Effects of personality traits in predicting grade retention in Brazilian students (Supplementary Material)

Anonymous Authors[1]

1. Data variables

The information we used from the surveys can be divided in three parts: (i) socioeconomic questions, that evaluate students profiles in 2012 or in previous years; (ii) mathematics and language standardized tests, taken by students in 2012, and that were prepared by professor Ricardo Primi using items of the National Institute for Educational Studies and Research Anísio Teixeira (INEP); (iii) a personality test called Big Five Inventory, taken by students in 2012, which measured students scores in 10 "facets".

1.1. Personalities

1. *Activity*: measures propensity to be energetic, fast-paced and vigorous.
2. *Aesthetics*: measures sensibility to art and beauty.
3. *Altruism*: measures propensity to kindness and gentleness.
4. *Anxiety*: measures propensity to feel nervous and tense.
5. *Assertiveness*: measures propensity to demonstrate a decisive and self confident behaviour.
6. *Compliance*: measures propensity to be deferential and obliging.
7. *Depression*: measures propensity to experience bad feelings like sadness, guilty, hopelessness.
8. *Ideas*: measures propensity to be analytical and theoretically oriented.
9. *Order*: measures propensity to be precise and efficient.
10. *Self-Discipline*: measures propensity to be organized and through.

1.2. Other features

1. *grade_2012*: Which school year the student was in 2012.
2. *grade_2017*: Which school year the student was in 2017. This item was used to calculate grade retention and was removed from the dataset to further analysis.
3. *follow_up*: Whether the student respond to 2017's surveys. Students who haven't being re-interviewed was removed from the data. This variable was removed in further analysis
4. *month*: The month student born.
5. *year*: The year student born.
6. *school_2012*: Which type of school the student have studied.
7. *ethnicity*: which ethnicity the student identifies with.
8. *gender*: whether the student identifies with male or female.
9. *failed_before_2012*: if the student has already repeated a grade before the year of 2012 .

   It's factors: 'sim' and 'não'

   Which translated means, respectively: 'yes' and 'no'
10. *mother_educ*:

    Description: student mother's school degree

    It's factors: 'nunca_estudou', or 'ef1', or 'ef2', or 'em', or 'superior', or 'nao_sabe'

    Which translated means, respectively: ef1 = 'Ensino fundamental 1' = It means that if the mother attended each grade at the ideal time, the mother left school at some point between 10 and 14 years old.

    ef2 = 'Ensino fundamental 2' = It means that if the mother attended each grade at the ideal time, the mother left school at some point between 14 and 17 years old.

    em = 'Ensino médio' = High school, composed of 3 school years, It's expected that the person studies it between 14 and 17 years old. So the mother left school at 17 years old and didn't pursue further education.

    superior = 'Ensino Superior' = It means the mother has finished a university degree.

    nao_sabe = 'Não sabe' = It means that the person who was interviewed doesn't know about their mother's education.

11. *pre_k_pub, pre_k_priv, kinder_pub, kinder_priv*:

    pre_k_pub: if the student used to frequent a public pre-kindergarden

    pre_k_priv: if the student used to frequent a private pre-kindergarden

    kinder_pub: if the student used to frequent a public kindergarden

    kinder_priv: if the student used to frequent a private kindergarden

    It's factors: sim, nao

    Which translated means, respectively: 'yes','no'

12. *delay*: It tells whether the student has repeated at least one grade in school. It's factors: 0 (so the student never repeated a grade in school) or 1 (the student repeated at least one grade in school). Delay was used only as target variable and was excluded from predictive variables

13. *delay_years*: It tells how many grades the student has repeated. This variable wasn't used in further analysis.

14. *Languages* Result of a skill test in language Portuguese collected in 2012. The grade used item response theory to permit more comparable grades between the students.

15. *Mathematics* Result of a skill test in mathematics collected in 2012. The grade used item response theory to permit more comparable grades between the students.

2. Hyperparameters

For the Logistic Regression we used a elastic-net regularization and choose the C parameter and l1_ratio by cross-validation with 3 folds and 100 iterations. Se Table 1 for values tested.

For the XGBoost classifier, we fix the use of 500 boosted trees. The other most important hyperparameters were validated according to the values in Table 2. The hyperparameters' names are reported as they are in the XGBoost Python API[1]. Just like we did in logistic regression, we choose hyperparameters by cross-validation with 3 folds and 100 iterations.

3. Additional graphics

3.1. SURVEL

We performed a SURVEL analysis with all features to have a general relevance order

[1]See https://xgboost.readthedocs.io/en/latest/python/python_api.html

Table 1. Possible Values hyperparameters for the Logistic Regression classifiers, this parameter grid was used in randomizedSearchCV function with 100 iterations and k-fold = 3

Hyperparameter	Possible Values
C	1000 values between $(1, 30)$
l1_ratio	100 values between $(1, 100)$

Table 2. Possible Values hyperparameters for the XGBoost classifiers, we tested more options and iterations, but this combination gave the best results. For this randomizedSearchCV we use 25 iterations with a k-fold=5

Hyperparameter	Possible Values
Gamma	$0.1, 0.32, 0.55, 0.77, 1.$
Lambda	$10., 16.67, 23.33, 30., 36.67, 43.33, 50.$

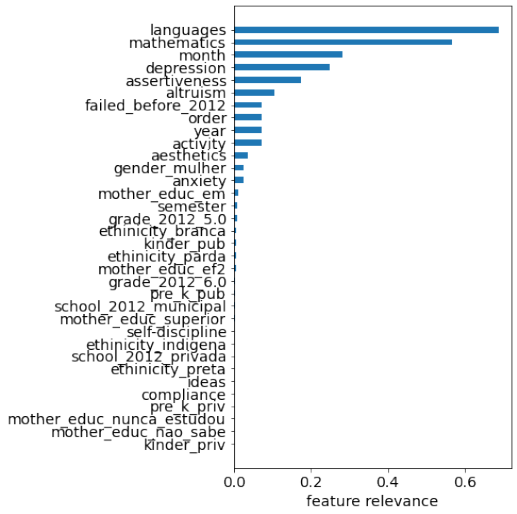

Figure 1. SUVREL relevance analysis with all features

3.2. Conditions for Student's t-test

We verify the conditions to Student's t-test by performin a Levenne's test for variance and performing a QQplot to verify normality of the data.

Model	p-value
Logistic Regression	0.902
XGBoost	0.520

Table 3. Levenne's statistic p-value

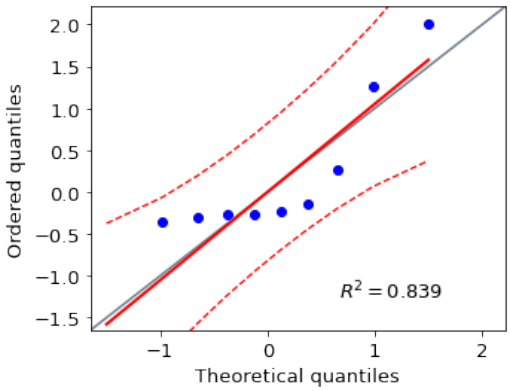

Figure 2. qqplot of gain distribution used to compare the similiarity of the distribution with normal distributed data

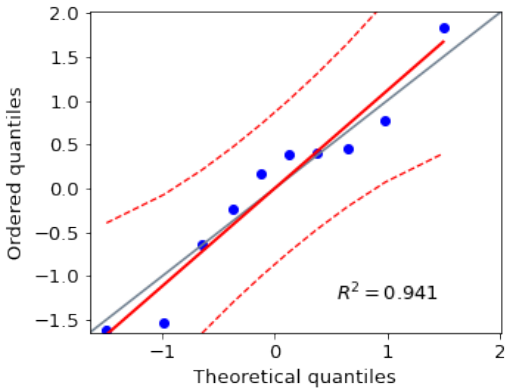

Figure 3.

4. Models and Code

Due to bureaucratic reasons, we cannot share the actual dataset used. In the link below, all the code used to generate this paper can be accessed.

https://github.com/grupo-atraso/paper-submission/tree/main