Student Performance

Rachel Bellflowers

Last compiled: Sep 11, 2020

Contents

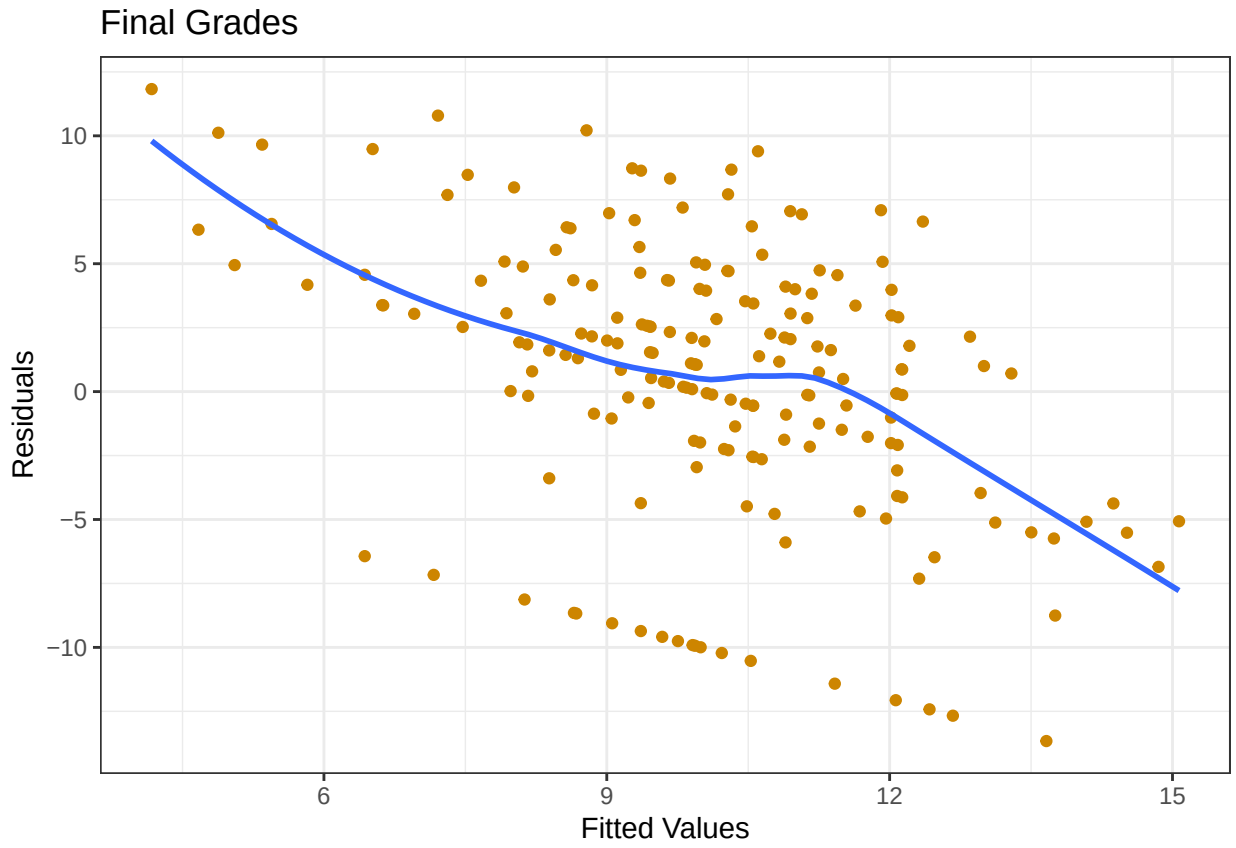
Dataset Description

This dataset examines how different factors affected 382 students' final grades for a math course. I used only the following variables from the dataset:

- **Dependent Variable**
  - `G3`: final grades for the class, ranging from 0-20
- **Independent Variables**
  - `school`: whether they attended Gabriel Pereira or Mousinho da Silveira
  - `sex`: female or male
  - `age`: ranged from 15-22
  - `address`: whether they lived in a rural or urban area
  - `famsize`: whether their family had less than or equal to 3 members or greater than 3
  - `Pstatus`: parents separated or living together
  - `nursery`: whether they attended nursery school
  - `internet`: if they had Internet access at home
  - `guardian`: if they are taken care of by a mother, father, both, or other
  - `studytime`: How many hours they studied a week
  - `famsup`: if their family supported their education
  - `paid`: if they were taking extra paid classes within the course subject (math)
  - `activities`: if they participated in extracurricular activities
  - `higher`: whether they intend to seek higher education
  - `romantic`: in a romantic relationship

Linear Model

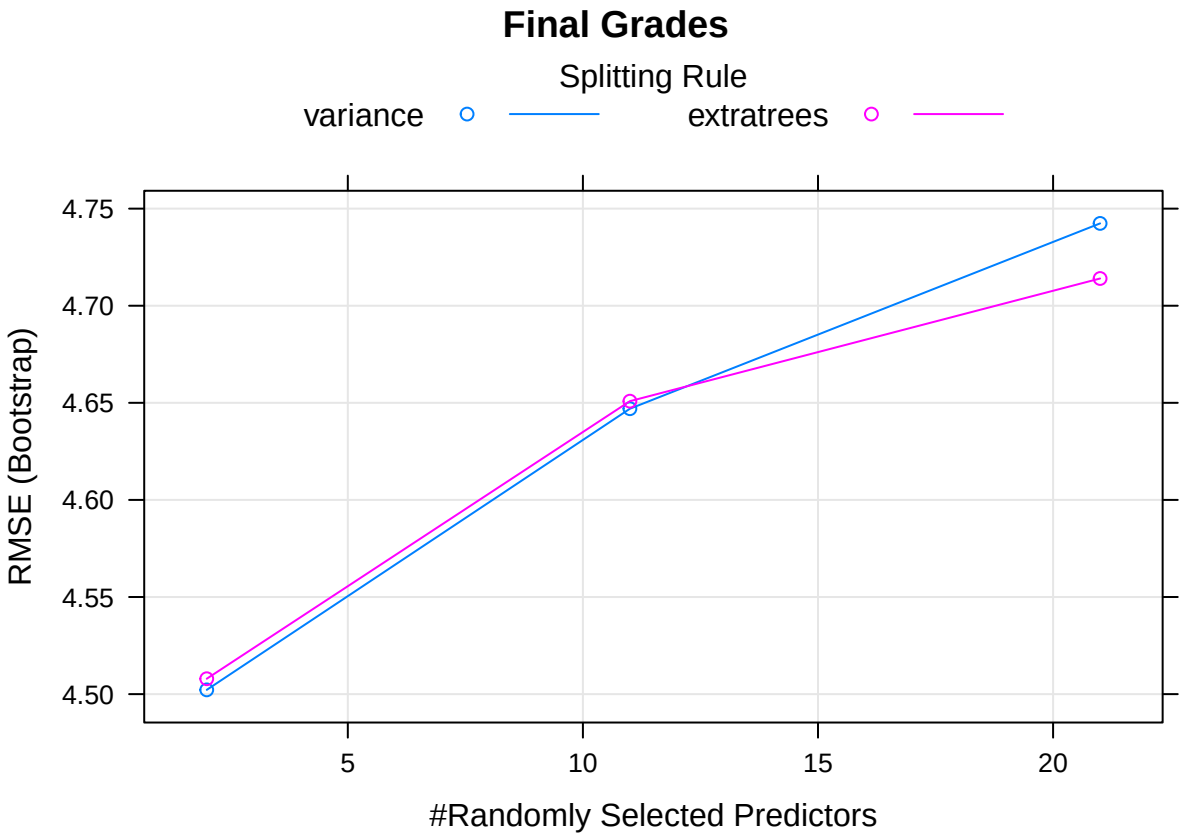
First, I created a simple linear model using a training/test split of 50/50. Given that most the curve is centered on "0" for the residuals, the linear model appears to be a good predictor for final course grades.



The plot in Figure **??** was created with `ggplot2` (Wickham et al. 2020).

Random Forest

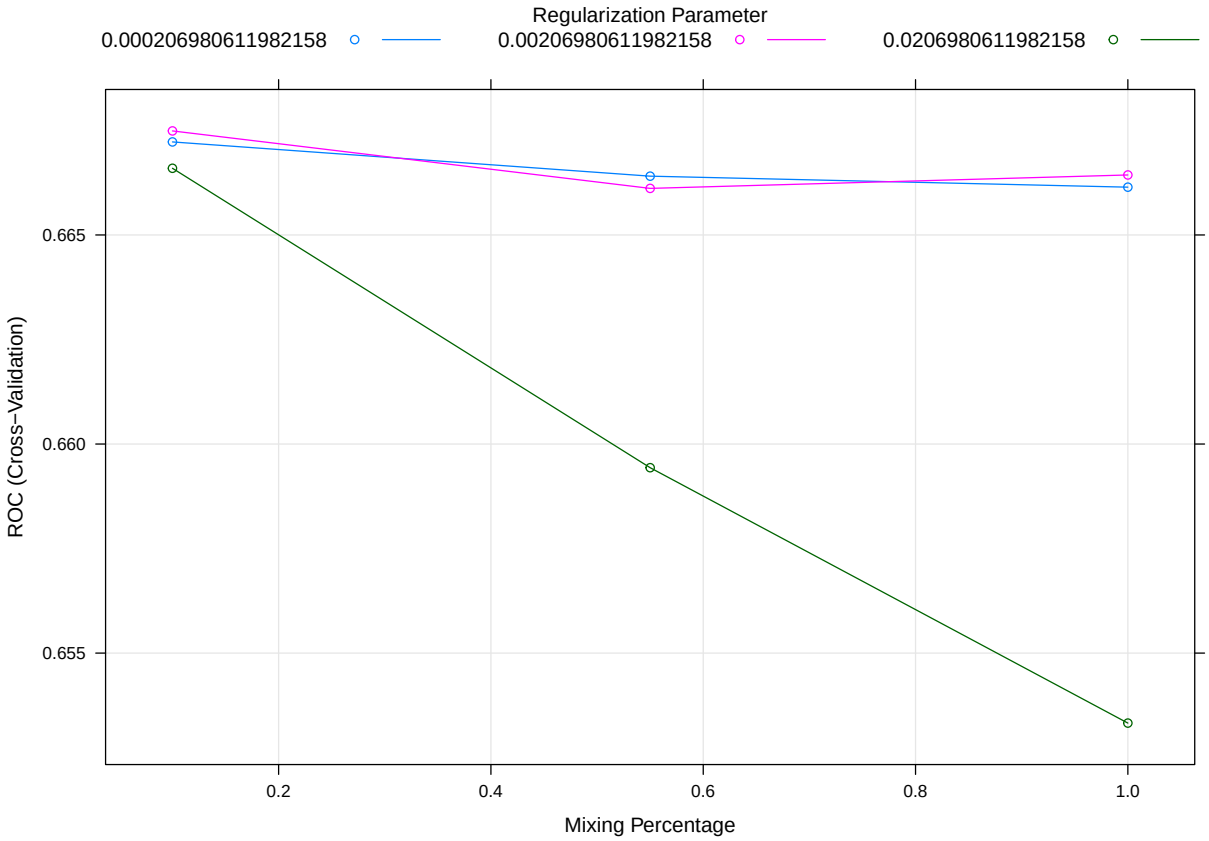
I had doubts concerning the results of my initial linear model given that I used a 50/50 split. Therefore, I decided to run a random forest model. According to the graph, the less randomly selected predictors, the lower the RMSE. As the grading scale ranged from 1-20, a RMSE ranging from around 4.4 to around 4.7 suggests that this model does not serve as a good predictor of final course grades.



The plot in Figure **??** was created with base R (R Core Team 2018).

Classification Model

To fit a classification model, I divided up grades into two categories: grades ranging from 1-10 and grades ranging from 11-20. I then used `method = glm` in the `train()` function. The green line, which has the highest lambda value, on average has a smaller ROC value than either the blue or pink lines.



The plot in Figure **??** was created with base R (R Core Team 2018).

Conclusion

Although the first graph appeared to be the best out of the three, I believe that none of these models were good predictors of the outcome variable. The package suggested for creating ROC curves in the Datacamp exercise was not compatible with the version of RStudio Server we are using, so I am not sure if my last graph was calculated correctly. Additionally, I believe creating a 50/50 split for the first graph perhaps resulted in overfitting. In the future, a larger sample size with a greater age range would probably help with finding a greater effect.

References

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.