Project: Socioeconomic factors that explain variation in exam scores

Georgina Wager

21/11/2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

This project is part of the assessment for the module C7081 Statistical Analysis for Data Science, that is ran at Harper Adams University College by Ed Harris for the MSc Course: Data Science for Global Agriculture, Food and Environment.

For this analysis secondary data was collected from Kaggle, a website that contains many data sets to be analysed for learning purposes. The exact source of the data was as follows: https://www.kaggle.com/spscientist/students-performance-in-exams (https://www.kaggle.com/spscientist/students-performance-in-exams). The raw data has been downloaded and is fully accessible from the authors git hub. The data contained information for 1000 students yet information on how this data was collected is unavailable.

GitHub Name: georginaanna GitHub Link: https://github.com/georginaanna/C7081-Assessment (https://github.com/georginaanna/C7081-Assessment) GitHub Contents are as follows: an Excel format of the data with a data dictionary, an R Script of the Explanatory Data Analysis (EDA) with contents and comments and an R Script of the Overall Analysis with contents and comments.

Project Outline

This project is proposed to use socio-economic factors, gender, and ethnicity to predict exam scores for each student whilst determining the overall factors that influence exam score.

Understanding the relationship between socio-economic factors and exam outcome can provide schools with the ability to allocate the correct resources to students who need it most. This is a challenging prospect as an individual's performance can be due to their current environment and not solely down to socio-economic variables. Research has supported that education is a vital part of society and its government but also to shape an individual's identity (Idris et al., 2012), therefore ensuring students succeed in an their educational setting could produce a stronger society, and widen that individuals prospects through life. Overtime UK education systems have done little to maximise resources directly for those who are deemed socially unequal in society (McLoyd, 1998), they have also indirectly avoided understanding how parents attitudes could improve a child's willingness to learn (Phillips et al., 1998).

Brynner and Joshi (2002) documented that students in a high socioeconomic position performed better in exams than those in a low socioeconomic position. It is thought that 35% of children who are eligible for free school meals (FSM), attained 5 or more A*–C grades (including English and Maths) compared to 62% of non-FSM children (Release, 2012). Social differences in attainment not only exist but are greater than gender or ethnicity differences; around six times larger than gender differences and three times larger than ethnicity differences (Strand, 2011). A study carried out by Hijazi and Naqvi (2006) identified that mother's education and student's family income were highly correlated with the student academic performance. Since much

research has been carried out in this area using Data mining (DM) being applied to education an interdisciplinary field has emerged this is Educational Data Mining (EDM) (Satyanarayana and Nuckowski, 2016).

Objectives

1. To find a model that accurately predict student exam scores based on socioeconomic information.
2. To state if a significant relationship exists between the independent variables and student exam score.

Method

The secondary data was downloaded as an excel file before further processing and handling in R (A language and environment for statistical computing (R Core Team, 2020)). General packages were installed this included the installation of ggplot, dplyr, random forest, caTools, miscTools, caret, lme4, rmarkdown, broom and shiny.

The data had variables that were classed as characters and these were converted to factors. The variable “Gender” was the students sex, and is represented in the data as a factor with 2 levels (Male or Female), followed by the variable “Ethnicity” which was the students ethnicity as a factor with 5 levels (Group A, Group B, Group C, Group D, and Group E). Next the socioeconomic variables were “Parental Education” that was students parental education as a factor with 6 levels (bachelor’s degree, some college, master’s degree, associate’s degree, some high school, high school) where by each level has been ranked based on the level of qualifications needed with some high school being “1”, high school being “2”, some college being “3”, associate’s degree being “ 4” , bachelor’s degree being “5” and a master’s degree being “6”. The next socioeconomic variable was “Lunch Cost” this was the cost of lunch for each student as a factor with 2 levels (Standard, free). The final socioeconomic variable was the “Exam Preparation Course” which was the student’s participation of an exam preparation course as a factor with 2 levels for whether the student completed the exam preparation course (yes/no). Next the final variables are the dependent variables that were numeric, these are the students marks obtained in each exam, marked as 1 – 100, the variable names are “Math exams score”, “Reading exam score” and “Writing exam score”.

Prior to analysis a principal component analysis (PCA) was carried out to reduce the dimensionality of the data set through minimizing variables, while still containing the information that characterizes that variable. For the analysis, the three dependent variables, maths exam score, reading exam score and writing exams score were used for Principal Component Analysis. The three dependent variables were highly correlated with an r squared between 0.8 and 0.9. In carrying out a principal component analysis the eigenvectors and eigenvalues are computed from the co-variance matrix. This provides a new principal component analysis that is instructed as an additional variable in the data set. The aim is to have most of the data within the first principal component analysis (PC1). The principal component one will contain the highest percentage of explained variance, this is the new dependent variable for further analysis (James et al., 2013). For the PC1 to be usable we are aiming for a value above 70% this means that the PC1 represents 70% or more data within those variables.

A multiple linear regression analysis was fitted as it enables one dependent variable and multiple independent variables to be used within a model.The multiple linear regression model will also provide answers to the objectives through the ability to assess if a group of independent variables can predict an outcome and identifying the variables that contribute most to the dependent outcome (Chamber, 1992).The linear model was put together using the “lm()” function, for each model the PC1 is the response variable, this is followed by a tilde(~), and then the remaining independent variables. In addition, the data was split into a train and test data set, this was carried out to be able to test the model accuracy on data the model was not initially trained on. Using the new dependent variable (PC1) in the data set a linear model was used to identify the relationship between the independent variables and PC1.

The linear regression models (supervised learning method) underwent multiple ANOVA, to examine the contribution of all the independent variables; gender, ethnicity, lunch price, parental level of education and exam preparation course to the variation in the dependent variable PC1. For each model, the Akaike’s (An

Information Criterion (AIC)) was calculated to test the quality of the statistical model. The model with the lowest AIC is deemed to be the better model as it is a better fit to the data set and avoids over fitting and maximizes the models generalizability (Sakamoto et al., 1986).On deciding the best model for predictions it was then evaluated evaluated using the Root Mean Squared (RMSE) (Witten and Frank 2005).

Random forest model (supervised learning method) was carried out to maximize the predictive accuracy and robustness of predictions compared to a multiple linear regression model.A random forest model is built from a randomly selected sample of the data and a random collection of features for node splitting before averaging the results from multiple trees to produce a prediction.Random Forest was used to lower the variance as this method takes a set of independent variables each with its own variance, then averages them over the total number of observations, this lowers the variance, leading to greater prediction accuracy (Breiman 2001).

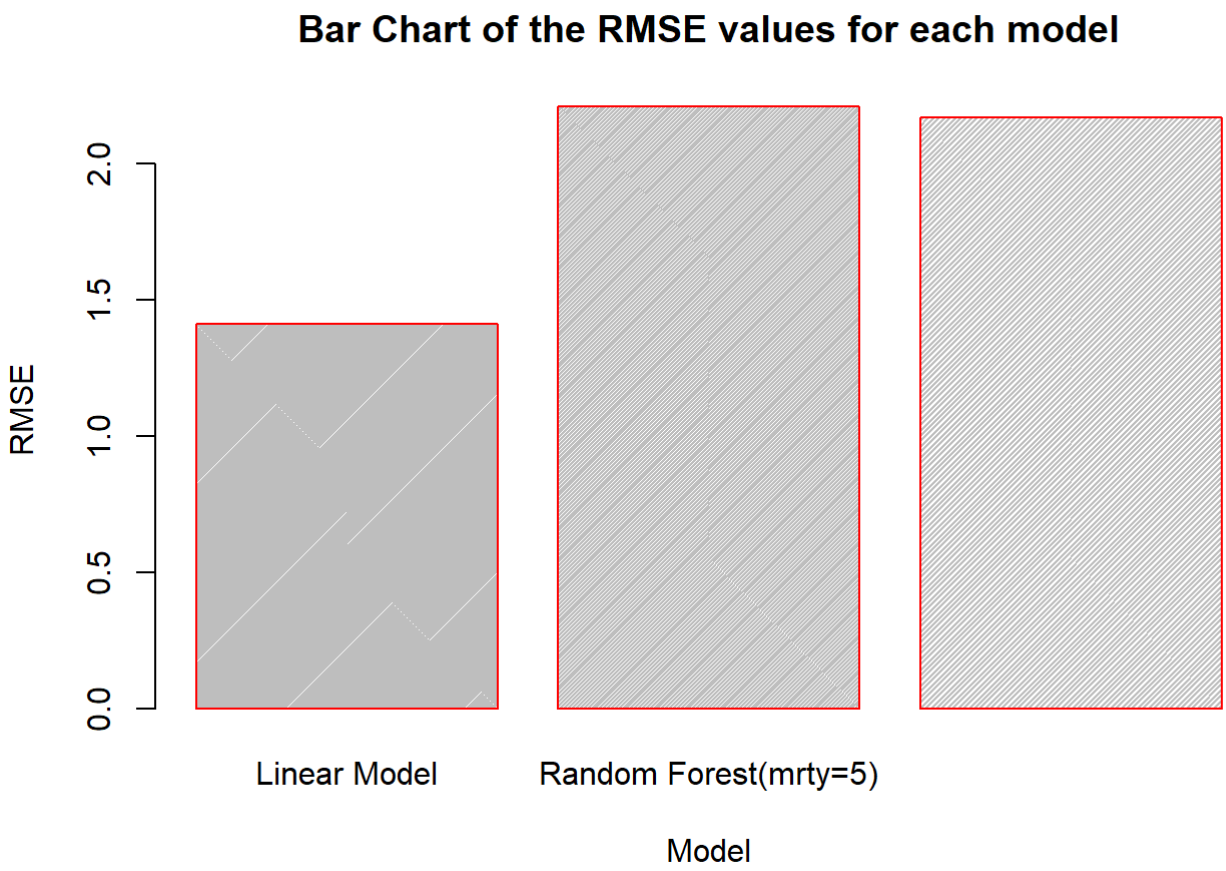
For the random forest model, the three dependent variables were amalgamated into a single exam score labeled “total score” before being converted into a categorical variable. This categorical variable was converted into a factor before being converted into a numeric class. For the models the total score and PC1 was removed because these dependent variables, would take up variable importance.The first model was produced using all 6 variables (mtry=6): Category, Ethnicity, Gender, Lunch Price, Exam Preparation Course and Parental Level of Education, the second model altered the mtry through (p/3) leading to mtry=2.

The for each of these models the variable importance was analyzed and plotted. The variable importance provides the % increase in the MSE. Increase in node purity is the decrease in node impurities from splitting on that specific variable.

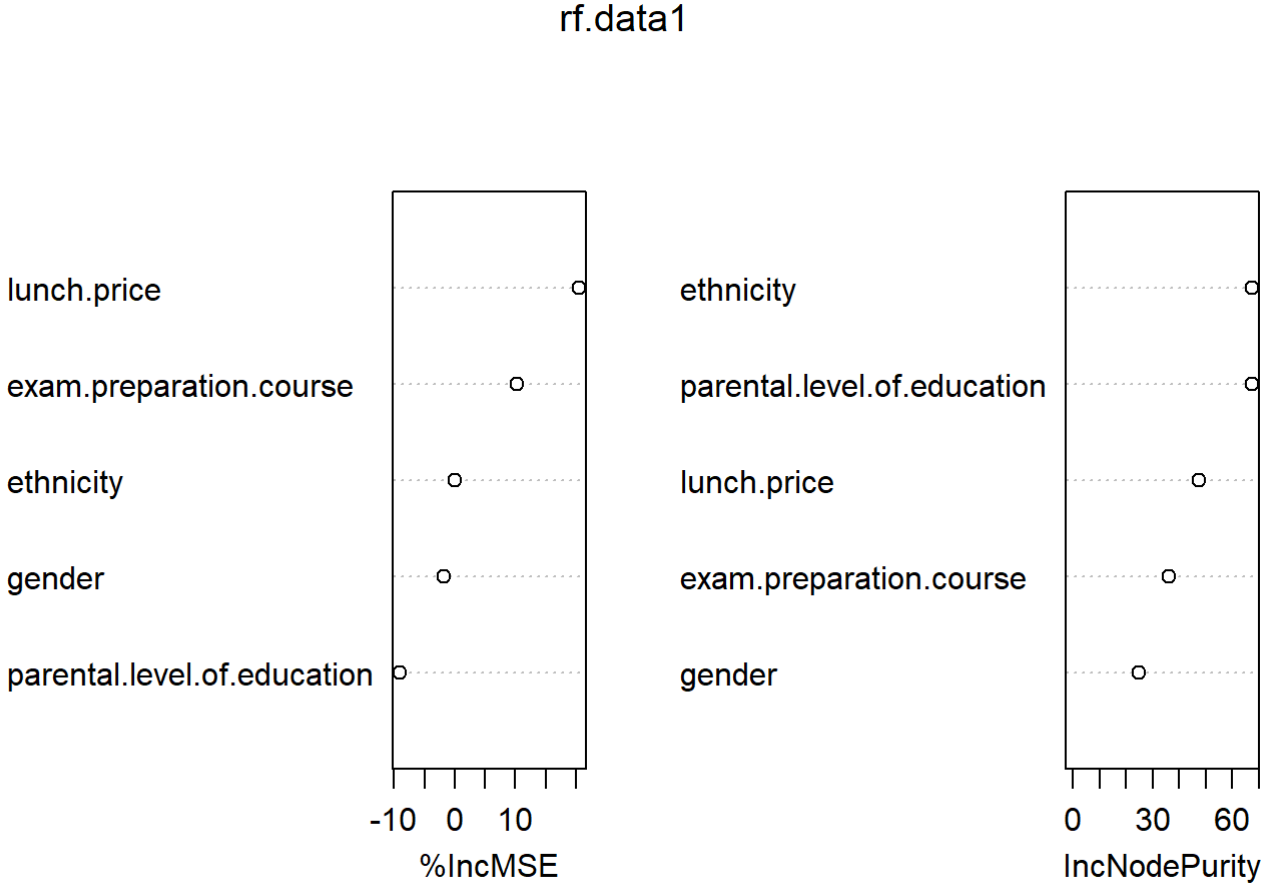
Results

1. To find a model that is best at predicting student exam scores based on socioeconomic information.

The PC1 obtained was 90.61% for proportion of variance demonstrating that more than 90% of the data is in the dependent variable “PC1” that was used for the multiple linear regression models. The multiple linear regression model chosen had an AIC of 2886.7, upon predictions an MSE (mean squared error) of 1.98, an RMSE (root mean squared error) of 1.40, in comparison to the random forest model that yielded an RMSE of 2.21 with all predictors and the random forest model with 2 predictors that yielded an RMSE of 2.21. The rsquared value for the linear regression model was 0.23, suggesting that the sum of square residuals is 77%.



Variable importance was calculated for the random forest model (mrty=2) showing lunch price and exam preparation course with the highest % increase in MSE with 20.5% and 10.27% respectively.Parental level of education was -9.05, despite this variable being correlated with total score when this variable is numeric and not grouped.



2. To state if a significant relationship exists between the independent variables and student exam score.

The anova from the linear model with the lowest AIC demonstrated that parental level of education significantly affected exam score ($p < 0.001$) along with gender ($p<0.001$), ethnicity ($p<0.001$), participation in exam preparation ($p<0.001$) and lunch price ($p<0.001$).

Conclusion

Too conclude the linear model had the lowest RMSE, however an r squared value of 0.23, this suggests an sum of squares residual error of 77%, concluding that despite the better RMSE value, this model is not the best fit for the data.In addition to this it also shows that there are other variables that aren't included in the data that have a have a bigger effect/ contribution to the dependent variable more than the variables in the data currently used in the analysis. A random forest was carried out however when the variable importance of each variable was calculated the variable importance for parental level of education was negative, from EDA a correlation does exist between parental level of education and exam score therefore we would conclude that some objects within the data have similar predictors leading to very different outcomes suggesting this model is not suited to predicting exam score.Overall, neither model provided a good fit for the data, but at of all of the models the multiple linear model with the use of principal component analysis seemed the better one.

References

Breiman L., 2001. Random Forests. Machine Learning, 45, no. 1, 5–32.

Bynner, J. and Joshi, H., 2002. Equality and opportunity in education: Evidence from the 1958 and 1970 birth cohort studies. Oxford Review of Education, 28(4), pp.405-425.

Chambers, J. M. (1992) Linear models. Chapter 4 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hijaz, S.T. and Naqvi, S.R., 2006. Factors affecting students' performance: A case of private colleges in Bangladesh. Journal of sociology, 3(1), pp.44-45.[Accessed 21 November 2020].

Idris, F., Hassan, Z., Ya'acob, A., Gill, S.K. and Awal, N.A.M., 2012. The role of education in shaping youth's national identity. Procedia-Social and Behavioral Sciences, 59, pp.443-450.[Accessed 21 November 2020].

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

McLoyd, V.C., 1998. Socioeconomic disadvantage and child development. American psychologist, 53(2), p.185.[Accessed 21 November 2020].

Phillips, M., Brooks-Gunn, J., Duncan, G.J., Klebanov, P. and Crane, J., 1998. Family background, parenting practices, and the Black–White test score gap.

R Core Team. 2020. The R Project for Statistical Computing. [Online]. The R Foundation. Available from: https://www.r-project.org/ (https://www.r-project.org/) [Accessed 21 November 2020].

Release, S.F., 2012. GCSE and equivalent attainment by pupil characteristics in England, 2010/11.[Accessed 21 November 2020].

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. 1986. Akaike information criterion statistics. Tokyo: KTK Scientific Publ.[Accessed 21 November 2020].

Witten I. and Frank E., 2005. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco, CA.