schools

June 3, 2021

1 Influence of social life and alcohol consumption on academic performance: a cross-sectional study

Rebecca Huang

Isabelle Kim

Anton Kim

1.1 Introduction

Educators, parents, students and policymakers have all been interested in being able to measure, explain and predict good academic performance at all levels of education. Countless papers have been written on the subject pontificating about education issues big and small. Resesearcher have been concerned about various issues: from effects of alcohol use on academic achievement in high school (see, for example, Balsa et al, 2011) to the effects of studying time on academic performance in high school (Stinebrickner and Stinebrickner, circa 2007).

In this paper, we contribute our two cents to the ongoing discussion by building a model that will help us predict academic performance of secondary school students in Portugal.

In particular, we want to predict how the final grade (G3, response variable) of each student depends on the amount of studying the student does (studytime), how diligently the student attends the school (absences), the degree to which he participates in social life (goout) and how much the student drinks on weekend (Walc).

The data we have come from the study by Cortez and Silva (2008) and includes cross-sectional data about secondary school students in two Portuguese schools. In addition to the variables mentioned above, information about a whole host of additional variables have been collected by these researchers, including whether a student plans to attend a university, commute time, parent education and the like. We do not use all these variables in our study as we believe that the most predictive power comes from the variables that we identified above.

1.2 Data

Following best traditions of tidyverse and DSCI 100, we load all necessary libraries and read in the data, cautiously looking only at the first six observations.

[1]:

```
# all the library we need to use
library(tidyverse)
library(tidymodels)
library(GGally)
library(gridExtra)
library(kknn)
```

```
  Attaching packages                          tidyverse
1.3.0

  ggplot2 3.3.3      purrr   0.3.4
  tibble  3.0.4      dplyr   1.0.2
  tidyr   1.1.2      stringr 1.4.0
  readr   1.4.0      forcats 0.5.0

Warning message:
"package ‘tibble’ was built under R version 4.0.3"
Warning message:
"package ‘readr’ was built under R version 4.0.3"
  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()
```

```
Error in library(tidymodels): there is no package called ‘tidymodels’
Traceback:

1. library(tidymodels)
```

[ ]:
```
#read file

temp <- tempfile()
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/00320/
 ↪student.zip",temp)
unzip(temp)
mat <- read.csv("student-mat.csv", sep = ';')
unlink(temp)

head(mat)
```

Slightly dizzy from the sheer abundunce of variables, we judiciously select only those of them that warrant exploration as explained in the introduction.

[ ]:
```
# clean data
mat_selected <- mat %>%
  select(G3, studytime, absences, goout, Walc)
head(mat_selected)
```

The above tamed dataset is much more tidy, minimalistic and yet fully sufficient for the data analysis that we are about to unleash on it in the sections to come.

1.3 Methods and Results

We use K-nearest neighbors regression to answer the question of interest as it is most methodologically appropriate method of analysis due to the fact that G3 is a quantitative variable.

Model Training With an eye toward evaluating how our knn regression model performs with real-world data, we split the dataset into training and testing components.

[ ]:
```
# Splitting into training set and testing set
set.seed(1)
mat_split <- initial_split(mat_selected, prop = 0.75, strata = G3)
mat_train <- training(mat_split)
mat_test <- testing(mat_split)
head(mat_train)
```

Predicting variables in our dataset feature different scales. For example, whereas the number of absences (absences) vary from 0 to 93, the frequency of going out with friends (goout) range from 1 (very infrequent) to 5 (very frequent). To avoid comparing two predictors of different scales we preprocess the data by standardizing all predictors before running our K-Nearest Neighbor regression.

[ ]:
```
# pre-processing data
mat_recipe <- recipe(G3 ~ ., data = mat_train) %>%
  step_scale(all_predictors()) %>%
  step_center(all_predictors())
```

We let the algorithm automatically tune to the optimal number of neighbors to achieve the highest prediction accuracy.

[ ]:
```
# specifying that we want to use knn regression with optimal K

mat_spec <- nearest_neighbor(weight_func = "rectangular", neighbors = tune())␣
 ↪%>%
  set_engine("kknn") %>%
  set_mode("regression")

mat_wkflw <- workflow() %>%
  add_recipe(mat_recipe) %>%
  add_model(mat_spec)
```

We ensure the highest possible model accuracy, we further split our overall training data into several training/validation sets to mitigate unlucky data split issues. Further, we identify the number of neighbors associated with the lowest RMSPE.

[ ]:
```
# Model Accuracy
set.seed(1)
mat_vfold <- vfold_cv(mat_train, v = 5, strata = G3)
```

```
gridvals <- tibble(neighbors = seq(1, 200))

mat_min <- mat_wkflw %>%
  tune_grid(resamples = mat_vfold, grid = gridvals) %>%
  collect_metrics() %>%
  filter(.metric == "rmse") %>%
  filter(mean == min(mean))

mat_min
```

It turns out that the cross-validation RMSPE is 4.398 and the number of neighbors that we should use in our knn regression to achieve this RMSPE is 101.

Test Set Model Evaluation We check how good our trained model is by asking it to predict the values for the testing set that we set aside earlier and calculating RMSPE for the testing set.

[ ]:

```
kmin <- mat_min %>% pull(neighbors)  # kmin = 101
mat_spec <- nearest_neighbor(weight_func = "rectangular", neighbors = kmin) %>%
  set_engine("kknn") %>%
  set_mode("regression")

mat_fit <- workflow() %>%
  add_recipe(mat_recipe) %>%
  add_model(mat_spec) %>%
  fit(data = mat_train)

mat_summary <- mat_fit %>%
  predict(mat_test) %>%
  bind_cols(mat_test) %>%
  metrics(truth = G3, estimate = .pred)
mat_summary
```

Our model's RMSPE is 4.899. At first, this seems not too far off (less than 11% higher) from the cross-validation RMSPE of 4.398 that we obtained in the previous subsection. However, RMSPE is expressed in response variable units, which in our case are the final grade points which range from 0 (lowest grade) to 20 (highest grade). So, roughly speaking, on average, we can expect that the model predicted final grade will be off by 4.88 points from the true final grade. This cannot be considered to be too accurate, considering that the average grade for the testing dataset is 10.22.

[ ]:

```
mean(mat_test$G3)
```

For sanity check, we plot our prediction line for all four predictors variables and alas we confirm our suspicions about poor model fit.

[ ]:

```
# PLOTS
set.seed(1)
mat_preds <- mat_fit %>%
  predict(mat_train) %>%
  bind_cols(mat_train)

plot_final1 <- ggplot(mat_preds, aes(x = studytime, y = G3)) +
  geom_point(alpha = 0.4) +
  xlab("Study Time (hours/week)") +
  ylab("Final Grade") +
  geom_line(data = mat_preds, aes(x = studytime, y = .pred), color = "blue") +
  ggtitle(paste0("K = ", kmin))

plot_final2 <- ggplot(mat_preds, aes(x = absences, y = G3)) +
  geom_point(alpha = 0.4) +
  xlab("Number of Absences") +
  ylab("Final Grade") +
  geom_line(data = mat_preds, aes(x = absences, y = .pred), color = "blue") +
  ggtitle(paste0("K = ", kmin))

plot_final3 <- ggplot(mat_preds, aes(x = goout, y = G3)) +
  geom_point(alpha = 0.4) +
  xlab("Going Out Frequency (low to high)") +
  ylab("Final Grade") +
  geom_line(data = mat_preds, aes(x = goout, y = .pred), color = "blue") +
  ggtitle(paste0("K = ", kmin))

plot_final4 <- ggplot(mat_preds, aes(x = Walc, y = G3)) +
  geom_point(alpha = 0.4) +
  xlab("Weekend Alchohol Consumption (low to high)") +
  ylab("Final Grade") +
  geom_line(data = mat_preds, aes(x = Walc, y = .pred), color = "blue") +
  ggtitle(paste0("K = ", kmin))
```

[ ]:

```
plot_final1
plot_final2
plot_final3
plot_final4
```

The blue line in each plot basically says that if we just look at studying hours, it makes no difference whether one studies for one hour per week versus four hours per week. The final grade does not depend on consistent weekly self-education. We see similar picture with other plots that have weekly alcohol consumption, number of absences, and frequency of going out with friends on horizontal axis.

1.4 Supplementary Analysis: Multiple Regression

To make sure that we leave no stone unturned, in addition to the core analyis above, we provide multiple regression as a supplementary analysis in this section. Following best practices of tidy data analysis, we first split the dataset into training and testing components.

[ ]:

```
set.seed(1)
mat_split <- initial_split(mat_selected, prop = 0.75, strata = G3)
mat_train <- training(mat_split)
mat_test <- testing(mat_split)
head(mat_train)
```

After we have our traning set, we need to create and assign the linear regression model specification and create a recipe for the model beased on traning data.

[ ]:

```
mat_spec <- linear_reg() %>%
  set_engine("lm") %>%
  set_mode("regression")

mat_recipe <- recipe(G3 ~ ., data = mat_train)
```

Then we can fit our simple linear regression model by putting them together in a workflow.

[ ]:

```
mat_wkflw <- workflow() %>%
  add_recipe(mat_recipe) %>%
  add_model(mat_spec) %>%
  fit(data = mat_train)
mat_wkflw
```

Our coefficients are (intercept) b0 = 10.89 and b1 = 0.554, b2 = 0.029, b3 = -0.54, b4 = 0.006.

This means that the equation of the line of best fit is:

G3 = 10.89 + 0.554 * studytime + 0.029 * absences - 0.54 * goout + 0.006 * Walc

Studytime, absences, Walc have the positive realtionship with G3, and goout has negative relationship with G3. Their relationships are mild because their coefficients are small and statistically insignificant. Lack of statistical significance falls outside of the scope of this class, but essentially it confirms the findings of the KNN regression that we performed in the section above.

[ ]:

```
# Calculate the RMSPE for training data to assess goodness of fit

mat_rmspe <- mat_wkflw %>%
  predict(mat_train) %>%
  bind_cols(mat_train) %>%
  metrics(truth = G3, estimate = .pred) %>%
  filter(.metric == 'rmse') %>%
  select(.estimate) %>%
  pull()
mat_rmspe
```

[ ]:

```
# Calculate the RMSPE for testing data to assess goodness of fit

mat_rmspe <- mat_wkflw %>%
  predict(mat_test) %>%
  bind_cols(mat_test) %>%
  metrics(truth = G3, estimate = .pred) %>%
  filter(.metric == 'rmse') %>%
  select(.estimate) %>%
  pull()
mat_rmspe
```

The RSMPE is similar to the RMSPE from the KNN regression in the previous section. More generally, overall results from the multiple regression analysis confirm and support the results done in our core analysis where we used KNN regression.

1.5 Discussion

While it is possible to predict with reasonable accuracy what final grade a Portuguese student will get, somewhat surprisingly, as the RMSPE statistics along with the graphs show the grades do not vary all that much with respect to individual predicting variables.

It is especially surprising that the amount of studying hours per week that a student puts into their education does not seem to affect the final grade. Obviously, this result should be surprising, if discouraging, for educators, parents and students. This is contrary to what we expected to find when we set out to do this study.

That suggests several explanations. The final grade may be sensitive to the amount of time that a student crams for the final exam more than on the variables that we used in our analysis. This can happen if the final grade for Portuguese schools is heavily dependent upon the final exam or the final project, which lend themselves to cramming. Another explanation is that students that reported studying for long hours every week may have exaggerated how much they are studying, perhaps because, they find studying boring and time seems to be dragging if that is the case. Finally, (heavily weighted) final exams in Portuguese schools do not reflected faithfully the material studied during the semester, which explains why it is almost irrelevant how long students study during the week. Incidentally, this explanation is congruent with the fact that an average final grade is slightly higher than 10 (out of 20).

A future set of studies could look at each of the three explanations that we suggest above and by collecting additional data tackle the question as to why the amount of studying per week seems to be a weak predictor of the final grade. If any of our explanations find support in the future studies, perhaps, this could be a motivation for an educational reform in Portuguese schools that can be affected by the schools themselves or the government.

2 References

Balsa, Ana I et al. “The effects of alcohol use on academic achievement in high school.” Economics of education review vol. 30,1 (2011): 1-15.

Cortez P. and Silva A. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Stinebrickner, Todd R. and Stinebrickner, Ralph. “The causal effect of studying on academic performance.” The University of Western Ontario:

https://economics.uwo.ca/people/stinebrickner_docs/paper2.pdf (circa 2007)