REPORT

The goal of this study is to understand the underlying reasons for lack of success in the core classes of Mathematics and Portuguese in Portugal's schools. Predictions/Insights from various machine learning models developed on real world data (e.g. demographic, social and school related features) collected using school reports and questionnaires would help in improving quality of education and enhancing school resource management.

To achieve the aforesaid goal, various machine learning models were developed to predict the grades of students and understand the key variables that affect educational performance. The impact of various features on Mathematics and Portuguese grades was studied separately.

Since predicting the Grade G3 (which is a continuous variable ranging between 0-20) is a regression task, the Test Root Mean Square Error (RMSE) metric is used to evaluate/compare the performance of the various models.

Mathematics Performance Analysis

Data Preparation:

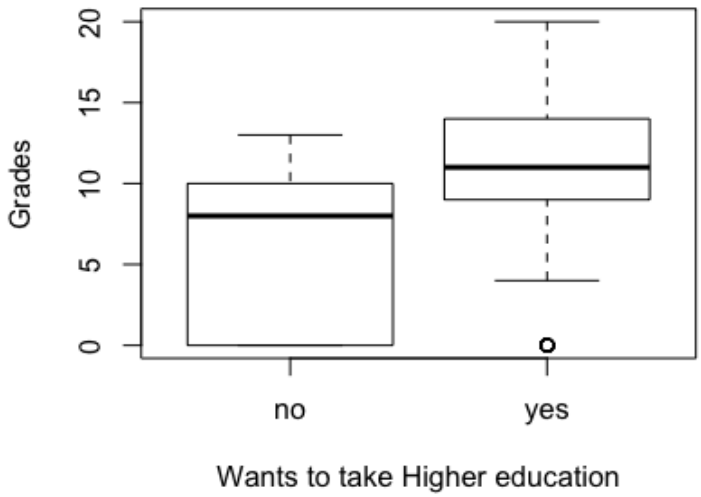
Before fitting the models, the data is made ready for further analysis. Various variables that are categorical in nature, like Mother's Education, Father's Education, study time, Daily alcohol consumption etc. are present in numerical format (signifying the ordering) and thus are converted into factor/categorical type.

Exploratory Data Analysis (EDA):

EDA insights for student's performance in Mathematics is as follows:

- Students who want to pursue higher education score (Median:11) much better than those who do not want to pursue higher education (Median:8). [Fig. 2.1]

Figure 2.1



- Student's whose Mothers work in the health sector (Median:13) perform better than other students (Median:11). [Fig. 2.2]
- Student's whose Fathers work in the teaching sector (Median:14) perform better than other students (Median:11). [Fig. 2.3]

- Student's whose guardian is either Father or Mother (Median: 10.5) perform better than other students (Median:9). [Fig. 2.4]
- Students with internet access at home (Median:11) perform better than those do not (Median:10). [Fig. 2.5]
- Gabriel Pereira's students perform better (Median:11) than their counterparts at Mousinho da Silveira (Median:10). [Fig. 2.7]
- Students that do not need extra educational support at school (Median:11) perform better than those who do (Median:10). [Fig. 2.9]
- Students taking extra paid classes have lesser variation in marks (middle 50% score between 9-13) than those who do not (mid 50% score between 8-14). [Fig. 2.10]
- Male students perform better (Median:11) than female students (Median:10) in mathematics. [Fig. 2.11]
- Urban students (Median:11) perform better than their counterparts in Rural Areas (Median:10). [Fig. 2.12]
- Performance of students with family size greater than 3 shows a wider variation (IQR:6) in comparison to those with family size of less than 3 (IQR:4.75). [Fig. 2.13]
- Students travel time is inversely proportional to their grades, with students who travel for 15-30 minutes outperforming (Median:11) others who travel for more duration (Median score:10). [Fig. 2.18]
- Students putting in 5 to 10 hours and above perform better (Median score:12) than those who do not (Median score:10.5). [Fig. 2.19]
- Number of past failures is inversely proportional to student's score. Students with no past failures are likely to perform better (Median:11) than others.
- Student who occasionally go out with friends – level 2 – perform the best. [Fig. 2.23]
- Students with very low weekday alcohol consumption perform better than the rest. [Fig. 2.24]

The EDA process helps in deriving key insights into the data and also aids in outlier detection. An outlier is a point that is distant from similar points and they can be handled using various approaches. Statistically speaking, a general rule of thumb is to treat any value outside 1.5 times Inter-Quartile Range (IQR) from 25_{th} and 75_{th} percentile range as an outlier. One approach is to remove them all together. However, this is not possible in the present scenario as the data set is relatively small (395 observations). Another approach involves imputing the values of outliers with maximum/minimum acceptable value or the mean/median/mode depending on the context. However, by imputing outliers in the present data set, we risk introducing quite a lot of bias as some of the variables have substantial number of outliers, e.g. address has 28 outliers (approx. 7% of data). Moreover, since some variables have high number of outliers, these outliers might hold some pattern which would be missed if the outliers are treated. Thus, it is expected that machine learning algorithms like Decision Trees and Random Forests that are much more robust to outliers will give better performance than linear regression models like lasso.

Predictive Performance:

After performing the Training/Test split of 80/20, the models can be fitted on the training dataset.

In regression tasks, one usually fits the standard linear model using least squares method. However, this model takes into consideration all the variables even if those variables are not effective in explaining the variations in the dependent variable Y. This results in low prediction accuracy and model interpretability.

To enhance the performance and interpretability of the least squares model, various methods like:

- Subset selection approach
- Shrinkage approach

were used. These approaches help in selecting the features that explain the dependent variables most effectively using the least number of features (for better interpretability).

One of the methods of subset selection is Best subset selection which fits a least squares regression for each possible combination of *p* predictors (30 predictors in the present dataset). However, this approach is computationally intensive as it searches through 2 to the power p models (approx. 1 billion in the present dataset). Thus, a better alternative to best subset selection is the stepwise selection approach. The backward Stepwise selection approach starts with the full least squares model containing all the *p* predictors, and then iteratively removes the least useful predictor, one at a time. The backward stepwise selection approach searches through *1+p(p+1)/2* models (466 models for p=30) to select the model that performs most optimally among them.

The criteria for variable selection could be Akaike information criteria (AIC), Bayesian information criterion (BIC), Mallow's Cp etc. Backward stepwise selection can be performed using the **stepAIC()** function in R.

The following variables are considered important for prediction by backward stepwise selection:

sex, age, famsize, Mjob, Fjob, reason, studytime, failures, schoolsup, romantic, freetime, absences

and the following relation is obtained:

G3 = 8.56 +1.02*sexM -2.94*Age +1.19*famsizeLE3 +0.22*Mjobhealth -0.55*Mjobother +0.86*Mjobservices -1.12*Mjobteacher +0.64*Fjobhealth -0.75*Fjobother -0.28*Fjobservices +1.78*Fjobteacher +0.84*reasonhome +1.07*reasonother +1.44*reasonreputation + ……. +3.47*absences

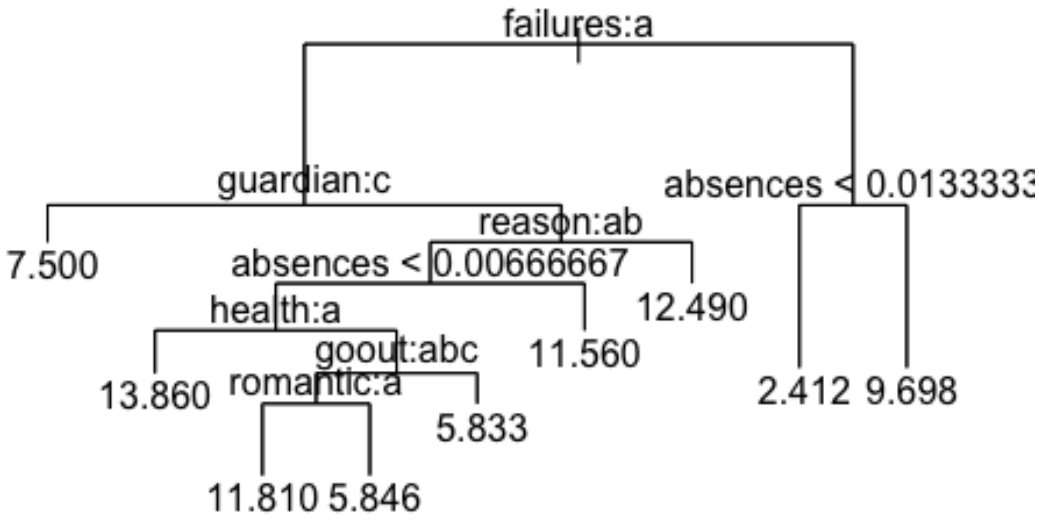
The main reason for using linear regression models is that they are easily interpretable. For example, just by looking at the model one can tell that – Male students get approx. 1.02 marks more as the coefficient of sexM is +1.02. The Backward stepwise AIC gives us a Test MSE of 19.29.

The lasso (shrinkage approach) gives a better performance and a more parsimonious model. The hyperparameter is tuned using cross validation on training dataset. The following model is obtained using lasso regression:

G3 = 8.26 +0.52*Medu.Q +0.28*Mjobservices +0.047*Fjobteacher -4.08*failures.L -0.24*failures.Q -0.25*freetime^4

Thus, lasso has considered only 5 variables to make it's model i.e., Mother's education, Mother's job, Father's job, no. of past failures and free time, resulting in a smaller and a more accurate model (Test MSE of 17.82).

Non-linear models like Decision Trees and Random Forests performed better on the data set as the relation between the independent variable (G3) and other features is not linear. The decision tree was pruned and the following DT was obtained:



The feature that results in maximum reduction in errors (RSS) is used to split the data. Thus, the most important features are split on first. The decision trees are read from top to bottom depending on the rules for a specific observation. For example, a student having failed 0 number of times previously, with a father/mother as guardian, having chosen a school for it's reputation is predicted to score 12.49 in mathematics. We would traverse the tree starting at node 1 (check if failures:0), then go to node 2 (check guardian: other), then go to node 5 (check if reason: reputation, other) and finally land on node 11 where the grade is predicted to be 12.49. The Decision tree results in a Test MSE of 16.7.

Decision trees suffer from high variance – meaning predictions of the model depend on the data set used to train the DT. Bootstrapping/Random Forests are approaches that can be used to bring down the variance of decision trees. Bagged decision trees made the most accurate predictions out of all the models and resulted in a Test MSE of 13.

Bagged decision trees also suggest that failures and absences are the most important features in predicting the student's grades.

Regression Results:

Mathematics (RMSE values)				
Backward Stepwise selection	Lasso	Decision Trees	Bagged DT	Random Forest
4.39	4.22	4.08	3.6	3.79

Similarly, following results were obtained for Portuguese performance prediction:

Portuguese (RMSE values)				
Backward Stepwise selection	Lasso	Decision Trees	Bagged DT	Random Forest
2.88	2.78	2.91	2.74	2.66