Modelling education: quantifying the effect of certain factors on school performance and graduate school admission

Rafid Akhyara Agung
200614920

Table of Contents

1. Background

As individuals, education can be one of the most important aspects of our lives. People who are educated tend to have more opportunities, earn more (Walker and Zhu, 2003) and even tend to be healthier (Cutler and Lleras-Muney, 2006). It is therefore of great interest to most people to focus on doing well on education and get those coveted 'good grades' so that they may open new doors for themselves. More and more people are willing to spend extra time in education—for instance, to obtain master's degrees (Snyder et al, 2016)—just because of the value that additional education and degrees can confer. Master's degree holders can earn $1,500 per week compared to around $1,300 per week for bachelor's degree holders. (U.S. Bureau of Labor Statistics, 2022)

This is why people are interested in knowing: what can one do to maximize their performance in school? And also: what factors will matter to maximize their chances of being accepted into graduate school? We can find out this information by creating models that allow us to quantify the effects of certain factors on school performance and graduate school admission, and not only that, but also allow us to predict an individual's potential school performance/chance of admission.

2. Data

In this analysis, we will look at data of performance in Portuguese secondary schools provided by Cortez and Silva (2008) (obtained here: https://archive.ics.uci.edu/ml/datasets/student+performance) and data on graduate admission outcomes (obtained here: https://stats.idre.ucla.edu/stat/data/binary.csv).

The data on graduate school admissions was collected from a sample of applicants to graduate school at UCLA. It contains data on the applicants' academic performance and the prestige of the school they went to in their undergraduate studies. There are 4 variables, our target variable:
admit - Whether they got admitted or not (Binary: 1 - admitted, 0 - not admitted)

And the 3 independent variables:
gpa - GPA of applicant (Numeric: from 0 to 4)
gre - GRE Score of applicant (Numeric: from 0 to 340)
rank - Prestige of undergraduate institution (Numeric: from 1 - most prestigious to 4 - least prestigious)

The data on secondary school performance in Portuguese schools looked at student achievement in Math in two Portuguese secondary schools. The data includes student grades, their demographic information, as well as social and school related features. The data was collected through information available in school reports, as well as questionnaires. It has 32 variables, our target variable:

G3 - Final grade (Numeric: from 0 - lowest possible grade to 20 - highest possible grade)

And the 31 independent variables, the complete list can be seen in the appendix below. Here only the 5 variables that ended up being used in the final model are shown:
G1 - First period grade (Numeric: from 0 to 20)
G2 - Second period grade (Numeric: from 0 to 20)
age - Student's age (Numeric: from 15 to 22)
activities - Extra-curricular activities (Binary: yes or no)
absences - Number of school absences (Numeric: from 0 to 93)

3. Methodology: Data Processing and Analysis

First, the data was cleaned. For the graduate admissions dataset, since the data was already quite tidy, there was not much cleaning needed. For the student performance dataset, most of the cleaning involved wrangling some numeric columns into the factor data type instead, as some were actually just categorical but being represented by a number. The wrangled variables were:

- *Medu*
- *Fedu*
- *traveltime*
- *studytime*
- *famrel*
- *goout*
- *Dalc*
- *Walc*
- *health*
- *freetime*

Then, before training on any of the datasets, the data was first split into a train set and a test set. Here, an 80-20 split was used, 80% of the data was the train set, 20% the test set. The train set was the part of the dataset used for training the algorithm and deciding on any parameters that needed to be tuned. The test set was used solely for evaluation at the very end. The reason used this split was done was to see how well the models would generalize to unseen data.

For the graduate admissions dataset, since the outcome is binary in nature, it means that the algorithm most appropriately suited for it would be a classification algorithm. The two classification algorithms that have been used are logistic regression and k-nearest neighbors (KNN).

Logistic regression is a linear binary classifier where a linear model is comprised of variables paired with a corresponding coefficient, and then those coefficients are estimated. The model is used to calculate the probability

of a given observation being in the positive class (i.e. being a '1') or in the negative class (i.e. being a '0'). From those coefficients, an odds ratio can be calculated, which quantifies the effect on the odds of an event when we change that variable. The odds of an event are calculated as the ratio of a probability and its complement: **p/(1-p)**. Hence, from the odds, we can obtain the probability.

K-nearest neighbors works in a different way. It takes a given observation and looks at the k nearest points around it. It then sees what class the majority of the points around it are and uses that to make a prediction. To perform KNN, the k parameter needs to be tuned, which affects how many nearest points will be considered when making a prediction. These parameters were decided through 5-fold cross-validation, where 5 different mutually exclusive validation sets were sampled from the train set and the average accuracy was counted from each value of the parameter on those 5 validation sets. The parameter that yielded highest accuracy was chosen for use in the final model.

The reason why two algorithms were used was to see whether a non-linear classifier would perform better than a linear one.

However, before training those algorithms on the data, principal component analysis was first conducted on the dataset. Principal component analysis is a method of dimension reduction that allows us to reduce the number of variables and can even 'group variables together'. It does this by making principal components which are linear combinations of the existing variables. We can see how much variance of the original data each principal component explains and choose the number of principal components from the total variance we want to be explained. PCA was done mainly to see if *gre* and *gpa* could be combined into one variable, since they both measure a similar attribute: academic performance.

For the student performance dataset, since the grade students receive is a continuous variable from 0 to 20, regression algorithms would be most suited. The two regression algorithms that were used were linear regression and regression trees. Again, two algorithms were used to see if a non-linear model would be able to perform better than a linear one.

Linear regression is similar to logistic regression, but whereas logistic regression results in a probability/prediction of the class, linear regression results in a prediction of the target variable's value, which in this case should be a number between 0 to 20. When deciding on the coefficients to use, linear regression will attempt to choose coefficients that minimize the residual sum of squares, a measure of error. Due to the large number of variables, some variables needed to be removed to keep the model simpler and prevent overfitting and allow the model to generalize better. Here, only the statistically significant variables (p-value < 0.05) were chosen.

A regression tree uses the variables—here we use variables from the final linear regression model—to create nodes. At these nodes there is a binary split. The value of the variable that is used to create this split is determined by the value that minimizes the error. The following nodes are either terminal nodes where a prediction is made or further internal nodes where more splits occur. The process of creating a regression tree starts with growing the tree fully and then pruning it so that it is simpler and is less prone to overfitting. The pruning criteria used here is from Breiman et al (1984), where the simplest tree is taken whose error is within 1 standard error of the minimum error tree.

4. Results

4.1 Classification

4.1.1 Principal component analysis

The results of the principal component analysis are shown below:

Principal Component Loadings

	PC1	PC2	PC3
gre	0.6882896	-0.1165073	-0.7160192
gpa	0.6645076	-0.2946878	0.6867232
rank	-0.2910104	-0.9484646	-0.1254105

Here, PC1 is highly correlated with *gre* and *gpa*, so it might be fitting to call PC1 ‘Academics’.

PC2 is highly correlated with *rank*, oso calling it ‘Rank’ is probably fitting. Do note that since the coefficient for rank is negative, it will be the inverse of the original column, meaning a low rank value represents a less prestigious undergraduate institution.

Total Variance Explained

No. of PCs	Total Variance Explained
1	0.4034067
2	0.7359301
3	1.0000000

Here, just 2 PCs are able to explain 73.6% of the variance in the dataset, so only two will be used from here on out.

Using the principal component loadings above as weights, scores are calculated using the value of *gpa*, *gre*, and *rank* in each observation, giving us the columns PC1_Academics and PC2_Rank

4.1.2 Logistic regression

The final logistic regression model contained the variables:

admit ~ PC1_Academics + PC2_Rank

The model had an accuracy of 68.75%. The estimated coefficients are listed below:

Estimated Logistic Regression Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9159690	0.1322996	-6.923443	0.0000000
PC1_Academics	0.5155550	0.1154571	4.465337	0.0000080
PC2_Rank	0.4154507	0.1328960	3.126135	0.0017712

However, for logistic regression we cannot directly interpret the coefficients. Euler's number (*e*) must be raised to the power of each coefficient first to obtain an odds ratio.

Odds Ratios for Each Factor

Coefficients	Odds Ratio
PC1_Academics	1.674568
PC2_Rank	1.515054

From the odds ratio, it shows that with a unit increase in *PC1_Academics*, the odds of being admitted are multiplied by around 1.675. And for a unit increase in *PC2_Rank*, the odds of being admitted are multiplied by around 1.515.

4.1.3 K-nearest neighbors

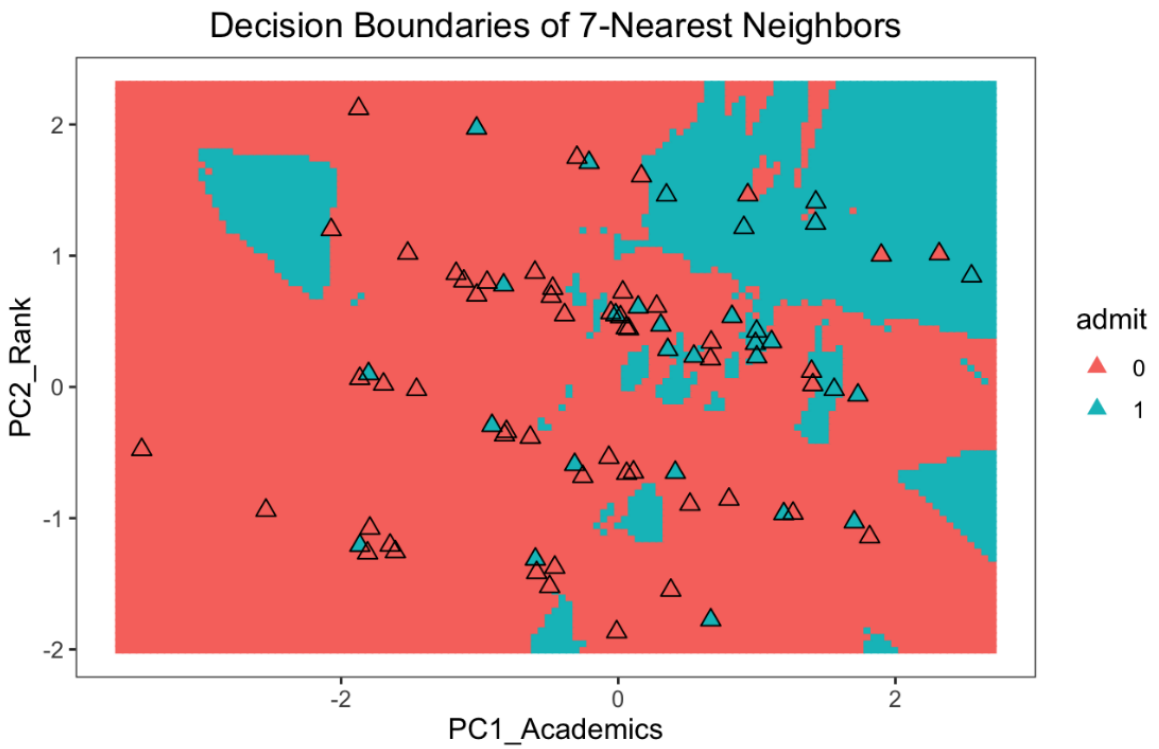
The cross validation results for the k-nearest neighbors model is shown below:

KNN CV Results

k	Accuracy
1	0.6247825
2	0.6219521
3	0.6184829
4	0.6715690
5	0.6434394
6	0.6589698
7	0.6839232
8	0.6747901
9	0.6811393
10	0.6714698

The model with the highest accuracy is the model with 7 neighbors. So the value of k was set to 7.

The decision boundary of 7-nearest neighbors is shown below:



The accuracy of the model was 68.75%. While similar to the logistic regression model the KNN model achieves this accuracy in a different way. It becomes more apparent when we look at the confusion matrices for each model.

LogReg CM

	0	1
0	34	8
1	17	21

KNN CM

	0	1
0	46	20
1	5	9

LogReg Spec. and Sens.

Sensitivity	0.6666667
Specificity	0.7241379

KNN Spec. and Sens.

Sensitivity	0.9019608
Specificity	0.3103448

The logistic regression model has better sensitivity and specificity balance compared to KNN's. This means that the accuracy of KNN is 'less valid' as it achieves this accuracy by predicting most observations as admit = 0 and is only able to achieve a decent accuracy simply due to the fact that around 60% of the data happens to have an admit of 0.

4.2 Regression

4.2.1 Linear regression

After dropping the insignificant variables, the final linear regression model contained the following variables:

G3 ~ age + activities + absences + G1 + G2

The model had a root mean squared error (RMSE) of 2.058844. The estimated coefficients of the model are listed below:

Estimated Linear Regression Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.6371475	1.5139311	1.741920	0.0825142
age	-0.2661717	0.0869121	-3.062539	0.0023871
activitiesyes	-0.4810731	0.2113669	-2.276010	0.0235273
absences	0.0506299	0.0142569	3.551259	0.0004429
G1	0.1869452	0.0592083	3.157416	0.0017484
G2	0.9473820	0.0525661	18.022683	0.0000000

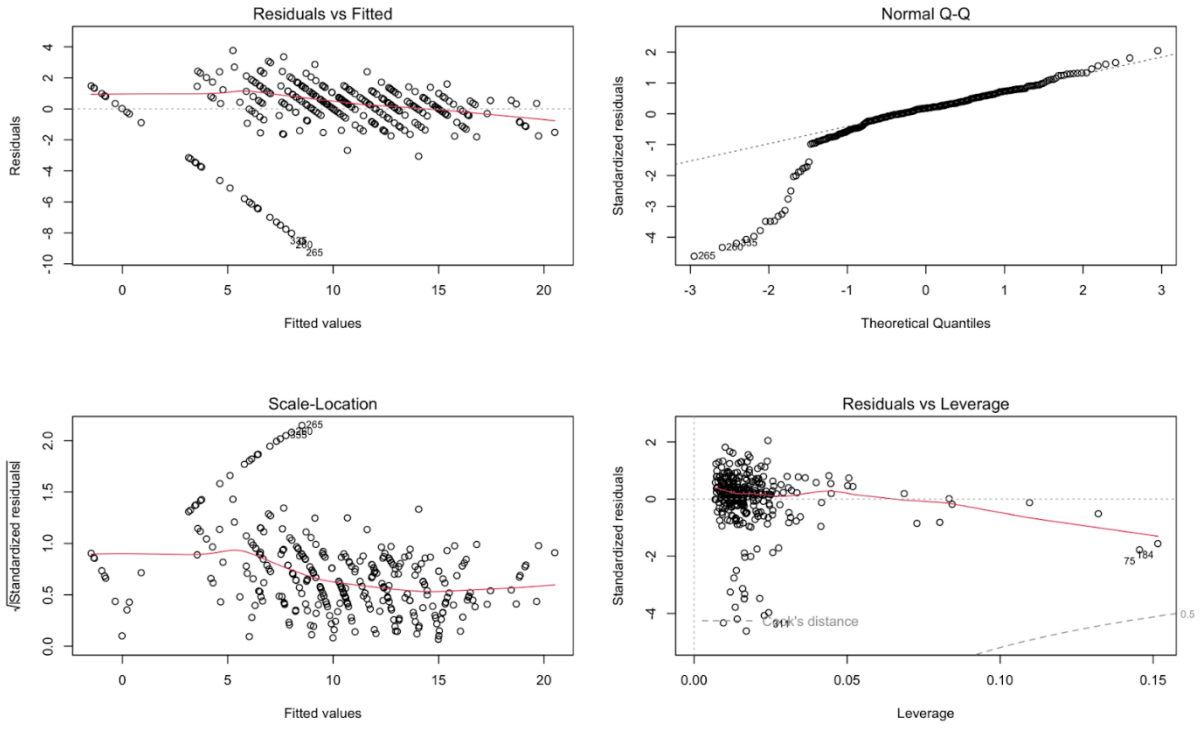
From here, we can see how the grade of a student changes with the factors. For instance, for *age*, for every increase in age, there seems to be a 0.266 decrease in final grades on average.

activities seems to negatively impact the grades of the student as well, with grades going down by 0.481 on average if the student partakes in activities.

For *absences*, it seems that those who are absent more often will have better grades, as their grades increase by 0.051 on average for each absence.

Finally, there are *G1* and *G2*. *G1* positively affects grades with final grades increasing by 0.187 for every unit increase of *G1*. *G2* positively affects grades with final grades increasing by 0.947 for every unit increase of *G2*

Then after conducting the linear regression itself, the four plots below were created to test whether it was appropriate to conduct linear regression on the data.



From the top left plot, we can see that the assumption of that the relationship between the target variable and explanatory variables is linear holds. The top right plot shows that the errors have a mostly normal distribution. The bottom left plot shows that the errors are independent of the explanatory variables and variance constant. The bottom right plot shows that none of the data points are influential outliers. Hence, from these diagnostic plots, it shows that the use of linear regression was appropriate.

4.2.2 Regression tree

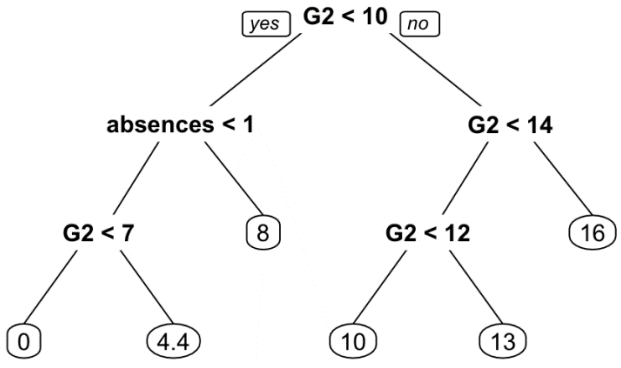
After growing the tree we obtain the following table showing the errors of the regression tree at differing levels of complexity:

Regression Tree Errors at Differing Complexities

CP	nsplit	rel error	xerror	xstd
0.5404760	0	1.0000000	1.0127548	0.0886714
0.1329240	1	0.4595240	0.4750642	0.0406622
0.1198958	2	0.3266001	0.4387440	0.0437819
0.0312017	3	0.2067043	0.2498161	0.0364127
0.0214236	4	0.1755026	0.2274169	0.0345701
0.0159883	5	0.1540790	0.1795961	0.0263531
0.0158777	6	0.1380907	0.1821058	0.0274257
0.0120331	7	0.1222130	0.1731303	0.0284504
0.0100000	8	0.1101798	0.1626427	0.0282540

The more splits the more complex the tree. Here, the tree with the lowest error is the tree with nsplit = 8 with an error of 0.163. The simplest tree with an error within 1 standard error of that is the tree with nsplit = 5.

The tree is shown below:



The RMSE of this model is 2.136505. Here, generally, as *G2* increases, the final grade of a student is predicted to increase.

The linear regression model performs better based on the RMSE value as it is lower so it is the preferred model here. This is not surprising as the diagnostic plots for the linear regression model shown earlier show that there

aren't any non-linear relationships. Hence, though regression trees have the advantage of being able to capture non-linear relationships, since there are none to begin with that advantage becomes moot.

5. Conclusions

For graduate school admission, it seems that one needs to perform both academically well and come from a reputable institution in order to maximize chances of being admitted. This is shown by the similar magnitudes of the odds ratios of both academics and rank. (Do note that we can make this comparison of magnitudes as PCA has scaled the variables) as well as the decision region of the KNN model, where the region for 'admit' is predominantly in the top right side, where academic performance and rank are high.

For further work on this dataset, getting more attributes would be quite important. There are many more things that go into a graduate school application than just grades and school reputation. Students provide letters of recommendation or showcase their research experience, and these and more, are aspects that could be used as variables to supplement the existing ones.

For school performance, one important thing to note is that it mostly shows correlation not causation. If we were to take the model at face value, it would seem that we should skip all days of school because more *absences* seems to be a positive thing, when intuitively that does not make much sense. It does not make the model necessarily wrong but it just means we cannot make such direct interpretations, so one should always maintain healthy skepticism. Now, from the results themselves, it seems that the best way to know how well someone in school is going to perform is from their past performance, especially more recent performance. *G2* has a much larger effect than *G1* in the linear regression model and *G2* is the predominant variable used in the decision tree.

Since this is quite an uninteresting finding considering past performance is not something one can control (it is something that is determined by one's 'talent'), getting a dataset with similar attributes but much more observations would be a great improvement. The problem with the linear regression model right now is that many of the other variables are not statistically significant due to having large p-values. Increasing the number of observations would help reduce those p-values and may result in more variables being included in the final model. Hopefully, those variables are factors that are actionable.

6. References

Walker, I., & Zhu, Y. (2003). Education, earnings and productivity: Recent UK evidence. *Labour Market Trends*.

Cutler, D., & Lleras-Muney, A. (2006). Education and health: Evaluating theories and evidence. *NBER Working Paper Series*. https://doi.org/10.3386/w12352

Snyder, T. D., Brey, de C., & Dillow, S. A. (2016). *Digest of Education Statistics 2016*. Claitor's Publishing Division.

U.S. Bureau of Labor Statistics. (2022, September 8). *Education pays*. U.S. Bureau of Labor Statistics. Retrieved February 27, 2023, from https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

Cortez, P., & Silva, A. (2008). *15Th European Concurrent Engineering Conference 2008, Ecec '2008 and 5th Future Business Technology Conference, Fubutec '2008: April 9-11, 2008, Porto, Portugal*. EUROSIS-ETI.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. CRC.

7. Appendix

Complete list of variables for student performance dataset from Cortez and Silva (2008):

G3 - Final grade (Numeric: from 0 - lowest possible grade to 20 - highest possible grade)
G1 - First period grade (Numeric: from 0 to 20)
G2 - Second period grade (Numeric: from 0 to 20)
school - Student's school (Binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex - Student's sex (Binary: 'F' - female or 'M' - male)
age - Student's age (Numeric: from 15 to 22)
address - Student's home address type (Binary: 'U' - urban or 'R' - rural)
famsize - Family size (Binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus - Parent's cohabitation status (Binary: 'T' - living together or 'A' - apart)
Medu - Mother's education (Numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu - Father's education (Numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

Mjob - Mother's job (Nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob - Father's job (Nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

reason - Reason to choose this school (Nominal: close to 'home', school 'reputation', 'course' preference or 'other')

guardian - Student's guardian (Nominal: 'mother', 'father' or 'other')

traveltime - home to school travel time (Numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

studytime - Weekly study time (Numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

failures - Number of past class failures (Numeric: n if 1<=n<3, else 4)

schoolsup - Extra educational support (Binary: yes or no)

famsup - Family educational support (Binary: yes or no)

paid - Extra paid classes within the course subject (Math or Portuguese) (Binary: yes or no)

activities - Extra-curricular activities (Binary: yes or no)

nursery - Attended nursery school (Binary: yes or no)

higher - Wants to take higher education (Binary: yes or no)

internet - Internet access at home (Binary: yes or no)

romantic - With a romantic relationship (Binary: yes or no)

famrel - Quality of family relationships (Numeric: from 1 - very bad to 5 - excellent)

freetime - Free time after school (Numeric: from 1 - very low to 5 - very high)

goout - Going out with friends (Numeric: from 1 - very low to 5 - very high)

Dalc - Workday alcohol consumption (Numeric: from 1 - very low to 5 - very high)

Walc - Weekend alcohol consumption (Numeric: from 1 - very low to 5 - very high)

health - Current health status (Numeric: from 1 - very bad to 5 - very good)

absences - Number of school absences (Numeric: from 0 to 93)