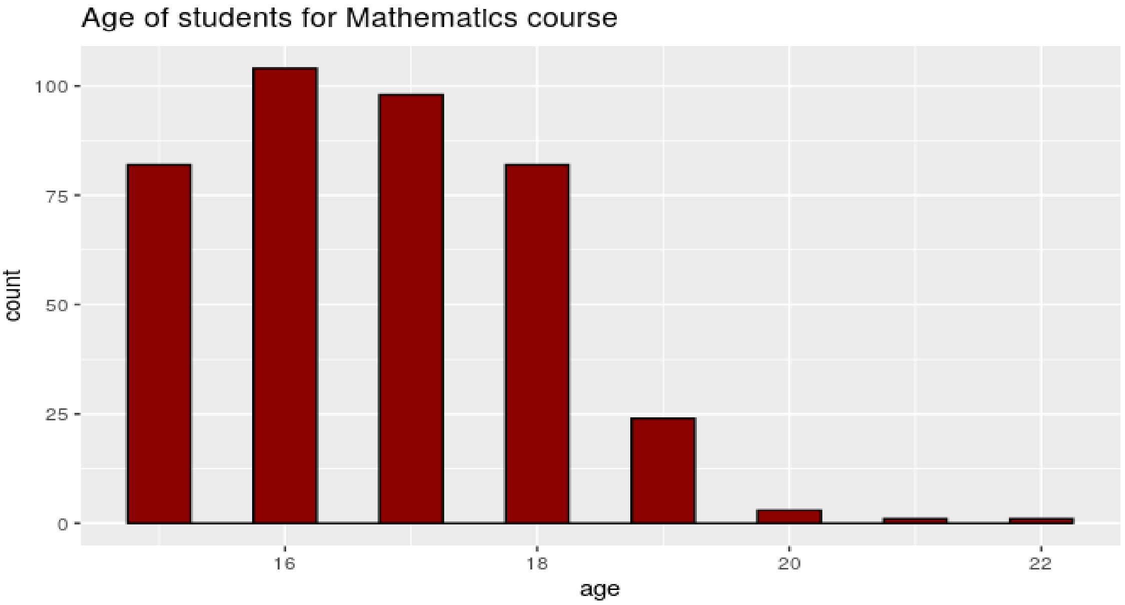


Figure 1

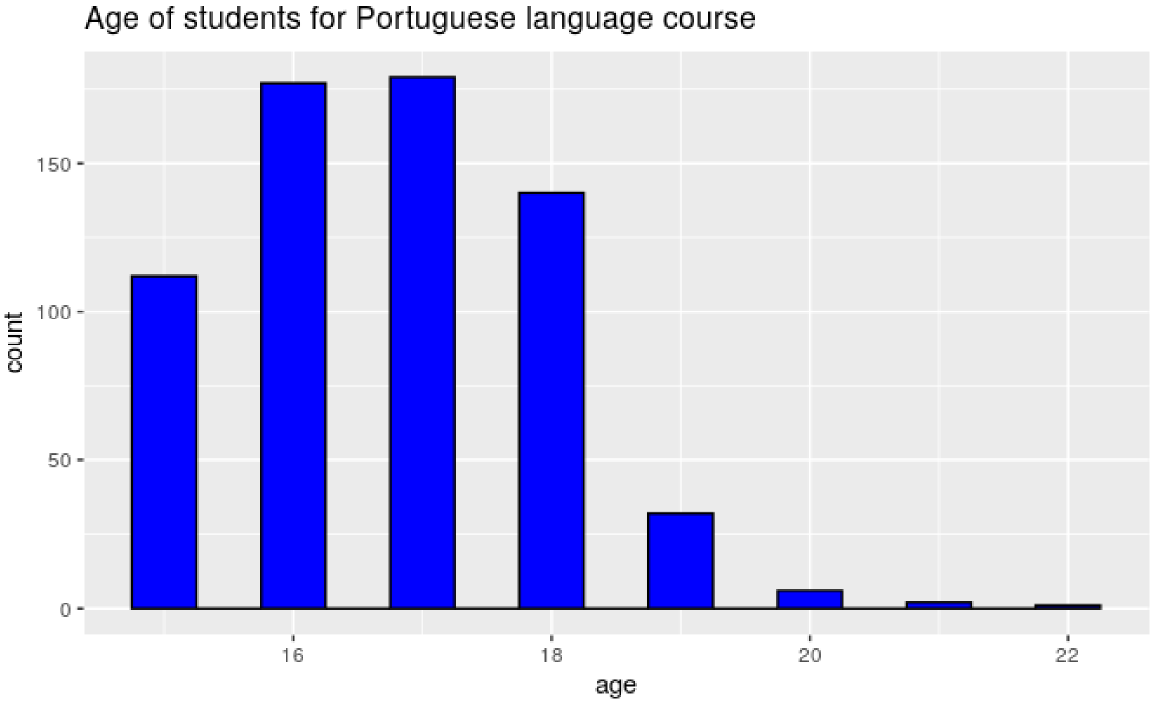


Figure 2

The table above as well as the plots shows the distribution of students age across the two courses. From the table, we can deduce that for Mathematics course, most of the students (104) are 16 years while for the Portuguese language course, most of them (179) were 17 years of age. Overall, the mean age for both course is 17 years. Since we have the same distributions for course, we can now proceed to answer the question bellow:

- We want to investigate if females performs better than males.

To do this, we'll explore gender differences in the class with respect to G3 (student final grade).

We will:

1) Estimate how many female and male student are present in the the school.

2) To examine the performance of male and female based on and age we will do the following:

a) Do male perform better than females?

b) Does age influence students performance at all?

Table 2

Courses / Ages	**Males**	**Females**
Mathematics	187	208
Portuguese language course	266	383

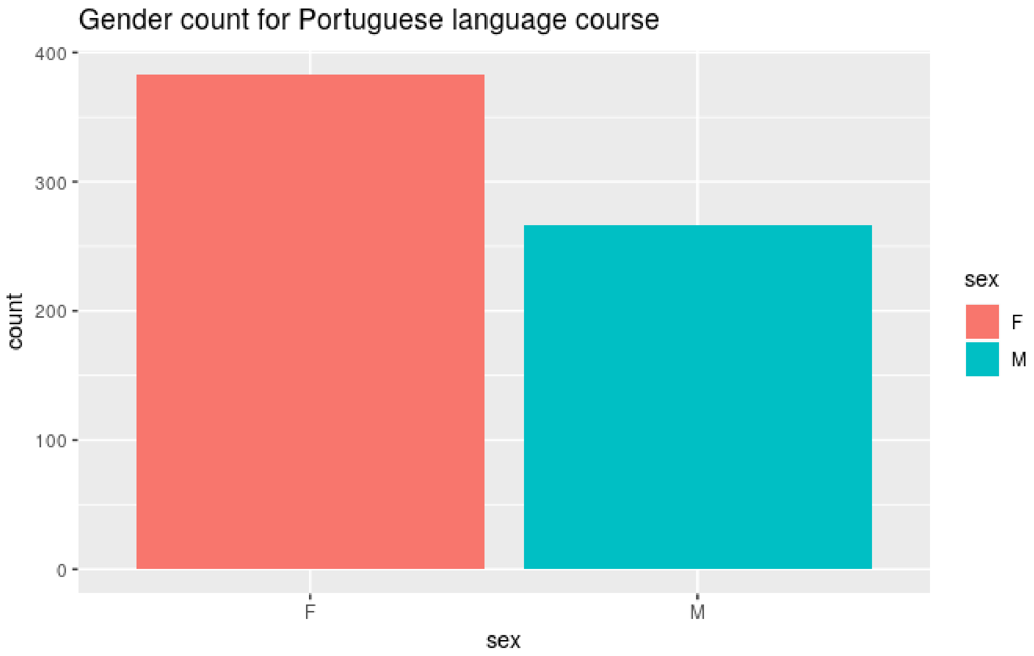


Figure 3

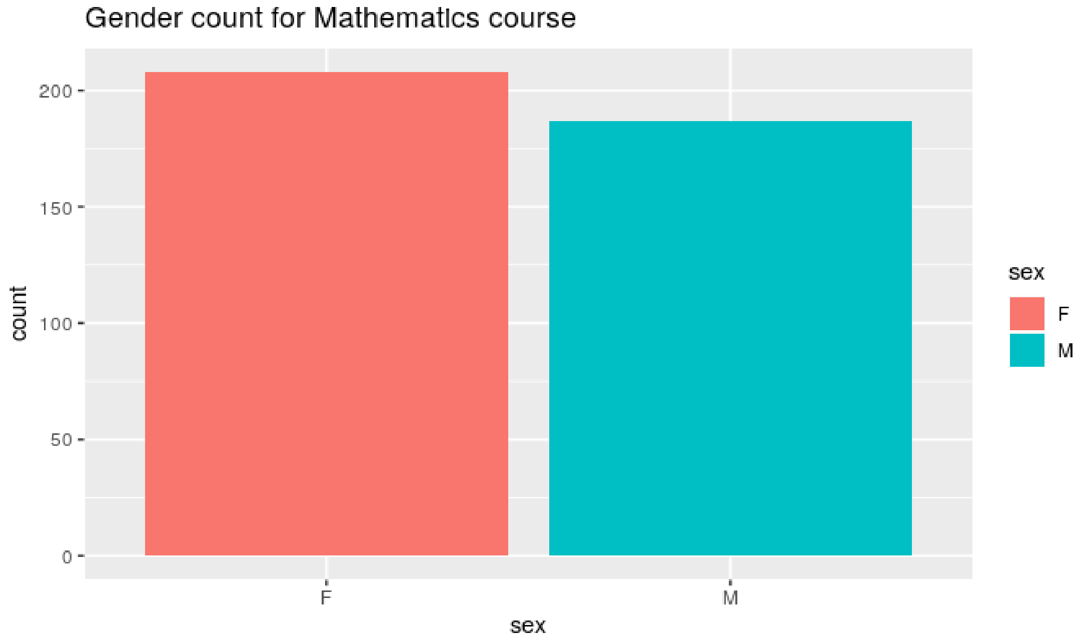


Figure 4

Table 2, Figure 3 and Figure 4 shows the count of students in each course. From the bar plots, it is obvious that we have more females than males in the two courses. However, the difference in that of Mathematics course was not as much as that of the Portugal language course.

a) Who does better at school? Do girls perform better or do boys gets better results than girls?

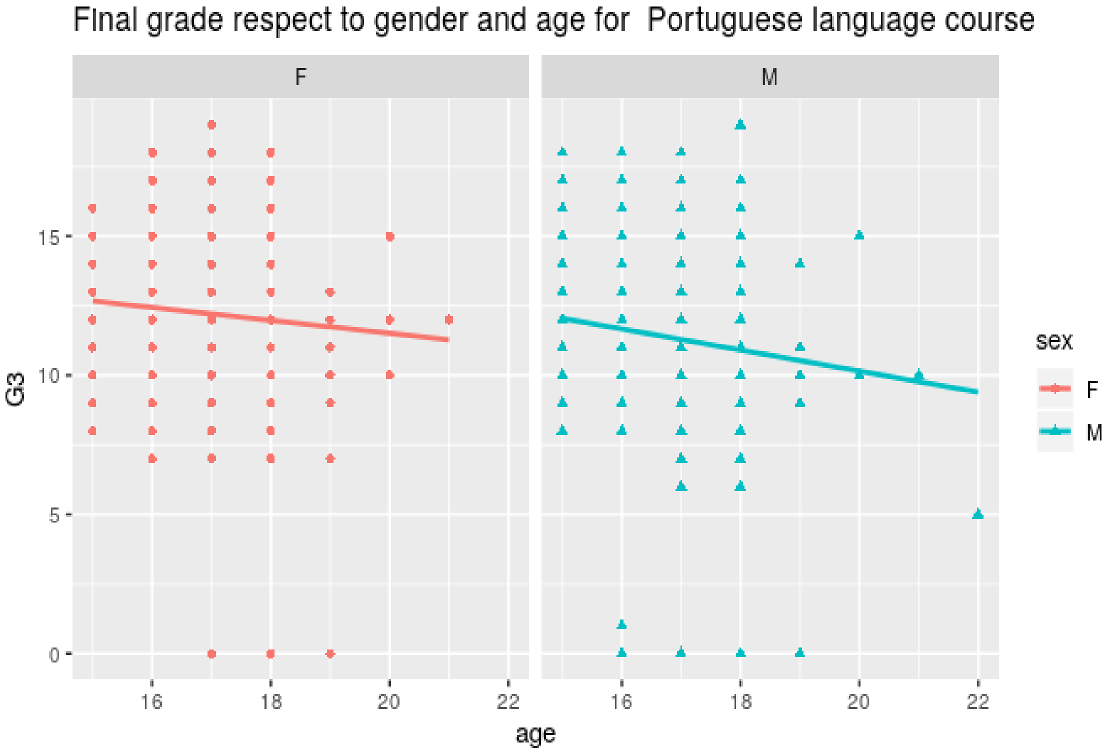


Figure 5

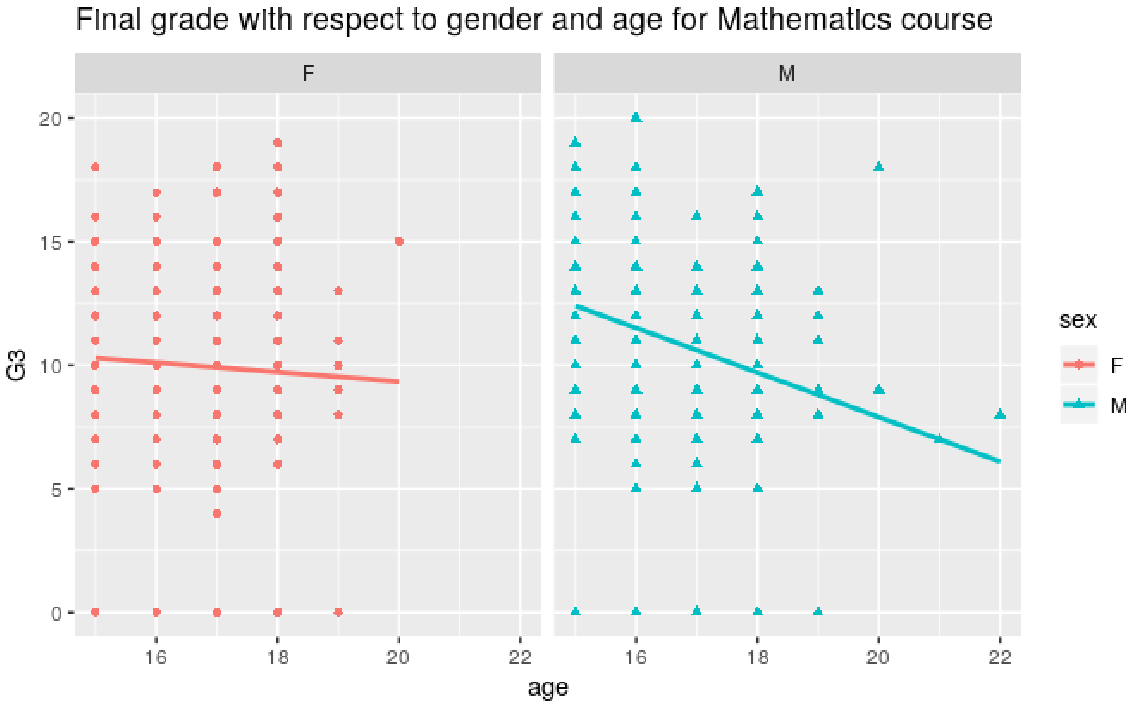


Figure 6

From Figure 5 and Figure 6, We can see that a sharp decrease in the males performance in Mathematics course. Hence, we would like to know why male students performance decrease with age. let's explore more.

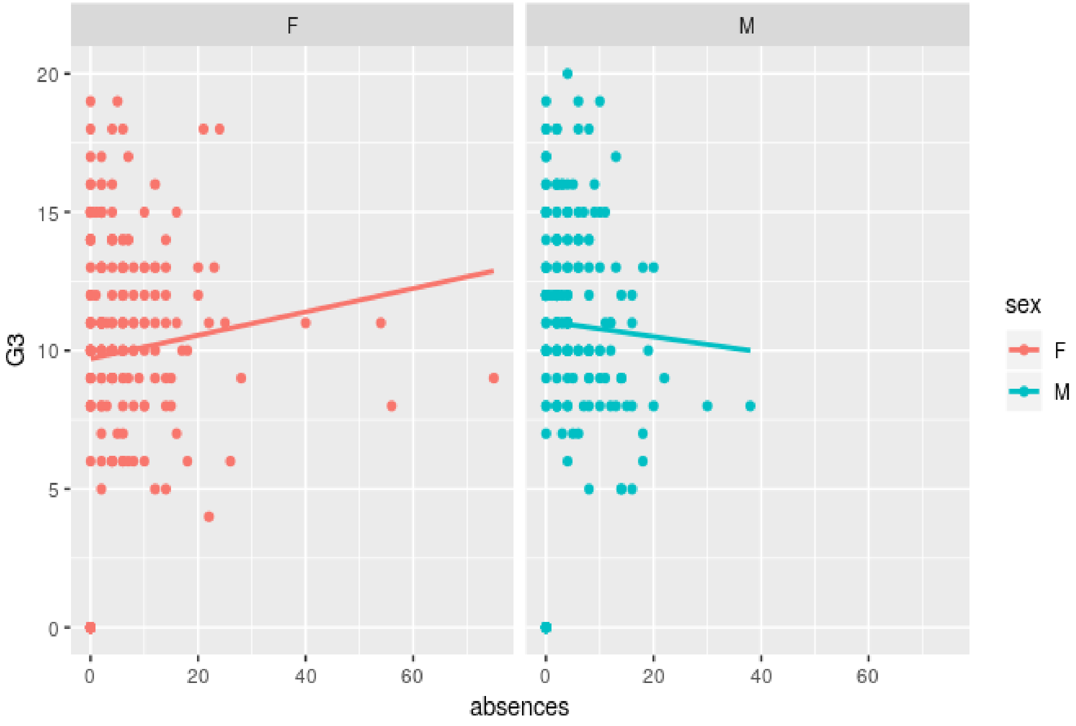


Figure 7

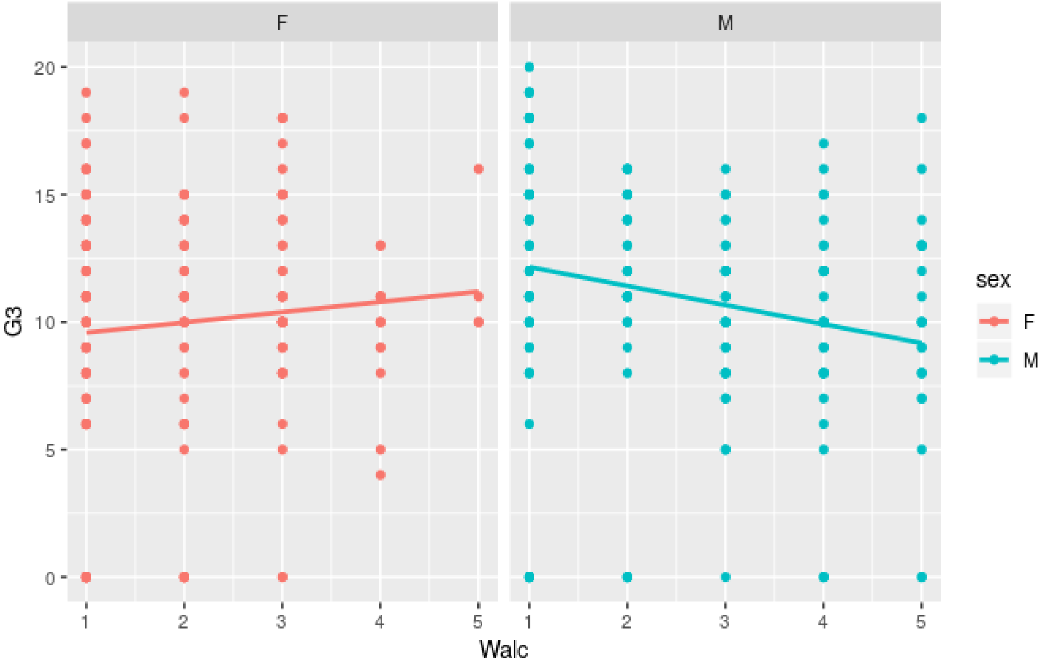


Figure 8

Obviously, we can observe that ***absences*** - (number of school absences) and ***Walc*** - (weekend alcohol consumption) might be the possible factors why male performance dropped in Mathematics course. On the other hand, absences - (number of school absences) and Walc -(weekend alcohol consumption) has no effect on the performance of the females.

- Do kids of divorced parents score lower in the exams?

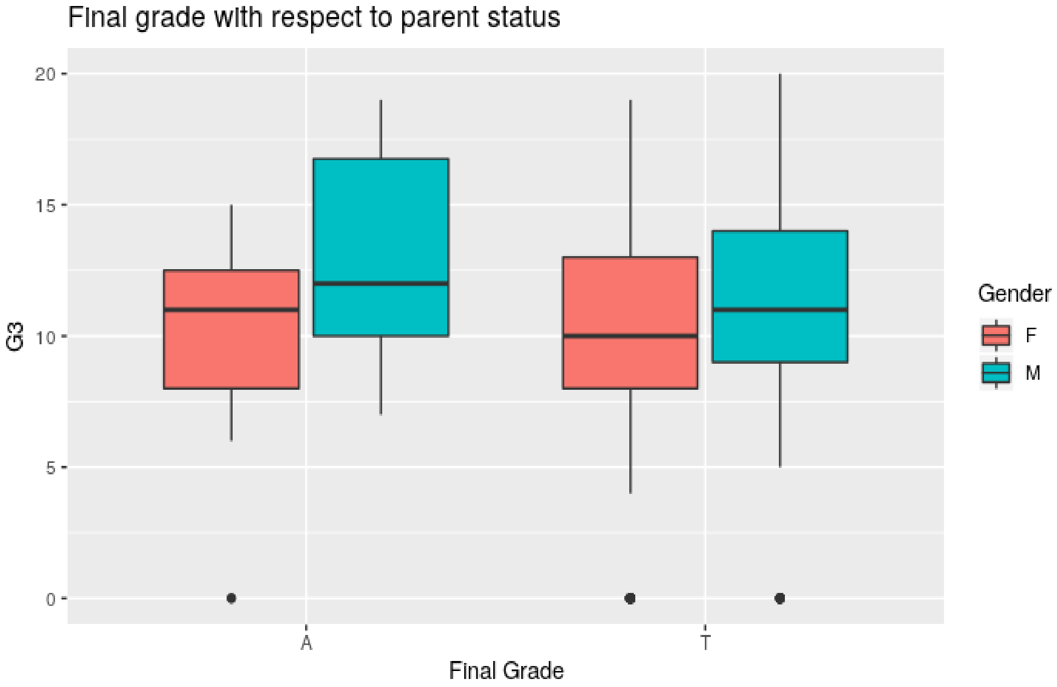


Figure 9

Figure 9 above shows the grade of student in relation with their parents cohabitation status (binary: 'T' - living together or 'A' – apart). From the above plot, we can deduce that, females whose parent are living together have higher score than than children whose parents are apart. This is applicable for both male student as well. This is for the mathematics course.

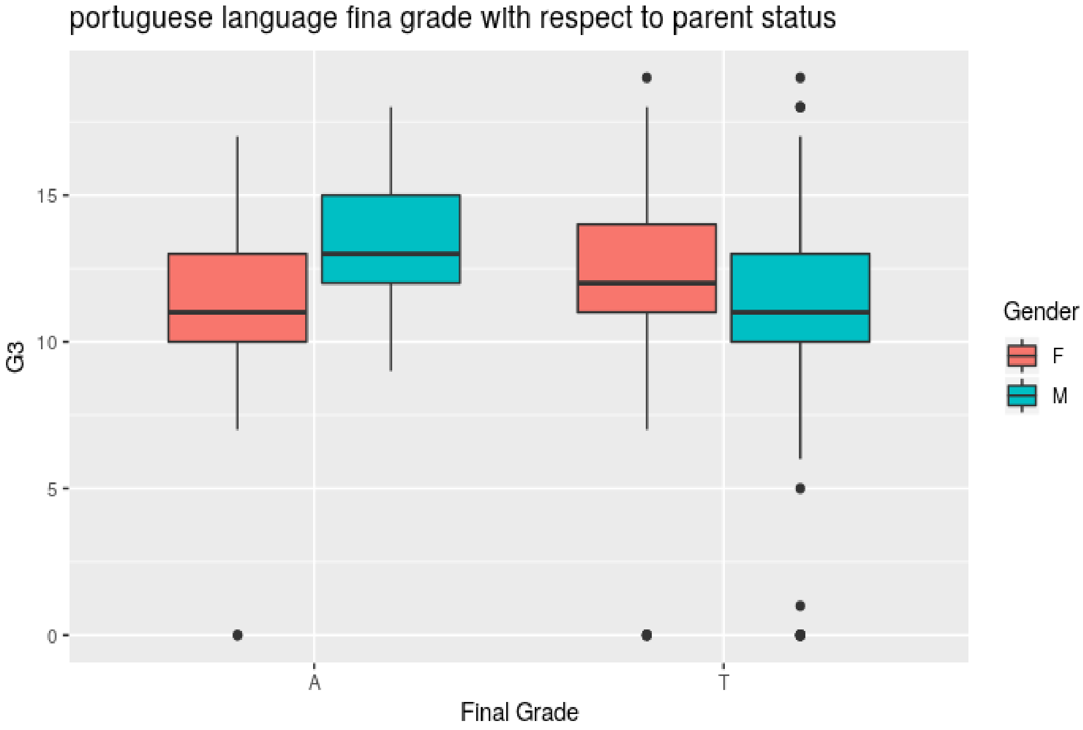


Figure 10

Figure 10 above shows the grade of student in relation with their parents cohabitation status (binary: 'T' - living together or 'A' – apart). From the above plot, we can deduce that, females whose parent are living together have higher score than than children whose parents are apart. However, This is not applicable for the male student , we can see that males students whose parents are apart has a higher score than those whose parent are living together. This is for the Portuguese language course. In conclusion, we can say that the females performs better than males in Mathematics course.

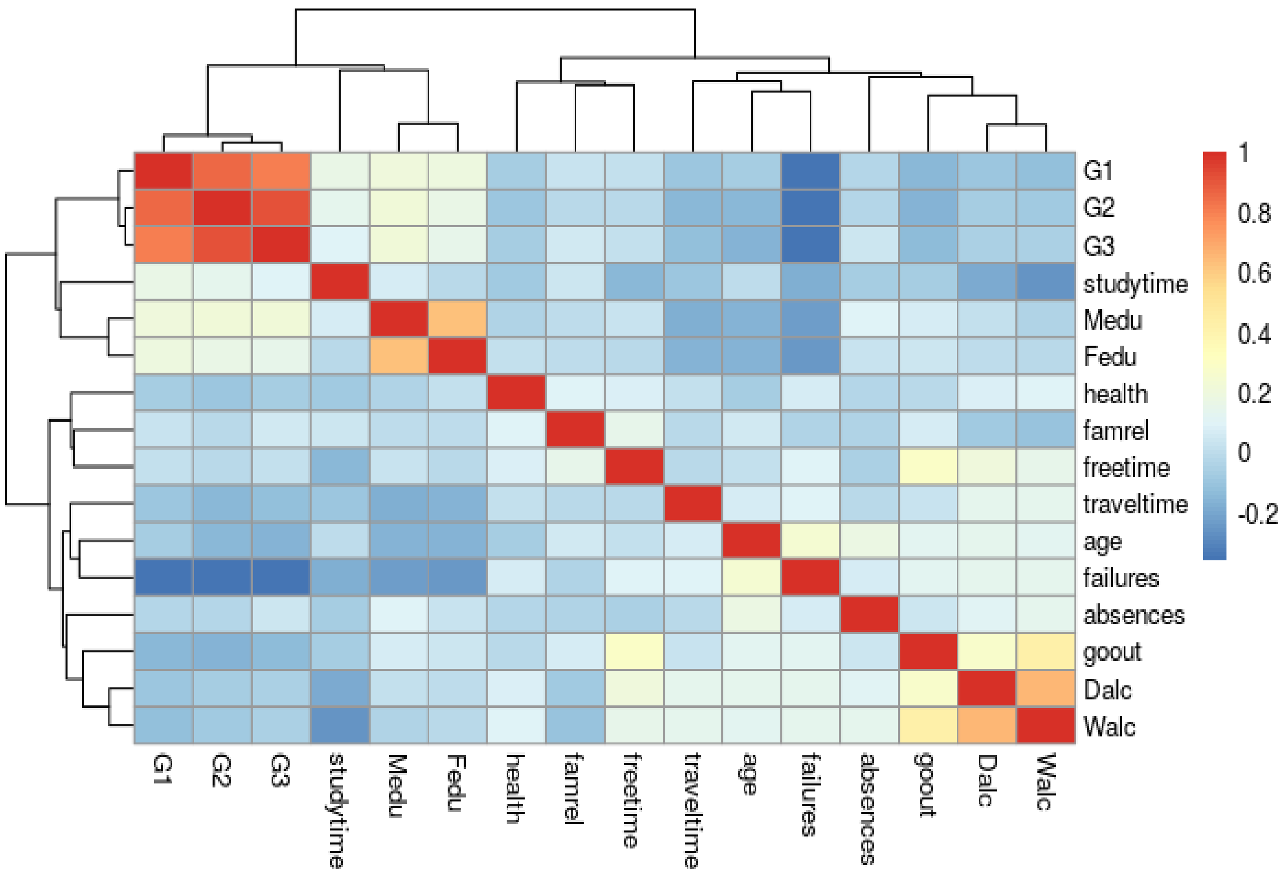
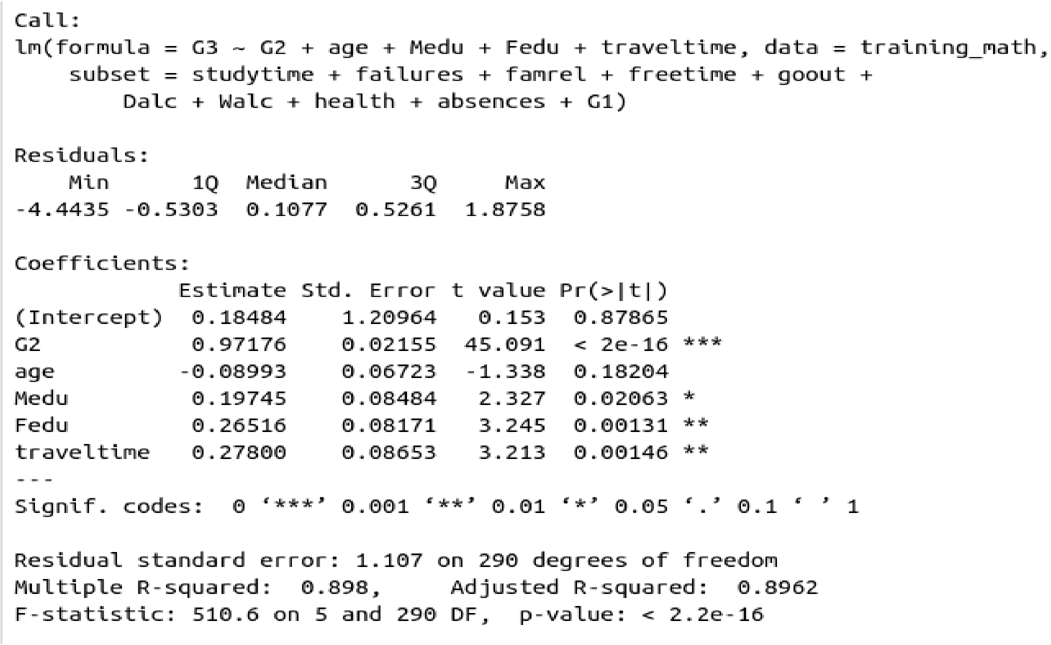


Figure 11

Figure 11 above, shows the correlation matrix for all the numeric variables in the Mathematics course data set. A correlation matrix plot shows the different variables as well as it's corresponding correlation with other variables. The red means highly correlated. Numbers from 0.8 up to 1 means a high correlation. Numbers from 0.3 up to 0.79 means moderately correlated, and lastly, numbers between 0 up to 0.29 is a week correlation. Numbers less than 0 means negative correlation. From the plot, we can infer that, there is a strong correlation between the final grade (G3) with G1, G2 which are first period grade and second period grade respectively.

Exploring the three different types of models

Model 1: Multiple Linear Regression Model



```
Call:
lm(formula = G3 ~ G2 + age + Medu + Fedu + traveltime, data = training_math,
    subset = studytime + failures + famrel + freetime + goout +
        Dalc + Walc + health + absences + G1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4435 -0.5303  0.1077  0.5261  1.8758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.18484    1.20964   0.153  0.87865
G2           0.97176    0.02155  45.091  < 2e-16 ***
age         -0.08993    0.06723  -1.338  0.18204
Medu         0.19745    0.08484   2.327  0.02063 *
Fedu         0.26516    0.08171   3.245  0.00131 **
traveltime   0.27800    0.08653   3.213  0.00146 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.107 on 290 degrees of freedom
Multiple R-squared:  0.898,     Adjusted R-squared:  0.8962
F-statistic: 510.6 on 5 and 290 DF,  p-value: < 2.2e-16
```

Figure 12

Above is the multiple linear regression of some of the numerical variables. A regression model is used to check for effects of 2 or more independent variables on one dependent variable. Here, in this case the dependent variable is G3 which is the student final score for Mathematics, while the independent variable here are G2, age, Medu e.t.c as seen from the plot above. From the model, we notice that the variance explained by the explanatory variables is 89.8%. We can also see that G2 has a very strong significant relationship in predicting students final score, as well as the Medu, Fedu and travel-time. Find below the residuals of the model.

Residual plots of the regression model.

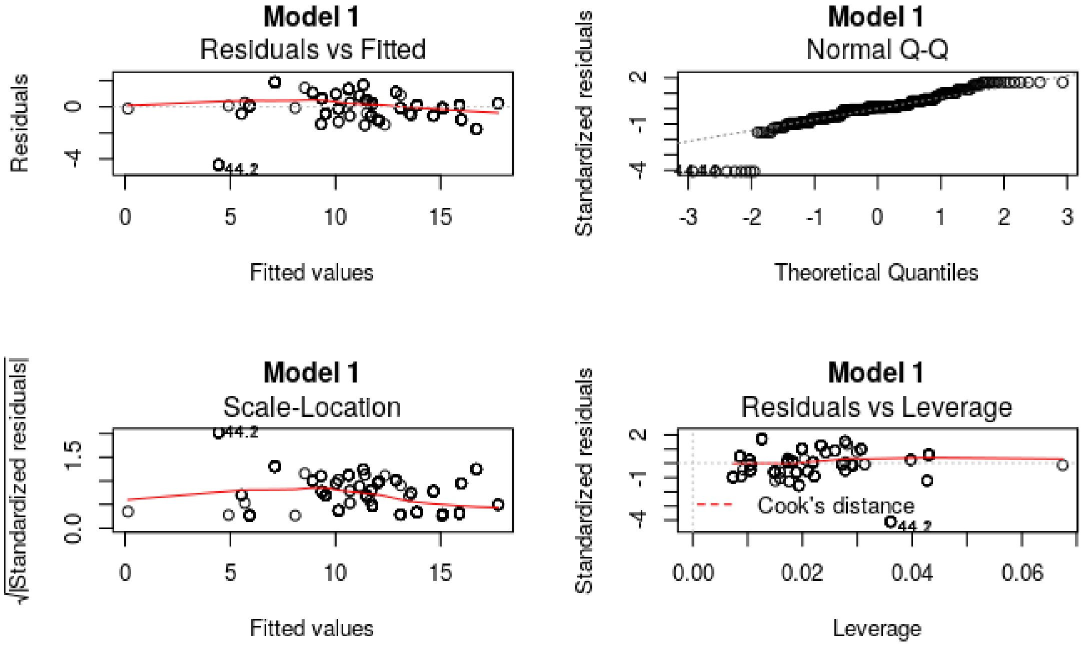


Figure 13

A Residual plots of the regression model is basically used to check for the assumption of OLS in regression. It helps us to know if our model is a true representation of our data. If possible, try to improve on it. The From the residual plots, one can observe that, the residual vs the fitted almost lies at 0, meaning that almost all the errors have zero mean. Also, the QQ plots shows that the data lies on the fitted line (although there are some outliers off the fitted line). So we can say it is a good representation of our data, however, we need to explore some other models to see if we can find one which is better than the regression model.

Model 2: Decision Tree Regression

A decision tree is a decision support tool whose final output is a tree with decision nodes. It can be used for either classification or regression problem. However, it's challenge is that, it is unstable, as a small change in the data can cause a very drastic change in the optimal decision of the tree. The decision is always made at the end of a node, and the approach is always a greedy approach or a recursive approach.

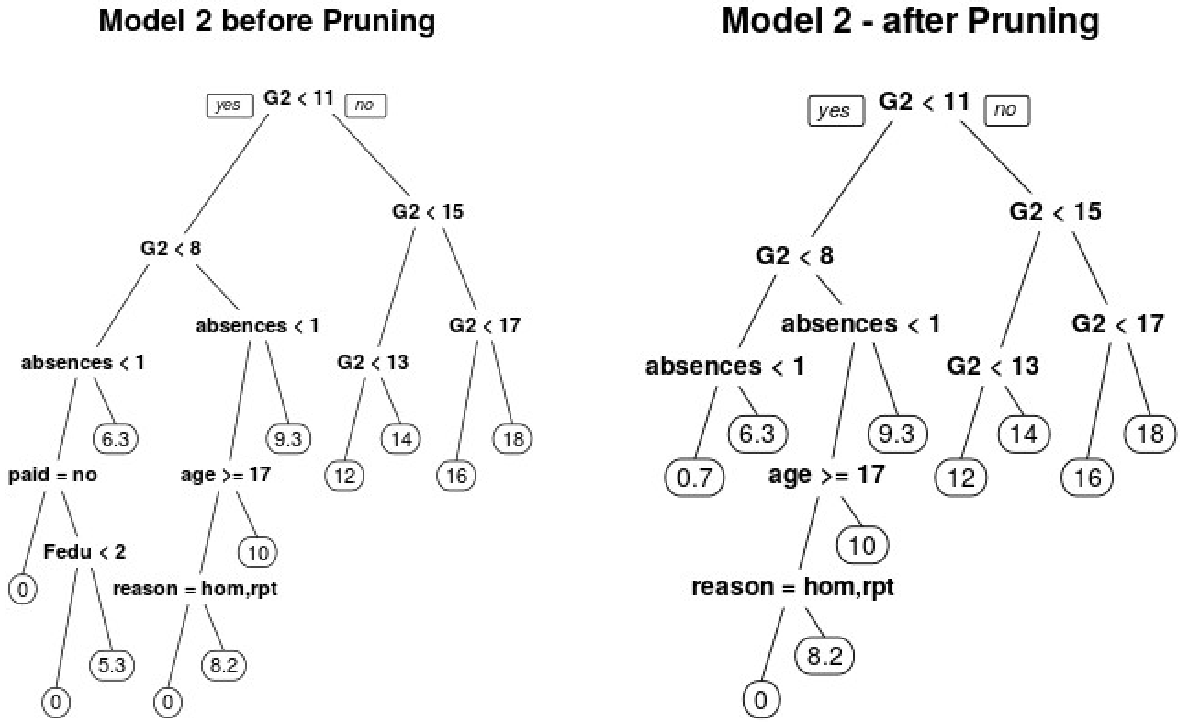


Figure 14

The graph on the left is the model before pruning was carried out, while that on the right is the model after pruning. **Tree pruning** methods address this problem as well as overfitting (a situation where the model cannot generalize well on new instances). From the right hand side, we realize that some unwanted nodes have been removed. We are now left with fewer nodes to make aid accuracy.

Model 3 : ***Random Forest Regression.***

The above model is a type of model that works additively, meaning that it's mode of making prediction is by combining several decisions trees together, this techniques is known as ensembling. With this, it can make better and more accurate predictions.

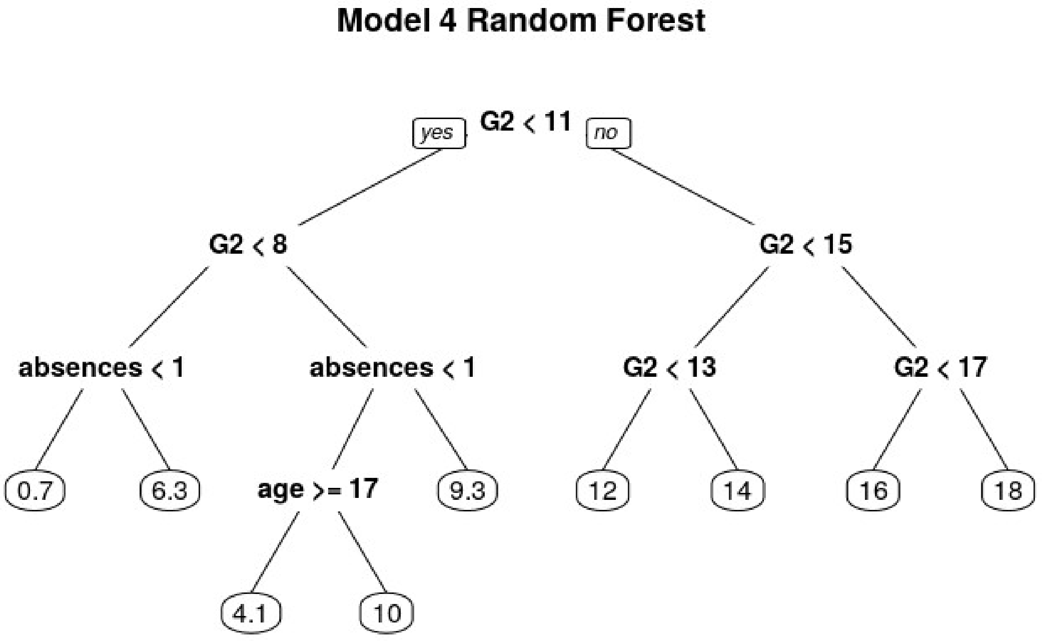


Figure 15

The above diagram is the output from the random forest, as stated above, the final decision is made by averaging several decision trees.

Model Evaluation

Table 3: Model Evaluation & Comparison

Model	RMSE	Correlation of actual value to predicted
Multiple Regression	2.583397	0.860
Decision Tree	2.127828	0.900
Random Forest	1.973326	0.914

After exploring the 3 models above, predictions were made using the test set. The correlation between the predictions made and the original test set were estimated. And from the table, is is observed that the Random forest model had the highest correlation compared to others. Also, in other to further validate our choice of model, the metric use for evaluation is the RMSE, we compere the RMSE and the Correlation of each of the models, and it was observed that, the Random Forest model, had the least RMSE as well as the highest correlation of (predicted vs actual), hence, we have sufficient reason choice the Random forest as the best model. This same analysis can be reproduce this same analysis for the Portuguese language course.

Variable Importance

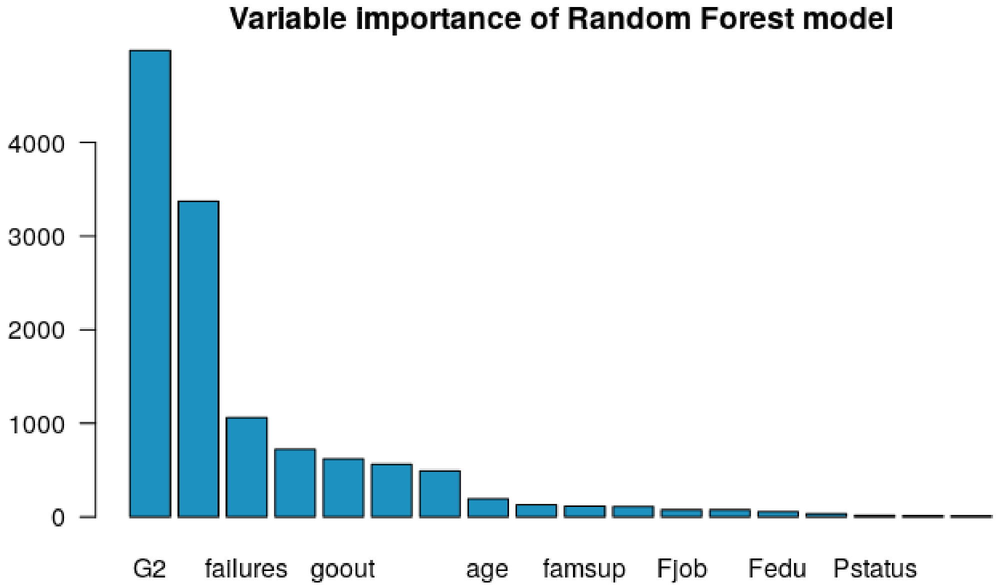


Figure 16

Figure 16 shows the variables that are important (in predicting students final grade in Mathematics Score, using the Random Forest Model). Out of the 36 features we had. From the table, we can see that **G2** was very much important, this is obvious because it is the student past grade (and it is highly correlated to his or her final grade as well, as seen in figure 11). Also, there were other factors like the **failures** - number of past class failures, **goout** - going out with friends, **age**, **famsup** - family educational support, **Fjob** - father's job (teacher, health care related and civil services), **Fedu** - father's education, and finally, the **Pstatus** - parent's cohabitation status.

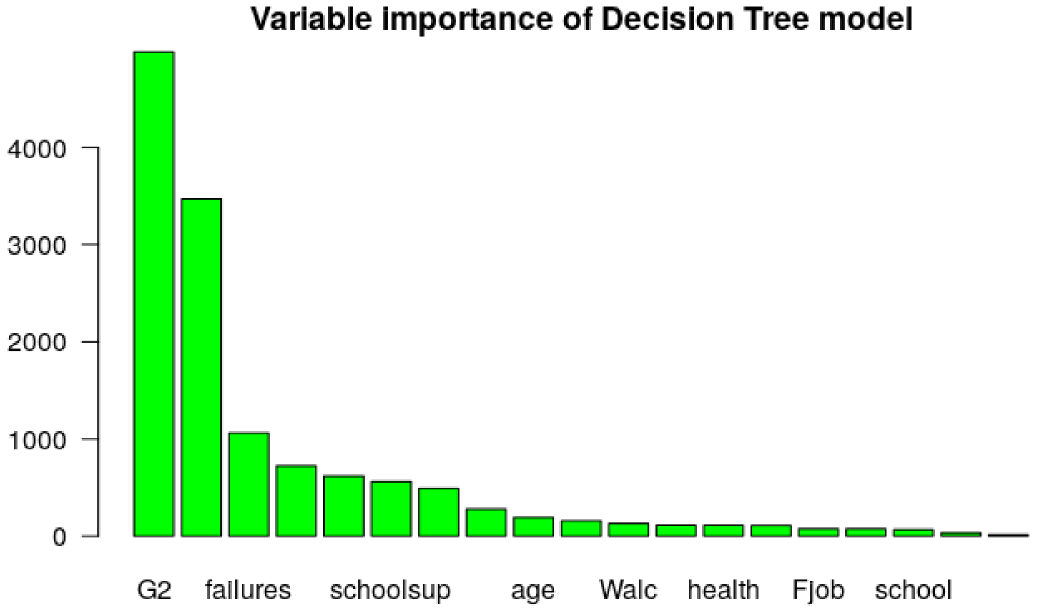


Figure 17

Figure 17 shows the variables that are important (in predicting students final grade in Mathematics Score) out of the 36 features we had. From the table, we can see that **G2** was very much important, this is obvious because it is the student past grade (and it is highly correlated to his or her final grade as well, as seen in figure 11). Also, there were other factors like the **failures** - number of past class failures, **schoolsup** - extra educational support, **age**, **Walc** - weekend alcohol consumption, **health** - current health status, **Fjob** - father's job (teacher, health care related and civil services), and finally, the **school** - student's school.

Summary

Survey was carried out and data was collected to analyze the performance of student for Mathematics and Portuguese language course. Some visualization was done in order to know the gender hwo performs best and why the other does not do well, from the visualization, we were able to see that most of the male did not perform well in Mathematics course, and when we investigated further, it was seen that, most male have reduction in their grades as they grow older, also, they absent from classes and finally, we were able to see that most of them take alcohol at week end. The main objective of the study was to have a good model that can predict students final score in mathematics given 36 features/explanatory variables. We explored three models namely, the (1) The Regression model, (2) Decision Tree and (3) Random Forest. After that, in order to choose the best model, we used 2 techniques to evaluate the model, RMSE and correlation. The model with the least RMSE, and whose correlation of (prediction vs true value) is highest, is chosen to be the best model. From the study, we found out that the random forest passed these criteria, hence, it was chosen as the best model. Note, for this study, we only experimented the three models only on the Mathematics score, however, the techniques can be reproduced for the Portuguese language course.

REFERENCES

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

2. http://www3.dsi.uminho.pt/pcortez/student.pdf

3. https://www.kaggle.com/micahshull/r-students-scores-linreg-tree-forest-svm

4. https://www.kaggle.com/hindelya/students-grade-prediction

Appendix

Loading the required libraries

```
#library(tidyverse)
require(readxl)
## Loading required package: readxl
library(RColorBrewer)
library(lattice)
#library(psych)    #
library(DataExplorer)
library(reshape2)
library(car)
## Loading required package: carData
library(caret)
## Loading required package: ggplot2
library(data.table)
library(e1071)
library(gridGraphics)
## Loading required package: grid
library(gridExtra)
library(cowplot)
library(lmtest)
library(gvlma)
```

loading the data for both maths and Portuguese language score

```
Math_students <- read_excel("student-mat.xls")
head(Math_students)

Por_students <- read_excel("student-por.xls")
head(Por_students)
str(Math_students)
str(Por_students)
```

Data cleaning

```
# the code bellow replaces the categorical variables to numerical
Math_students$sex = as.factor(Math_students$sex)
Math_students$school = as.factor(Math_students$school)
Math_students$famsize = as.factor(Math_students$famsize)
Math_students$Fjob = as.factor(Math_students$Fjob)
Math_students$Mjob = as.factor(Math_students$Fjob)
Math_students$Pstatus = as.factor(Math_students$Pstatus)
Math_students$reason = as.factor(Math_students$reason)
Math_students$guardian = as.factor(Math_students$guardian)
Math_students$schoolsup = as.factor(Math_students$schoolsup)
Math_students$famsup = as.factor(Math_students$famsup)
Math_students$paid = as.factor(Math_students$paid)
Math_students$activities = as.factor(Math_students$activities)
Math_students$nursery = as.factor(Math_students$nursery)
Math_students$higher = as.factor(Math_students$higher)
Math_students$internet = as.factor(Math_students$internet)
Math_students$romantic = as.factor(Math_students$romantic)
Math_students$school = as.factor(Math_students$school)
```

Data Visualization or EDA

```r
# what are the age distribution across the two courses
table(Math_students$age)
mean(Math_students$age)
ggplot(aes(x=age), data=Math_students)+
  geom_histogram(binwidth = 0.50, fill='darkred', color='black')+
  ggtitle("Age of students for Mathematics course")
table(Por_students$age)
mean(Por_students$age)
ggplot(aes(x=age), data=Por_students)+
  geom_histogram(binwidth = 0.50, fill='blue', color='black')+
  ggtitle("Age of students for Portuguese language course")
# check the number of female and male students in the school.
table(Math_students$sex)
ggplot(data=Math_students,aes(x=sex,fill=sex))+geom_bar()+
  ggtitle("Gender count for Mathematics course")
table(Por_students$sex)
ggplot(data=Por_students,aes(x=sex,fill=sex))+geom_bar()+
  ggtitle("Gender count for Portuguese language course")
# Final grade with respect to gender and age
ggplot(data=Math_students,aes(x=age, y=G3, col=sex, shape=sex))
+geom_point()+geom_smooth(method="lm",se=F)+facet_grid(~sex)+
  ggtitle("Final grade with respect to gender and age for Mathematics
course")
# Final grade with respect to gender and age
ggplot(data=Por_students,aes(x=age, y=G3, col=sex, shape=sex))
+geom_point()+geom_smooth(method="lm",se=F)+facet_grid(~sex)+
  ggtitle("Final grade respect to gender and age for  Portuguese
language course")
# Attendance and Final grade for both gender in Mathematics
ggplot(data=Math_students,aes(x=absences, y=G3, col=sex))+geom_point()
+geom_smooth(method="lm",se=F)+facet_grid(~sex)
# alcohol consumption and Final grade for both gender
ggplot(data=Math_students,aes(x=Walc, y=G3, col=sex))+geom_point()
+geom_smooth(method="lm",se=F)+facet_grid(~sex)
```

c) Do kids of divorced parents score lower in the exams?

```r
ggplot(data=Math_students,aes(x=Pstatus, y=G3, fill=sex))
+geom_boxplot()-> obj1
obj1+labs(title="Mathematics final grade with respect to parent status",
x="Final Grade", fill="Gender")
ggplot(data=Por_students,aes(x=Pstatus, y=G3, fill=sex))+geom_boxplot()-
> obj1
obj1+labs(title="portuguese language fina grade with respect to parent
status", x="Final Grade", fill="Gender")
library(pheatmap)
numeric_features <- Filter(is.numeric, Math_students)
pheatmap(cor(numeric_features))
```

Machine Learning Model Preprocessing ######

Splitting into the Training set and Test set, 75% of data as sample from total ‘n’ rows of the data

```
set.seed(101) # Set Seed so that same sample can be reproduced in future
also
sample <- sample.int(n = nrow(Math_students), size =
floor(.75*nrow(Math_students)), replace = F)
training_math <- Math_students[sample, ]
testing_math  <- Math_students[-sample, ]

dim(training_math)
dim(testing_math)
```

##### First Model is Multiple Regression Analysis. #######

```
mod2 <- lm(G3 ~ G2 + age + Medu+ Fedu + traveltime, studytime + failures
+ famrel + freetime + goout+ Dalc + Walc + health + absences + G1, data
= training_math);
summary(mod2)
```

Diagnostic Plots

```
options(repr.plot.width=14, repr.plot.height=7)
par(mfrow = c(2,2)); plot(mod2)
```

Second model is Decision Tree Regression

```
# r part has built in 10 fold cross validation
# method must be set to anova for regression
library(rpart)
library(rpart.plot)

Mod3 = rpart(formula = G3 ~ .,
           data = training_math,
           method = "anova")
```

Grow the Decision Tree

```
mod3_2 = rpart(formula = G3 ~ .,
             data = training_math,
             method = "anova",
             control =rpart.control(minsplit = 5, cp=0.005))
```

Prune the Tree by setting error rate

```
mod3_2_pruned <- prune(mod3_2, 0.01)

# compare the before and after pruning models
par(mfrow = c(1,2));
prp(mod3_2, main = "Model 2 before Pruning");
prp(mod3_2_pruned, main = "Model 2 - after Pruning")
```

Third Model Random Forest Regression#######

```
set.seed(1234)
mod4 = rpart(formula = G3 ~ .,
           data = training_math,
           method = "anova")

#####   Visualize the Random Forest     #####
prp(mod4, main = "Model 4 Random Forest")
```

Evaluate and comparing model accuracy ######

Predicting the Test set results

```
testmod2 = predict(mod2, newdata = testing_math[0:32]);
testmod3 = predict(mod3_2_pruned, newdata = testing_math[0:32]);
testmod4 = predict(mod4, newdata = testing_math[0:32])
```

correlation actual and predicted

```
#strong association between the predicted and actual
# over 75 correlation is a good model
cor_mod2 = cor(testmod2, testing_math$G3);
cor_mod3 = cor(testmod3, testing_math$G3);
cor_mod4 = cor(testmod4, testing_math$G3)
```

RMSE = root mean square error

```
library(Metrics)
##
## Attaching package: 'Metrics'
## The following objects are masked from 'package:caret':
##
##     precision, recall
rmse_mod2 = rmse(testmod2, testing_math$G3);
rmse_mod3 = rmse(testmod3, testing_math$G3);
rmse_mod4 = rmse(testmod4, testing_math$G3)
```

Correlation of actual value to predicted

```
cat("\nModel 1 RMSE = ", rmse_mod2, "Correlation of actual value to 
predicted = ", round(cor_mod2,3));
cat("\nModel 2 RMSE = ", rmse_mod3, "Correlation of actual value to 
predicted = ", round(cor_mod3,3));
cat("\nModel 3 RMSE = ", rmse_mod4, "Correlation of actual value to 
predicted = ", round(cor_mod4,3))

# library(caret)   # To see the most important variables
caret::varImp(mod2);
caret::varImp(mod3_2)
caret::varImp(mod4)
```

Feature Importance

```
#par(mfrow = c(1,2))
options(repr.plot.width=30, repr.plot.height=7)
vi1 <- mod4$variable.importance;
barplot(vi1, horiz = F, las = 1, col = "#1D91C0",
        main = "Variable importance of Random Forest model")
#par(mfrow = c(1,2))
options(repr.plot.width=30, repr.plot.height=7)
vi1 <- mod3_2_pruned$variable.importance;
```

```
barplot(vi1, horiz = F, las = 1, col = "green",
        main = "Variable importance of Decision Tree model")
```

Comparing the 3 predictions from the different models

```
dt_prediction = (data.frame((testmod2), (testing_math$G3)))

colnames(dt_prediction) <- c("Predicted final score from regression
model","Real Score")
head(dt_prediction,10)
dt_prediction = (data.frame((testmod3), (testing_math$G3)))

colnames(dt_prediction) <- c("Predicted final score from Decision Tree
model","Real Score")
head(dt_prediction,20)
dt_prediction = (data.frame((testmod4), (testing_math$G3)))


colnames(dt_prediction) <- c("Predicted final score from Random
forest model","Real Score")
head(dt_prediction,20)
```