ML Project Status Report - Grade Predictor

Nagabharan Nagendran (nxn141730)

Pavithran Chittoor Rajarajan (pxc140530)

Description of the data set:

This is a student performance data set.

Data set information from the source mentions this “this data approach student achievement in secondary education of two portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: mathematics (mat) and portuguese language (por). In [cortez and silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute g3 has a strong correlation with attributes g2 and g1. This occurs because g3 is the final year grade (issued at the 3rd period), while g1 and g2 correspond to the 1st and 2nd period grades. It is more difficult to predict g3 without g2 and g1, but such prediction is much more useful (see paper source for more details).”

URL: https://archive.ics.uci.edu/ml/datasets/Student+Performance

Number of instances:

|  |  |  |  |
| --- | --- | --- | --- |
| **Data set characteristics:** | Multivariate | **Number of instances:** | **649** |
| **Attribute characteristics:** | Integer | **Number of attributes:** | 33 |

Number and brief description of features:

|  |  |  |
| --- | --- | --- |
| **Attribute number** | **Attribute names** | **Attribute information** |
| **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) |
| **3** | Age | Student's age (numeric: from 15 to 22) |
| **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 (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education) |
| **8** | Fedu | Father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education) |
| **9** | Mjob | Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. Administrative or police), 'at\_home' or 'other') |
| **10** | Fjob | Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. Administrative or police), 'at\_home' or 'other') |
| **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 min., 2 - 15 to 30 min., 3 - 30 min. To 1 hour, or 4 - >1 hour) |
| **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 (numeric: from 0 to 93) |
| **31** | **G1** | **First period grade (numeric: from 0 to 20)** |
| **31** | **G2** | **Second period grade (numeric: from 0 to 20)** |
| **32** | **G3** | **Final grade (numeric: from 0 to 20, output target)** |

Note:

***G1 and g2 i.e; midterm scores are strong attributes that correlate strongly with the class label g3 which is the final grade. Using them will give accurate g3 classification, but makes little contribution. Learning from the rest of the attributes gives true characteristic and earning of a students pattern and prediction.***

***G3 is the final class label that we are going to predict.***

Type of output variable:

32 g3 - final grade (numeric: from 0 to 20, output target)

***G3 is the final class label that we are going to predict which is the grade and is a real value.***

Experimental methodology:

How are you going to train your classifiers?

1. Decision tree
2. Neural network
3. SVM
4. Naive Bayes
5. K-NN
6. Bagging
7. Boosting(Adaboost & gradient boosting)
8. Random Forest
9. Deep Learning(based on time available)

How will you prevent overfitting?

1. Removing redundant features
2. Using 2-6 levels of cross validation for (test/train/validation)
3. Reduce dimensionality of model if needed

How will you test your classifiers?

**Train** 70%, **Test** 20% & **Validation** 10%

Metrics to be used: accuracy, precision, recall and f-score

How do you ensure that the training data doesn't have null value or redundant data?

Null - Using R’s na.omit(dataset) function we can remove null entries

Redundant – We isolate needed columns and run unique(dataset[,n:m]) function in R

How will you visualize your results?

Histogram, barplot, iplots, rocr

How will you validate your results to make sure they make sense?

Manually run through a sample set and see if the classification suits the data in the attributes. Since performance of a student is subjective in nature, human deduction is the way to make sure the results make sense.

Coding Techniques:

Language - R

Packages - rpart, e1071, party, class, neuralnet, nnet, ipred, adabag, gbm, randomforest, h2o, deepnet

References:

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN978-9077381-39-7. [[Web Link]](http://www3.dsi.uminho.pt/pcortez/student.pdf)