Using machine-learning model techniques to predict Secondary School Student Portuguese Language Performance from large numbers of input variables

Xinyi Li, Wenlei Li, Zhaochen Gao, Nyu Chai

2017/12/10

Abstract

In order to look into the overall education level in Portugal, we focus on students performance on Portuguese language since it acts as important factor in Portugal's core courses. We aim at finding the best machine-learning model to predict the relationship between Portuguese final grade with different factors. We try six different regression models and find Random Forest performs the best. Then we transforms the final grades into two categories (poor/good) and try five classification models and find Random Forest still performs the best.

INTRODUCTION

Although Portugal's overall education level has improved in recent years, overall the country is facing 35% (Cortez and Silva 2008) unemployment rate among youths. Based on our research, the main reason of high unemployment rates of Portugal youths is actually due to the low grades("Education in Portugal" 2017) of core classes: Mathematics and Portuguese Language. Moreover, the portuguese language is recognized worldwide as granting credits for access to higher education, especially in United States. Hence, Portuguese language actually plays a significant role in Portugal's core courses("The Education System in Portugal," n.d.).

In this study, we set out to address the question: Is it possible to construct a function which can predict Portugal youths' portuguese language final performances using related social and school factors. Results from functions created using regression modeling and classification modeling are reported.

The random forest function in both regression modeling and classification modeling are the recommended choice based on test error-rate and favourable model characteristics.

MATERIALS AND METHODS

Data Collection

This study's data was collected from two public Portuguese schools during the 2005-2006 school year. Since our analysis is to predict students' portuguese language performances, we can use the variable final grades(G3) as response and other useful demographic or social factors such as sex, study time or weekly alcohol consumption as predictors. The data called "student-por.csv" is retrieved from the UC Irvine Machine Learning Repository. It was created by several researchers and institutions including Paulo Cortez, University of Minho and GuimarÃ£es.The dataset can be found at https://archive.ics.uci.edu/ml/datasets/Student+Performance.

Here's the detailed list of variables and their descriptions.

#	Attributes	Description
1	school	student's school (binary: GP - Gabriel Pereira or MS - Mousinho da Silveira)
2	sex	student's sex (binary: F - female or M - male)

#	Attributes	Description
3	age	student's age
4	address	student's home address type (binary: U - urban or R - rural)
5	famsize	family size (binary: LE3 - less or equal to 3 or GT3 - greater than 3)
6	Pstatus	parent's cohabitation status (binary: T - living together or A - apart)
7	Medu	mother's education
8	Fedu	father's education
9	Mjob	mother's job
10	Fjob	father's job
11	reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
12	guardian	student's guardian (nominal: mother, father or other)
13	traveltime	home to school travel time (numeric: 1 - <15 m., 2 - 15 to 30 m., 3 - 30 m. to 1 hr, or 4 - >1 hr)
14	studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15	failures	number of past class failures (numeric: n if 1<=n<3, else 4)
16	schoolsup	extra educational support (binary: yes or no)
17	famsup	family educational support (binary: yes or no)
18	paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19	activities	extra-curricular activities (binary: yes or no)
20	nursery	attended nursery school (binary: yes or no)
21	higher	wants to take higher education (binary: yes or no)
22	internet	Internet access at home (binary: yes or no)
23	romantic	with a romantic relationship (binary: yes or no)
24	famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25	freetime	free time after school (numeric: from 1 - very low to 5 - very high)
26	goout	going out with friends (numeric: from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28	Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29	health	current health status (numeric: from 1 - very bad to 5 - very good)
30	absences	number of school absences
31	G1	first period grade (numeric: from 0 to 20)
32	G2	second period grade (numeric: from 0 to 20)
33	G3	final grade (numeric: from 0 to 20, output target)

Exploratory Analysis

Exploratory analysis was performed to identify the quality of data(such as ensuring no missing values) and to determine the proper terms for the regression model in order to model a relationship between Portuguese language final grades(G3) and other variables. Our exploratory analysis consisted of a)transforming the raw data(e.g.,identifying missing values, splitting dataset into train and test data(Table 1), converting variables into numeric and factor formats, as needed); and b) studying data plots to identify high correlations between factors.

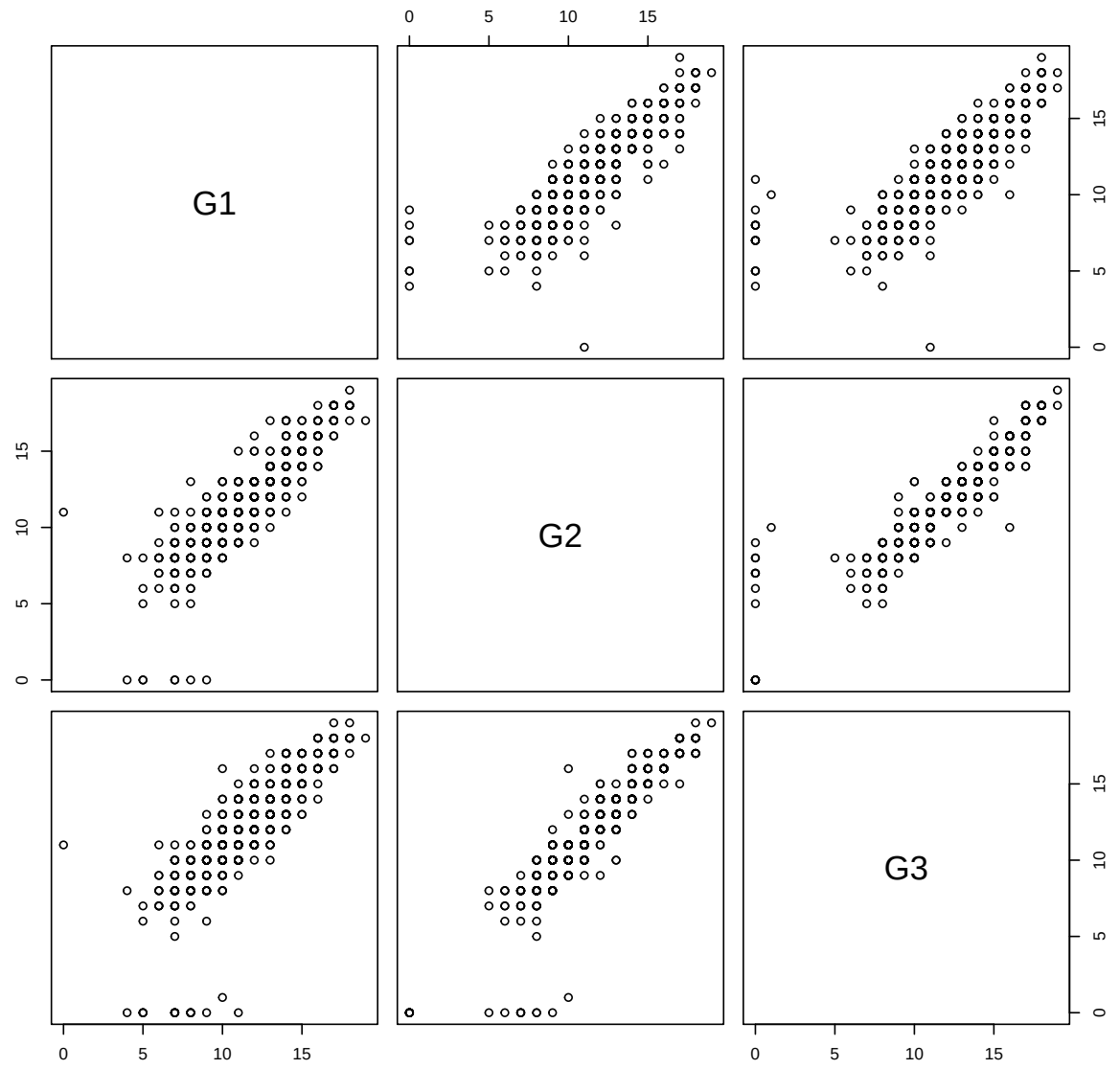
Statistical Modeling

An important part of predictive modeling is the careful partitioning of available data. A prerequisite of the study is that 80 percent of data is used for training and 20 percent of data is used for testing. In order to identify the relationships between portuguese languages final grades(G3) and other variables of interest, we used both regression and classification manner.

Figure 1: Portuguese Class

	Name of data-set	No. of obs
1	Training Set	520
2	Test Set	129

Since G1 and G2 stands for the first period grade and second period grade, there's no doubt that G3, the final grade, is composed of the first two grades. And G1 and G2 should not be good explanatory variables to explain the students' performance. In order to prove our assumptions, we look into the correlation between G1, G2 and G3.



A) Regression modeling:

Following background reading and exploratory analysis models from the general class of 'Statistical Learning' were chosen for further investigation. Six models were investigated: Linear Regression Model, K-Nearest Neighbors, K-Nearest Neighbors with scaling, Elastic Net Model, Random Forest Model, Boosted Tree Model.

At the first stage of modeling, we removed highly-correlated variables(G1 and G2) based on correlation plots(Figure 1(a)). Then models were developed, tuned and cross-validated using 5-fold split of the data training set to minimise overfitting risk. Cross Validation is a method to estimate test metrics with training data. Repeats the train-validate split inside the training data. Bespoke R functions were built to measure performance of each model. Models were evaluated by using Root Mean Square errors.

Mean Squared (RMSE) is a popular metric (Witten and Frank 2005). A high PCC (i.e. near 100%) suggests a good classifier, while a regressor should present a low global error (i.e. RMSE close to zero). (Cortez and

Silva 2008)

$$RMSE = \sqrt{\sum_{i=1}^{N}(y_i - \hat{y_i})^2/N}$$

Fit a total five models:

- An additive linear regression.

$$f(x) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_p * x_p$$

- A well tuned k-nearest neighbors model.
  - Do **not** scale the predictors.
  - Consider $k \in \{1, 5, 10, 15, 20, 25\}$

$$\hat{f}_k(x) = \frac{1}{k} \sum_{i \in N_k(x,D)} y_i$$

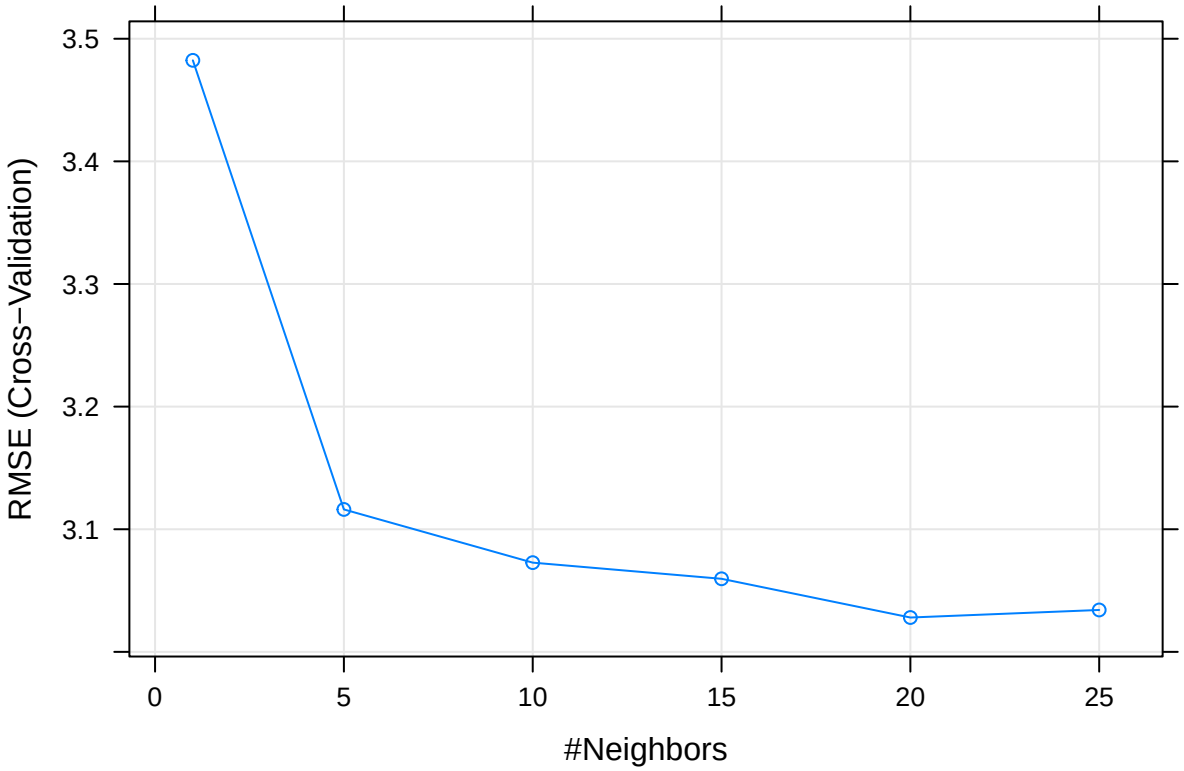
- Another well tuned k-nearest neighbors model.
  - *Do* scale the predictors
  - Consider $k \in \{1, 5, 10, 15, 20, 25\}$
- A elastic net model.

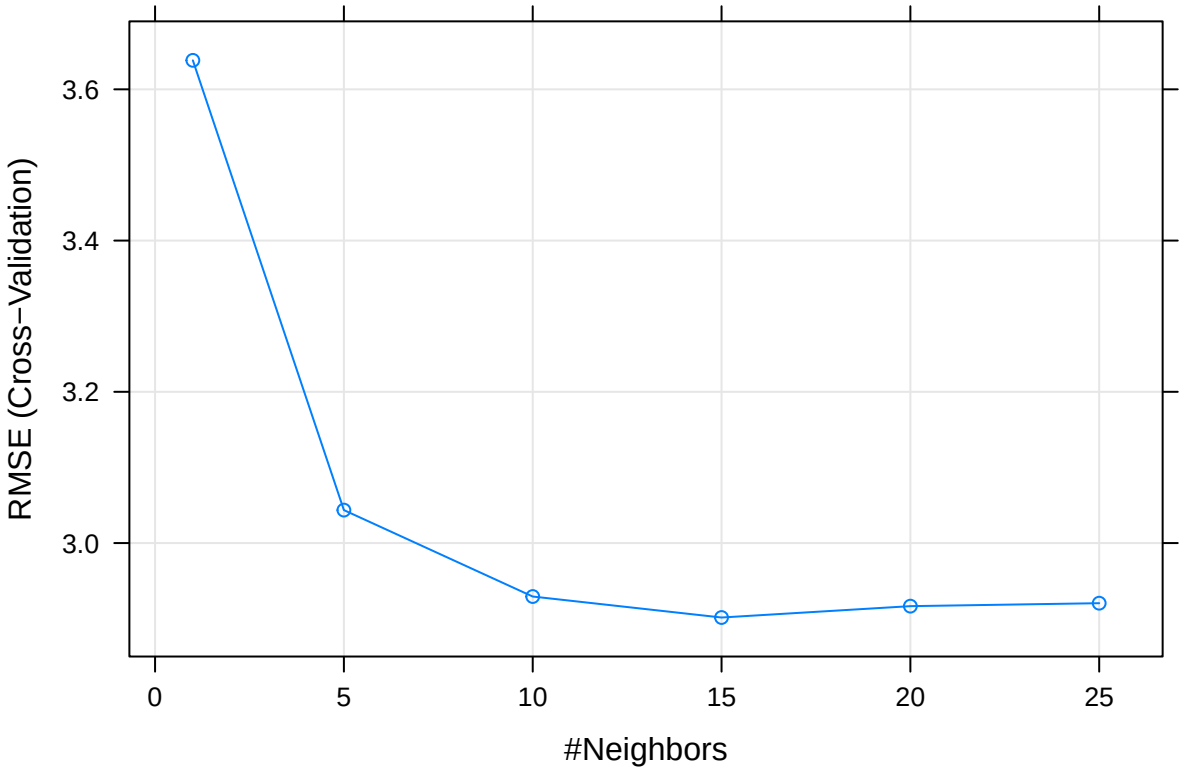
$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1]$$

$$l_1 norm : ||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$$

$$l_2 norm : ||\beta||_2 = \sqrt{\sum_{j=1}^{p}\beta_j^2}$$

- A random forest.
- A boosted tree model. Set $\hat{f}(x)$=0 and residuals $r_i = y_i$ for all i in the training set. For b = 1, 2, ..., B, repeat: Fit tree $\hat{f}^b$ with d splits to the training data (X, r)[3]. Update $\hat{f}$ by adding shrunken version of the new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda\hat{f}^b(x)$. Update the residuals $r_i \leftarrow r_i - \lambda\hat{f}^b(x_i)$ Output the boosted model, $\hat{f}(x) = \sum_{b=1}^{B}\lambda\hat{f}^b(x)$ (λ is called the shrinkage parameter. Typically, $\lambda \ll 1$)


RMSE (Cross−Validation)
3.5
3.4
3.3
3.2
3.1
0
5
10
15
20
25
#Neighbors



As we can see, knn model without scaling chooses 20 as k, and knn model with scaling chooses 15 as k.

```
#random forest
set.seed(uin)
reg_rf = train(G3 ~ ., data = student_trn_reg, method = "rf", trControl = cv_5, importance = TRUE)
reg_rf_cv = get_best_result(reg_rf)$RMSE
reg_rf_tst = calc_rmse(actual = student_tst_reg$G3,
                       predicted = predict(reg_rf, student_tst_reg))
```

1. Linear regression model is a statistical method that allows us to summarize and study relationships between the predictors and the response variable. In our case, we want to use linear regression model to check if our data follow linear relationship. That's because it assumes the linear relationship between predictors and the response variable and our data does not strictly follow linear relationship.

2. *k*-nearest neighbors model is using averages of k nearest neighbors. To make a total of two models, we consider both scaled and unscaled X data. The major weakness of knn is that we have to determine the value of parameter K. Knn is non-parametric model and doesn't have any specific assumptions.

3. In statistics and, in particular, in the fitting of linear or logistic regression models, the elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. Since we fit a model involving all p predictors, we want to use regularization methods to let some estimated coefficients be shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.

4. The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. In random forests, all the base models are constructed independently using a different subsample of the data. A fresh selection of m randomly selected predictors is presented

at each split of each tree.

5. The Boosted Trees Model is also a type of additive model that makes predictions by combining decisions from a sequence of base models. Unlike Random Forest which constructs all the base classifier independently, each using a subsample of data, boosted trees model uses a particular model ensembling technique called gradient boosting.

B) Classification modeling:

In the classification modeling, we aimed to predict if possible to use functions to identify the key variables that affect students' educational conditions(Poor/Good). Logistic regression, Generative models, Decision trees, k-Nearest Neighbors will be tested. Since there are high correlation variables G1 and G2, we removed them in classification modeling as well. Then we also used 5 fold cross validation method to minimum overfitting risk. Model size and complexity and the cross-validation error and test error were reported via following functions.

Fit a total of five models:

We use accuracy to evaluate the performance of classifications. Classification accuracy is defined as "percentage of correct predictions". That is the case regardless of the number of classes. Therefore, the higher the accuracy, the better of the performance.

- An Logistic regression.

$$log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_p * x_p$$

- A well tuned k-nearest neighbors model.

$$\hat{C} = \begin{cases} 1 & \hat{p}_{k0}(x) > 0.5 \\ 0 & \hat{p}_{k0}(x) < 0.5 \end{cases}$$

- A QDA model.

$$\hat{G}(x) = argmax(\delta_k(x))$$

- A random forest.
- A boosted tree model.

```
#random forest
rf_mod = train(
  grade ~ .,
  data = student_trn,
  trControl = trainControl(method = "cv", number = 5),
  method = "rf",
  importance = TRUE
  )
rf_cv_acc = get_best_result(rf_mod)$Accuracy
rf_tst_acc = calc_acc(predicted = predict(rf_mod, student_tst),
                      actual    = student_tst$grade)
```

In the classification modeling, we introduced two more new methods: Logistics regression and Quadratic discriminant analysis.

(1) Logistics regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

(2) Quadratic discriminant analysis(QDA), which is closely related to linear discriminant analysis and both as bayesian classifiers, is made to class predictors with highest estimated posterior probability.

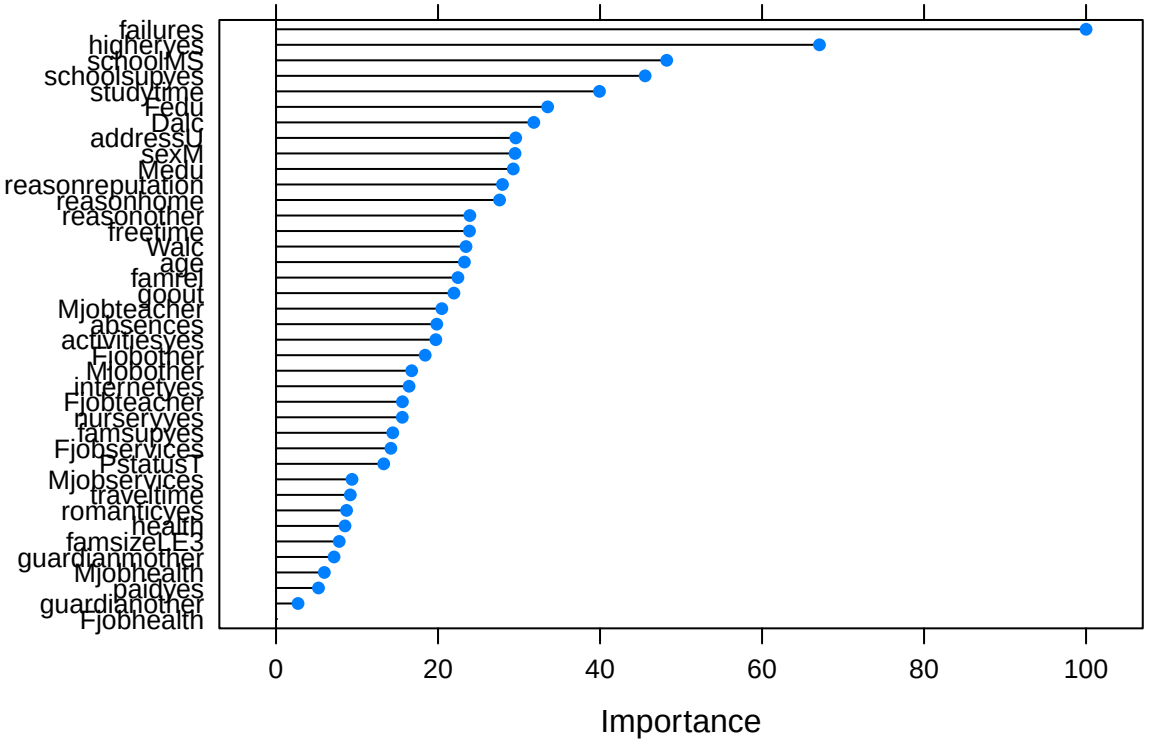
RESULTS

A) Regression Part

Table 3: Regression Analysis Results Summary Table

Model	Cross Validated RMSE	Test RMSE
Additive Linear Regression	2.832687	2.687318
KNN Without Scaling	3.028010	2.720860
KNN With Scaling	2.901539	2.670762
Elastic Net	2.768747	2.635908
Random Forest	2.737203	2.383894
Boosted Tree	2.724232	2.470888

Figure 1, Variable Importance for Regression Modeling



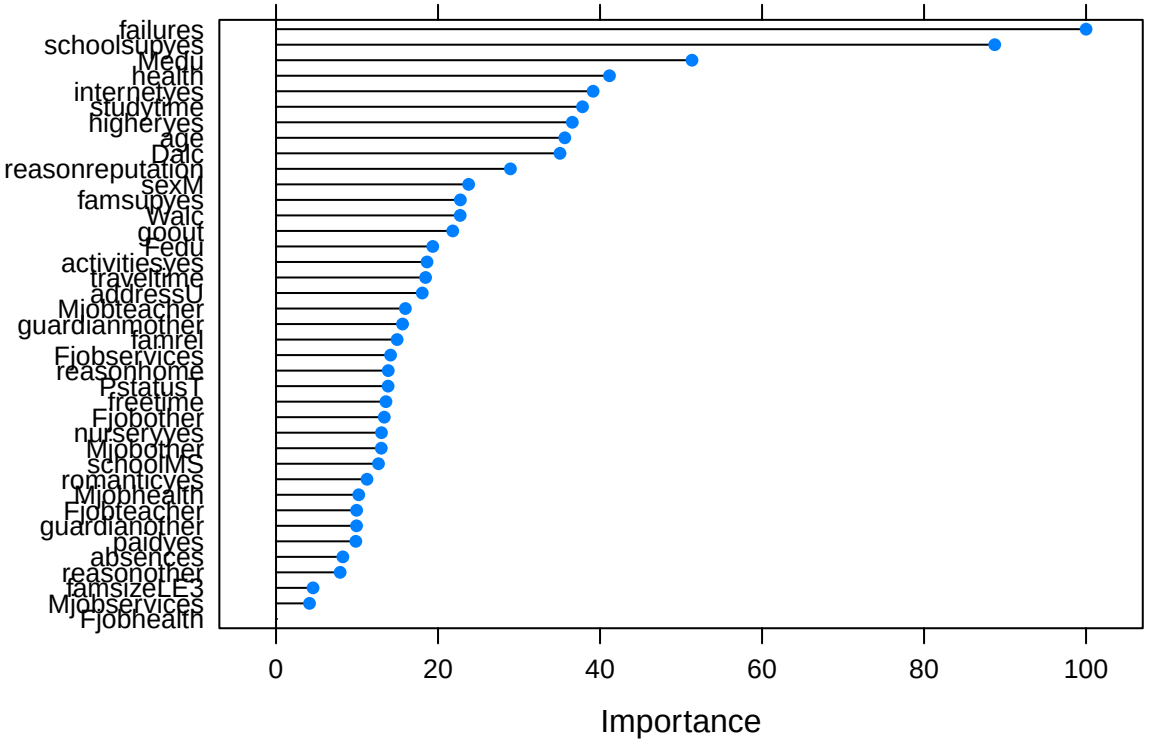
We can look into the variable importance based on the best performed Random Forest Model. The most important three variables are number of past class failures, whether wants to take higher education and whether get extra educational support.

B) Classification Part

Table 4: Classification Analysis Results Summary Table

Model	Cross Validated Accuracy	Test Accuracy
Additive Logistic Regression	0.7018205	0.6976744
KNN Without Scaling	0.7153556	0.6589147
QDA	0.7066712	0.7054264
Random Forest	0.7520618	0.7441860
Boosted Tree	0.7268381	0.6976744

Figure 2, Variable Importance for Classification Modeling



Again, we can look into the variable importance based on the best performed Random Forest Model. The most important three variables are still number of past class failures, whether wants to take higher education and whether get extra educational support.

The source data contained a total of 649 observations and 33 variables. The full codebook is available via ("430project-11.Rmd"). From the information provided and validated through basic checks, no missing data were observed through 33 columns. In this study, we splitted the data into 80% training set and 20% testing set. Based on our exploratory data analysis, we found that highly-correlated columns 'G1' and 'G2' could be described as "confounders" in the conventional modeling. Additionally, background reading suggested to develop our analysis in several directions: decision trees, K nearest neighbour(KNN) modeling, linear models, generative models, Elastic net and logistic regression. Therefore, we removed these two variables and then performed regression modeling and classification modeling.

Regarding regression modeling, it was observed during modeling that Random forest tree performed best with lowest Test RMSE. Next, our classification modeling resulted in giving conclusions that random forest tree model performed best among all models with lowest Test RMSE.

DISCUSSION

A) Regression Part

Based on our results Table 3 (Classification Analysis Results Summary Table), we determine the best model based on test RMSE, which represents the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. According to our results, random forest model performs the best. The final model is a non-linear method. And it's non-parametric method. It's discriminant. And we continue discussing the results for each model and the possible reasons behind the results.

1. Linear regression model doesn't perform well.That's because it assumes the linear relationship between predictors and the response variable and our data does not strictly follow linear relationship.
2. In our case, two knn models perform the worst due to the curse of dimensionality. And there's no big difference in test RMSE for scaled and not-scaled models. Therefore, scaling is not appropriate.
3. For elastic net model, in our case, alpha is between 0 and 1 and it's closer to 1. So it's closer to lasso method. Elastic Net seemed to do better under correlated true variables situations, did worse in the case of correlated true and noise predictors.
4. In our case, random forest model performs the best with the lowest test RMSE.
5. Boosted tree model doesn't outperform random forest model. Boosting works best if d (the size of each tree) is small.

B) Classification Part

Based on the test accuracy Table 4 (Regression Analysis Results Summary Table), we select the Random Forest as the best model. The final model is a non-linear method. And it's non-parametric method. It's discriminant.

From the accuracy we got, logistic regression does not perform well. Because it assumes the linear relationship between predictors and our data is not in linear relationship. k-nearest neighbors model is using averages of k nearest neighbors. Knn models perform the worst because of the curse of dimensionality. Therefore the accuracy is low. QDA does not perform well. Because the predictors X is not normal in each of the classes. Moreover, the sample size of test data is only 129, which is less than 5 times the number of predictors. The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. Same as the regression, boosted tree will perform better if n is small.

According to Figure 1 (Variable Importance for Regression Modeling) and Figure 2 (Variable Importance for Classification Modeling), two random forest models pick up the same three most important variables.

CONCLUSION

In the current Portugal education system, Portuguese language still plays significant roles in influencing students' professional and educational development. Therefore, our group' purpose of this project is to construct functions to predict Portugal's secondary school student Portuguese language grades. In the procedure of analysis, we performed statistical modelings with detailed functions and methods. Our analysis has shown it is feasible to predict Portugal youths' portuguese language final performances, based on known and accepted machine learning techniques - Random Forest tree model- which have excellent Test RMSE result.

References

Cortez, Paulo, and Alice Maria Gonçalves Silva. 2008. "Using Data Mining to Predict Secondary School Student Performance." EUROSIS.

"Education in Portugal." 2017. *Wikipedia*. Wikimedia Foundation. https://en.m.wikipedia.org/wiki/Education_in_Portugal.

"The Education System in Portugal." n.d. *Education | Expatica Portugal*. https://www.expatica.com/pt/education/Education-in-Portugal_105195.html.