Analysis and Predictive Modeling of Final Grade

Brent Knight

Abstract

Through the use of supervised learning, the project consisted of a classification problem using a dataset that contained the social, gender, and study data from secondary school students in Portugal. My main focus was how all these many factors of a student affect their final grade in the class. Through the data mining techniques of logistic regression and the decision tree method, I came up with two models predicting whether a student passed or failed their math class. After careful analysis and model comparison, the pruned tree method gave better results and was able to accurately predict a student's success.

Introduction

The main goal for this project was to create a model for the dataset that would accurately predict whether or not each student would pass or fail their math class. This was done using explanatory variables which were the attributes of the students in order to classify which of these students would receive a passing grade represented by the response variable of "pass." There have been python projects dealing with this data trying to find a correlation between students' alcohol consumption and its relation to their final grade in the class along with visualizing the data and comparing variables. While python has some benefits, I used the software language R in order to create the models. I also didn't just try and see how alcohol affects a student's grade but instead used many variables such as gender and study time to see how a student's final grade is affected.

The dataset used comes from the UCI Machine Learning database and was obtained by surveying Portuguese students in both math and Portuguese language classes. I found this dataset to be very interesting considering it includes many attributes of a student that wouldn't normally come to mind when thinking of different factors affecting their in-class performance. For example, the dataset includes whether or not a student has internet access at home. It is important to note that the data is obtained from a survey done in the country of Portugal. So, I made sure to keep this in mind when choosing the cutoff point for my newly created variable. The response variable "pass" depended heavily on this fact since the grading system in Portugal is much different than my own. It is based off of a scale from 0-20 with a score of 10 or higher being a passing grade. I also decided to analyze just the math class in order to avoid variability between the two classes and to also focus more on a smaller sample.

Based on the many explanatory variables, I wanted to create the best model in order to classify each student as accurately as possible into one of two groups: pass or fail. The problem was finding a model that would use the explanatory variables to conclude whether or not a student received a passing grade in their math class. Methods such as Principal Component Analysis would be hard to apply to this dataset since it consists mostly of categorical variables. Logistic regression and the decision tree method were chosen in order to find a model to fit the dataset. After running these models, there were fairly positive results finding that it accurately predicts the success of a student in the math class with a true positive rate of 94.12% and a false positive rate of 64.29%. The source data can be found here: https://www.kaggle.com/uciml/student-alcohol-consumption. The software used to conduct the analysis and create the models is called RStudio and the programming language is called R.

Data and Methods

All the variables in the dataset were used.

- school - the student’s school (Gabriel Pereira or Mousinho da Silveira)
- sex - male or female
- age - student’s age (15 - 22)
- address - student’s home address type (Urban or Rural)
- famsize - student’s family size (Less than 3 or greater than 3)
- Pstatus - parent’s cohabitation status (Living together or apart)
- Medu - mother’s education (primary education, 5th - 9th grade, secondary education or higher education)
- Fedu - father’s education (same as Medu)
- Mjob - mother’s job (teacher, health care related, civil services, at home, or other)
- Fjob - father’s job (same as Mjob)
- reason - reason for choosing the school (close to home, school reputation, cmyse preference or other)
- guardian - student’s guardian (mother, father, or other)
- traveltime - home to school travel time (< 15 min, 15 - 30 min, 30 - 60 min or > 60 min)
- studytime - weekly study time (< 2 hrs, 2 - 5 hrs, 5 - 10 hrs, > 10 hrs)
- failures - number of past class failures (1, 2, 3 or 4+)
- schoolsup - extra educational support (yes, no)
- famsup - family educational support (yes, no)
- paid - extra paid classes for cmyse subject (yes, no)

- activities - extracurricular activities (yes, no)
- nursery - attended nursery school (yes, no)
- higher - wants to attend higher education (yes, no)
- internet - has Internet access at home (yes, no)
- romantic - in a romantic relationship (yes, no)
- famrel - quality of family relationships (very bad to excellent)
- freetime - free time after school (very low to very high)
- goout - going out with friends (very low to very high)
- Dalc - workday alcohol consumption (very low to very high)
- Walc - weekend alcohol consumption (very low to very high)
- health - current health status (very bad to very good)
- absences - number of school absences (0 - 93)
- G1 - first period grade (0 - 20)
- G2 - second period grade (0 - 20)
- G3 - final grade (0 - 20)

I decided to transform some of the variables for the tests. Instead of using G3, I used a new variable called “pass” which took the values of G3 and replaced them with “yes” or “no”. “yes” for a value greater than or equal to 10 and “no” for a value less than 10. This was done in order to know whether the other variables had anything to do with influencing whether or not a student passed. I researched the grading system in Portugal and came up with 10 as the cutoff point in their schools. I also ordered a few of the variables in the data set.

The ones I ordered were famsize, Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, and health since they were categorical variables containing some order.

For my analysis, I decided to do a logistic regression and a decision tree. I decided to use these two methods because my purpose for this project was to find out which variables were the greatest factors in determining whether a student would pass their class. The dataset has many variables that are categorical and these methods support these kinds of variables while others like the PCA method would have trouble dealing with categorical variables.

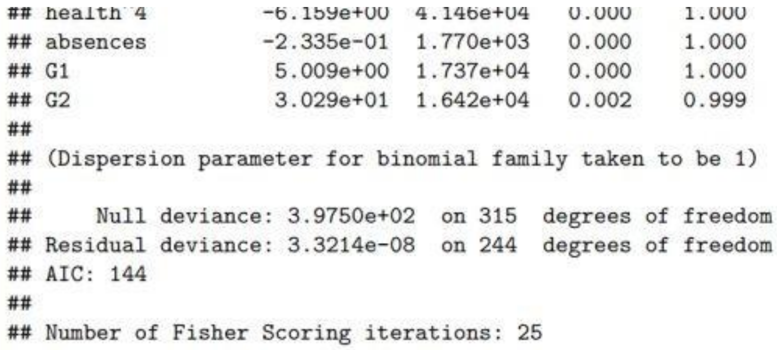
Luckily, there were no missing values in the dataset and I did not run into problems when figuring out the correct way to transform the variables. There was a time where I changed the age variable which did not need to be changed and that resulted in strange logistic regression results and a very small tree. At first I thought it was just an error in the code for the logistic regression method as the tree did make sense even though being quite small. I originally transformed the "age" variable into a variable called "legal" which would result in "yes" or "no" with the cutoff being 18 in Portugal. In the end I figured that it was an unnecessary change as that was more focused on the "alcohol" variable rather than the "pass" variable which was what the project is based on. I also did not include some variables like G1 and G2 (first period grade and second period grade) for one of my testing sessions as I thought those variables were extremely significant and wanted to see if the other variables played any role in determining final grade without the them. After more careful testing, I concluded that the dataset could not be properly analyzed without these variables so I left them in.

I predicted and observed that even though there were many variables in the dataset, only a few variables were significant in determining whether a student passed their class or not. Some of the results were obvious like G1 and G2 playing a huge factor in my response variable. However, there were a few surprise results in the decision tree method that I did not expect which will be shown later in my results.

The data mining process for my project began with loading the data into RStudio and downloading the necessary libraries for the code. I then transformed some of the variables in the data into ordered factors as these variables had numeric values that were connected to specific levels (very low, very high, job names, etc.). I also created a new factor response variable called “pass” which gave “yes” if the value in G3 (final grade) was 10 or greater and “no” if the value was less than 10. I then created a new dataset called “success.subset” which included all the variables besides G3 from the original dataset and the new “pass” variable. I created a training and test dataset from success.subset. The training dataset has 316 randomly sampled observations (80%) and the test dataset has the rest (20%). I took a part of the “success.subset” data in order to train my models then the part left over was used as the test data that I tested the trained models on. I needed to do this step for my logistic regression method and decision tree method. For the logistic regression method, I used the glm() function to create my model and the summary() function to get my results. For the decision tree method, I created a tree using the tree() function. I then used the cv.tree() function to utilize a 10-fold cross validation and find the best possible size for my tree. After I found the best size, I pruned my original tree to the best size to get my pruned tree. I then created a confusion matrix of the tree model for my test dataset. I did this so I could find the true positive rate and false positive rate which will be explained more in the analysis. Finally, I found the ROC curves and AUC values for both the logistic regression model and decision tree model. For both of these techniques, the result that is closer to 1 is better. This was done so I could figure out which of the following models performed better.

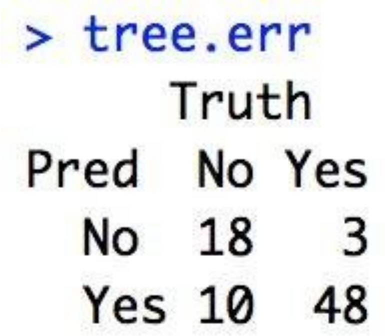
Results

Because most of my variables were categorical, the output for my summary of glm was very large. Here is a portion of my results. For example, I can see the coefficient estimates of the variables in the first column along with their p-values in the last. Larger coefficient estimates carry more weight in predicting the response variable as seen in G1 and G2.



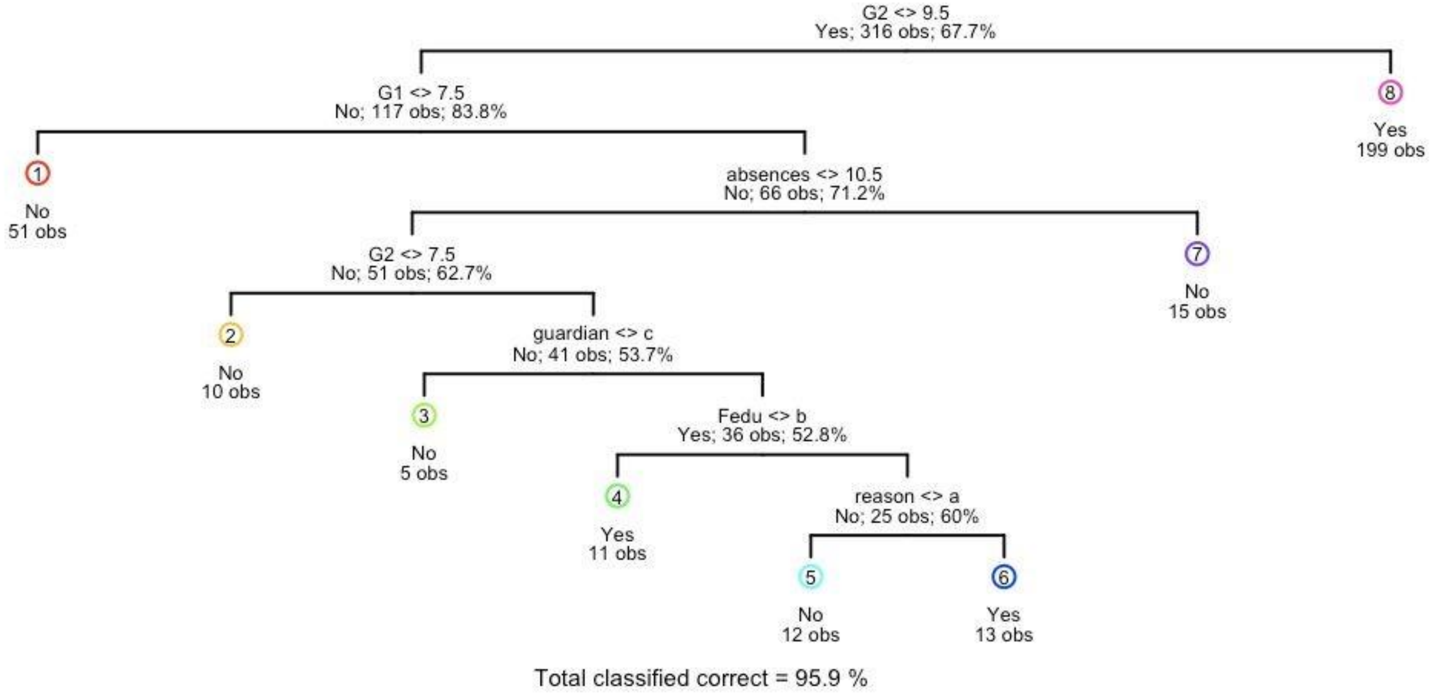
```
## health 4         -6.159e+00  4.146e+04   0.000    1.000
## absences         -2.335e-01  1.770e+03   0.000    1.000
## G1                5.009e+00  1.737e+04   0.000    1.000
## G2                3.029e+01  1.642e+04   0.002    0.999
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3.9750e+02  on 315  degrees of freedom
## Residual deviance: 3.3214e-08  on 244  degrees of freedom
## AIC: 144
##
## Number of Fisher Scoring iterations: 25
```

Before looking at the Pruned Tree Diagram, I first created a confusion matrix to check my true positive rate and my false positive rate. “Pred” stands for predicted values from the decision tree method while "Truth" stand for the true values from my test set. So, each number in the table represents different outcomes. I can calculate the true positive rate (the percentage of students that were correctly identified) and the false positive rate (the percentage of students incorrectly identified) by using the formulas TPR = TP/(TP+FN) and FPR = FP/(TN+FP). In this case TPR <- 48/(48 + 3) = 0.94 and FPR <- 18/(18 + 10) = 0.64.

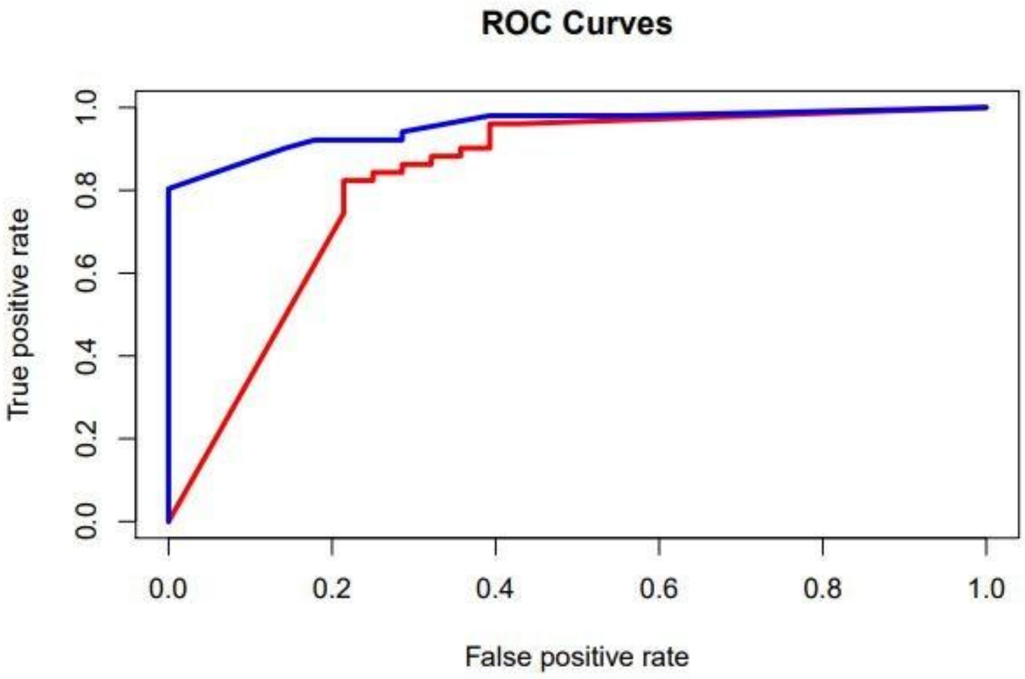


```
> tree.err
     Truth
Pred  No Yes
  No  18   3
  Yes 10  48
```

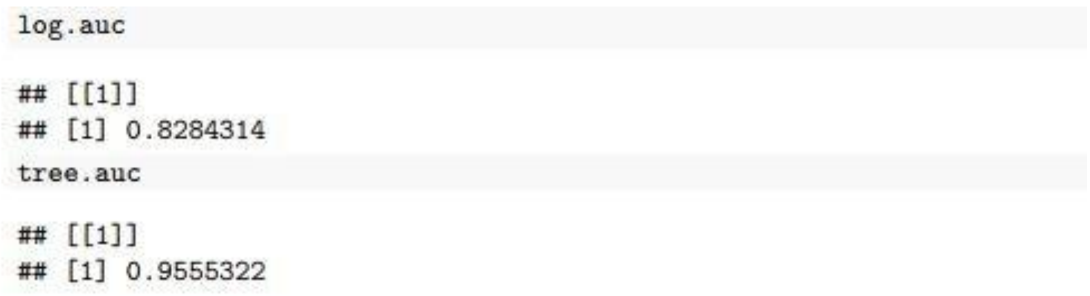
The decision tree map shows the most important variables about a student, represented in branches, to the conclusion of "yes" or "no" which is pass or fail. I can see that the variable G2 classified a good majority of the observations into "yes" then other variables come into play such as G1 and absences to explain the rest of the pass/fail split. The variables that surprised us were student's guardian, father's education level, and reason for choosing their school.



I plotted my two models on the same plot to compare their performance. When analyzing ROC curves, I am generally looking for a curve that hugs the top left corner of the graph. The logistic regression model is shown in red while the decision tree model is in blue. Clearly, the decision tree method does better with the blue line hugging the top left.



AUC works hand in hand with ROC curves as it calculates the area under the curve. It is another model performance indicator. The closer to 1 the better. Here I see that the pruned tree and logistic regression both performed well but pruned tree has a higher value.



```
log.auc

## [[1]]
## [1] 0.8284314
tree.auc

## [[1]]
## [1] 0.9555322
```

Discussion

In the end, I did achieve my goal in figuring out which of the following variables were significant in determining how the students would do in their class. Some challenges I faced were realizing that some of the variables needed to be ordered and really just finding a dataset that had enough interesting information. Using the tree model which was the best model as determined by the ROC curves and the AUC values, I see that G2 is the most significant variable in determining whether a student would pass or fail. This makes sense since a student would receive their second to last grade before their final grade in the class and depending on what grade they received, they would change their habits in order to pass the class. The second significant variable is G1. This is the first grade the students received and similarly to the G2 variable, it may cause the students to change their habits in order to pass. It makes sense that this is less significant than G2 as the first grade students get is near the beginning of the school semester and some students may not be alarmed by a poor grade received so early in the term knowing that there will be time to raise it. Another influential explanatory variable is “absences” which makes logical sense. If students are absent to class, how are they supposed to know the class material? Attendance also may be a part of the grading system and multiple absences may affect final student grades negatively. If I were to turn this report into a long term project I could try adding in the other data from the Portuguese language class and compare results. I could also go even bigger by gathering data from other school around Europe and compare which factors affect grade depending on their country.

Citations

https://www.kaggle.com/uciml/student-alcohol-consumption

https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

An Introduction to Statistical Learning, James et al AKA “ISLR” (for all students)

R/RStudio

Appendix

~~~~
```{r, message=FALSE, warning=FALSE, include=FALSE}
library(readr)
library(tidyverse)
library(ROCR)
library(tree)
library(maptree)
library(class)
library(lattice)
library(dplyr)
```

```{r, message=FALSE, warning=FALSE}
success <- read.csv("C:/Users/Brent/Downloads/student-mat.csv", header = TRUE)
```

```{r, message=FALSE, warning=FALSE}
success <- success %>%
  mutate(pass = as.factor(ifelse(G3 < 10, "no", "yes")))
```

```{r, message=FALSE, warning=FALSE}
success$famsize <- ordered(success$famsize, levels = c("LE3", "GT3"))
success$Medu <- ordered(success$Medu, levels = c(0:4), labels = c("0", "1", "2", "3", "4"))
success$Fedu <- ordered(success$Fedu, levels = c(0:4), labels = c("0", "1", "2", "3", "4"))
success$traveltime <- ordered(success$traveltime, levels = c(1:4), labels = c("1", "2", "3", "4"))
success$studytime <- ordered(success$studytime, levels = c(1:4), labels = c("1", "2", "3", "4"))
success$failures <- ordered(success$failures, levels = c(0:3), labels = c("0", "1", "2", "3"))
success$famrel <- ordered(success$famrel, levels = c(1:5), labels = c("1", "2", "3", "4", "5"))
~~~~

~~~~
success$freetime <- ordered(success$freetime, levels = c(1:5), labels = c("1", "2", "3", "4", "5"))
success$goout <- ordered(success$goout, levels = c(1:5), labels = c("1", "2", "3", "4", "5"))
success$Dalc <- ordered(success$Dalc, levels = c(1:5), labels = c("1", "2", "3", "4", "5"))
success$Walc <- ordered(success$Walc, levels = c(1:5), labels = c("1", "2", "3", "4", "5"))
success$health <- ordered(success$health, levels = c(1:5), labels = c("1", "2", "3", "4", "5"))
```

```{r, message=FALSE, warning=FALSE}
success.subset <- success %>% select(school:G2, pass)
```

```{r}
set.seed(1)
randomsamp <- sample(1:nrow(success.subset), 316)
success.train <- success.subset[randomsamp,]
success.test <- success.subset[-randomsamp,]
dim(success.train)
dim(success.test)
```

```{r, echo=TRUE, message=FALSE, warning=FALSE}
success.glm <- glm(pass ~ ., data = success.train, family = binomial)
summary(success.glm)
```

```{r, message=FALSE, warning=FALSE}
set.seed(1)
success.tree <- tree(pass ~ ., data = success.train)
success.cv <- cv.tree(success.tree, K = 10, FUN = prune.misclass)
success.cv
best.size <- 8
```

```{r, message=FALSE, warning=FALSE}
success.prune <- prune.tree(success.tree,  best = best.size, method = "misclass")
draw.tree(success.prune, nodeinfo = TRUE, cex = .5, cases = "obs")
```

```{r, message=FALSE, warning=FALSE}
~~~~

~~~
predictions <- predict(success.prune, success.test, type = "class")
truth <- success.test$pass
success.confusionmatrix <- table(truth, predictions)
success.confusionmatrix
TPR <- 48/(48 + 3)
FPR <- 18/(18 + 10)
TPR
FPR
```

```{r, message=FALSE, warning=FALSE}
success.predict <- predict(success.glm, success.test, type = "response")
log.pred <- prediction(success.predict, success.test$pass)
success.tree.pred <- predict(success.tree, success.test, type = "where")
tree.pred <- prediction(success.tree.pred, success.test$pass)
log.perf <- performance(log.pred, measure = "tpr", x.measure = "fpr")
tree.perf <- performance(tree.pred, measure = "tpr", x.measure = "fpr")
plot(log.perf, col = "red", lwd = 3, main = "ROC Curves")
plot(tree.perf, col = "blue", lwd = 3, add = TRUE)
```

```{r, echo=TRUE, message=FALSE, warning=FALSE}
log.auc <- performance(log.pred, "auc")@y.values
tree.auc <- performance(tree.pred, "auc")@y.values
log.auc
tree.auc
```
~~~