Part2- Regression

In this part of the project, we have been provided with real-world data of secondary education of two Portuguese schools with an aim for us to predict variable G3, the final grade of the student, by extract high-level knowledge from the two datasets using multiple algorithm techniques such as Linear Regression models and Data Mining models(Decision Tree, Random Forest, etc). Both of the datasets contain student's performance in two subjects i.e. Mathematics and Portuguese. We have been asked to predict G3 without G1 and G2, 1st and 2nd-period grades, as there is a strong correlation that exists between the final and last year grades. Although G1 & G2 are strong predictors in evaluating a student's final marks, there are other relevant features also such as failures, absences, whether the student is getting support from the school or not and so on. We will be further exploring this rich dataset by finding valuable information for predicting G3.

A summary of datasets:

The dataset includes 395 and 649 students samples in Mathematics and Portuguese subjects respectively by using school reports and questionnaires with a total of 33 features including the target variable i.e. G3. The input variable includes student's demographic, school-related, social and family information like sex, age, school support, absences, family size, mother's job, parent's status whether they are living together or apart, alcohol consumption, etc. By running different linear and non-linear algorithms we found that all features available in the dataset are not significant in predicting student's final grades in both the subjects.

Before fitting any model, data preprocessing is very important such as checking missing value, standardization, etc. "Data preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that the machine can easily parse it" (Pandey 2019) [6]. I have performed below data Pre-processing steps before fitting the

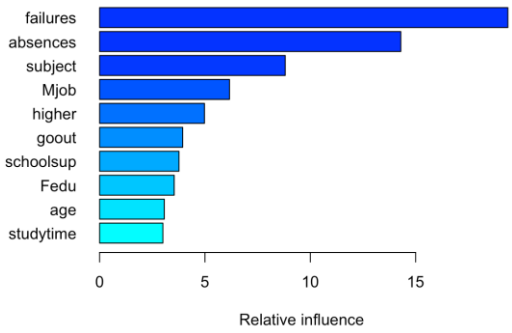
models. 1) Both subject files have been combined into one file and a new variable named "Subject" has been added with categories, Mathematics and Portuguese. 2) Continuous features like age, failures, and absences have been standardized as some machine learning algorithms are sensitive to the magnitude. 3) A few of the variables like Medu, Fedu, etc. have been converted as factors as they are categorical. In statistical modeling, implementing variables with the correct data type is very important. 4) Last, the dataset has been divided into training and test sets with a split of 75/25. We will be using the training dataset to train our model and test set will be used to evaluate model accuracy

After running Linear and Tree-based supervised learning algorithms, below model accuracy have been found on test data test:

Models	Train Error (rmse)	Test Error (rmse)
Linear Regression	3.135	3.4425
Hybrid Stepwise Selection	3.1991	3.4719
Lasso Regression	3.3166	3.3307
Decision Tree Regression	3.5068	3.136
Random Forest Regression	1.4519	3.0975
Gradient Boosting Machine	2.6446	3.0491

Conclusively, we can say Gradient Boosting has outperformed all other techniques and the best test dataset results were found with a root mean squared error i.e. RMSE of 3.04. RMSE is a standard way to measure the accuracy of the model. "It's the square root of the average of squared differences between prediction and actual observation" (JJ, 2016) [5]. From the above table, it is clear that tree-based models i.e. Decision Tree, Random Forest & Gradient Boosting have better accuracy than the Linear models. "Tree-based algorithms empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map non-linear relationships quite well" (Vidhya 2016) [9]. Boosting is a method of creating an ensemble. "Ensemble methods involve a group of predictive models to achieve a better accuracy and model stability"(Vidhya 2016) [9]. Ensembling starts by fitting a model, a tree in our case, to the data and then the next model is built based on the performance of the previous model. Let's say if the first model has performed poorly then the

later will focus on improving the performance of the first one. In general, the combined accuracy of models is expected to perform better than the single model alone [4]. "Gradient boosting is a type of machine learning boosting" (Hoare) [4]. "It is a machine learning technique that builds an ensemble of shallow and weak successive trees with each tree learning and improving on the previous" (Github) [7]. It is based on the intuition that the best possible new model when combined with the previous model, minimizes the overall error and each next model takes a step in the direction to minimize the prediction error[4]. To improve the prediction accuracy, I have used a combination of parameters in GBM function and the lowest rmse was found when the algorithm uses 110 trees. I have plotted the top 10 variables chosen by the algorithm and past failures are by far the most important predictor in predicting the final grade of the student.



The key point to note is, subject, a new variable added in the algorithm, plays an important role in predicting final grades. Mathematics and Portuguese are two very different subjects, they can't be measured on the same scale when we talk about a student's interest in subjects. Student's final grades also depend upon how many times they have missed the classes, what is their aim in their career and the occupation of their parents.