Capstone 2 Project

Student Performance Analysis and Prediction

Zhaoqi Fan

Table of Content

1. Introduction

A survey about student achievement was implemented in the secondary education of two Portuguese schools. The data attributes included student grades, demographic, and social- and school-related features) The target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued in the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. The board of school board would like to propose new policies based on the survey data to enhance students' performance in the future.

1.1 Problem statement

What measures can be taken in the secondary school to increase students' final year grades (G3) by 10% in the next period by introducing new policies?

1.2 Scope of solution space

- This project will focus on the categorical and numerical features of the survey to analyze the key factors in the G3.
- Regression models will be developed to represent the dependence of G3 on the key features and test possible policies.

1.3 Constraints

- Some features in the dataset can be treated as either categorical or numerical. The use of various data types brings uncertainties to the modeling.
- Some features, e.g., the education of parents, may impact students' performance but not be able to consider in the policy

1.4 Project plan

This project is implemented by following the procedure below.

a) Problem Identification
b) Data Wrangling
   - Check data integrity
   - Clean data: missing values, duplicates, etc.
c) Exploratory Data Analysis
   - Summarize important features
   - Statistical property analysis
   - Visualize data

- Identify trends and patterns

d) Pre-Processing and Training Data Development
   - Finalize the dataset by removing dispensable features
   - Splitting the dataset into testing and training subsets
   - Train the model

e) Modeling
   - Retrain the model with a whole set of data
   - Predictions and result analysis

f) Documentation
   - A project report, a slide deck, and Jupyter notebooks will be generated and shared in a GitHub repo.

1.5 Data Source

The dataset was scraped from Kaggle. It consists of 33 Column Dataset Contains Features like gender, age, size of family, Father education, Mother education, Occupation of Father and Mother, Family Relation, Health, Grades, etc. A more detailed explanation of column names can be found in Appendix and also in UCI Machine Learning Repository.

2. Data wrangling

The objectives of this section are to explore:

- Identify any issues that will require data cleaning
- Analyze features in the dataset

2.1 Data Integrity

Target variable: G3 is a continuous grade with a range of 0-56. There are 38 records with G3 = 0. It may be a result of the student missing the exam. Therefore, all the rows with G3 of zero are dropped.

Missing data: there are no missing values in the dataset

Duplicates: there are no duplicated rows in the dataset

2.2 Feature Analysis

The data types of all features are summarized in Table 2.1 below.

Table 2.1 Data types of features

Data types	Number of features
Nominal object	17
Ordinal int	11
Continuous int	5 (including G3)

Categorical features like "school" can be used to analyze students' performance (G3. i.e., final grade) in two schools. The dependence of G3 on these categorical features will be analyzed in the next section.

Numerical features consist of 11 ordinal features and 5 numerical features. The statistics of numerical features are tabulated in Table 2.2. The ordinal features range from 0 to 5, which only represent limited information about these features. It may negatively affect the model's performance later. It has also been found that

- Traveltime, failures, Dals features are far away from a normal distribution. A transformation may be needed in the next section.
- G3 is strongly related to G1 and G3 as shown in Figure 2.1.

Table 2.2 Statistics of numerical features

	count	mean	std	min	25%	50%	75%	max
age	395.0	16.696203	1.276043	15.0	16.0	17.0	18.0	22.0
Medu	395.0	2.749367	1.094735	0.0	2.0	3.0	4.0	4.0
Fedu	395.0	2.521519	1.088201	0.0	2.0	2.0	3.0	4.0
traveltime	395.0	1.448101	0.697505	1.0	1.0	1.0	2.0	4.0
studytime	395.0	2.035443	0.839240	1.0	1.0	2.0	2.0	4.0
failures	395.0	0.334177	0.743651	0.0	0.0	0.0	0.0	3.0
famrel	395.0	3.944304	0.896659	1.0	4.0	4.0	5.0	5.0
freetime	395.0	3.235443	0.998862	1.0	3.0	3.0	4.0	5.0
goout	395.0	3.108861	1.113278	1.0	2.0	3.0	4.0	5.0
Dalc	395.0	1.481013	0.890741	1.0	1.0	1.0	2.0	5.0
Walc	395.0	2.291139	1.287897	1.0	1.0	2.0	3.0	5.0
health	395.0	3.554430	1.390303	1.0	3.0	4.0	5.0	5.0
absences	395.0	5.708861	8.003096	0.0	0.0	4.0	8.0	75.0
G1	395.0	10.908861	3.319195	3.0	8.0	11.0	13.0	19.0
G2	395.0	10.713924	3.761505	0.0	9.0	11.0	13.0	19.0
G3	395.0	10.415190	4.581443	0.0	8.0	11.0	14.0	20.0

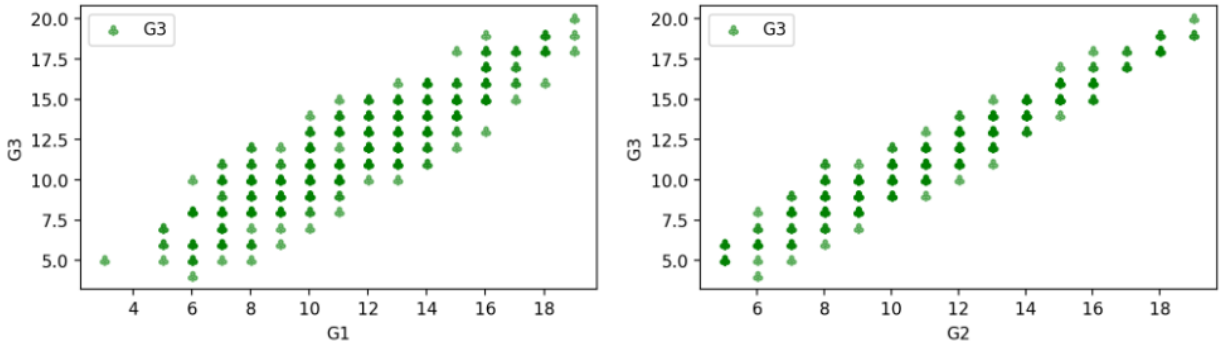

Figure 2.1 Correlation among G3 and G1, G2

3. Exploratory Data Analysis

The objective of EDA is to explore the characteristics of features and the correlations among features and target variable G3.

Barplot is used to explore the correlation between categorical features and G3 as shown in Figure 3.1. The variances of these features are quite large, which smears their correlation with G3.

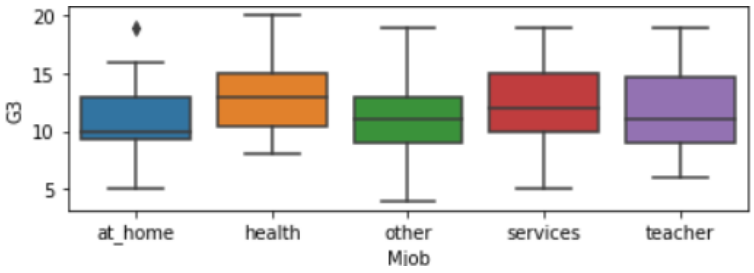

Figure 3.1 G3 vs. Mjob

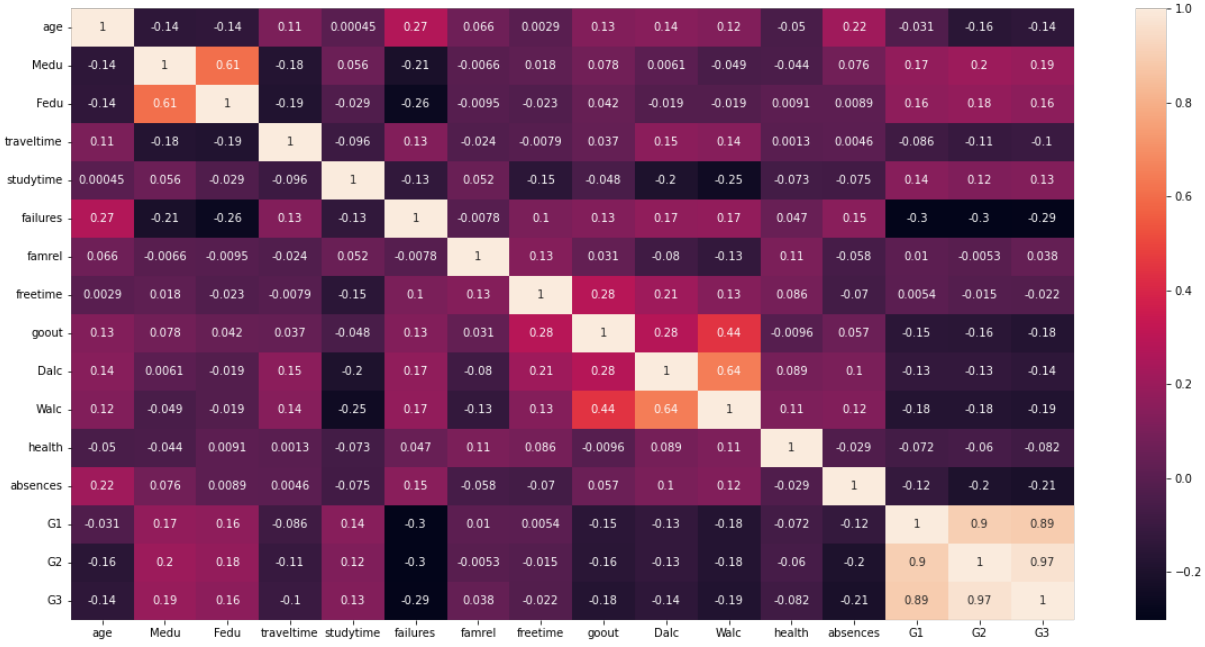

Figure 3.2 Heat map

The correlation between G3 and numerical features is shown in Figure 3.2. The heat map shows that most of the features are poorly related to G3 except G1 and G2. The Pearson correlations of the rest of the features are no larger than 0.3. The strong correlation between G1, G2, and G3 may depress the weights/effects of other variables. Therefore, G1 and G2 are removed from the dataset. We only use the rest of the features to train the model. In addition, rows with age = 20, 21, and 22 are moved due to the small number of observations.

4. Pre-Processing and Training Data Development

The objectives of this step are

- Create dummy or indicator features for categorical variables
- Split your data into testing and training datasets
- Standardize the magnitude of numeric features using a scaler

4.1 Pre-Processing

All the "object" features in the dataset are decoded by dummy features. As a result, the number of features increased from 33 to 59.

We will split the data with test_size = 0.25, i.e., training uses 75% of data. A random_state is defined to make sure anyone can repeat the procedure. Thereafter, the StandardScaler is used to standardize the train and test datasets for modeling.

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=47)

4.2 Training data

We tried multiple models to train the data. The metric of the r2 score is used to compare the performance of various models. Note that cross-validation is applied in the model comparison procedure. Table 4.1 list the r2 score of various models. The r2 score of test data is the result of cross-validation.

Overall, we can observe that the r2 score obtained from training is much higher than the r2 score obtained from test data. It seems an overfitting problem. However, the simplest linear model also yields a training r2 score much higher than a test r2 score.

To improve the modeling performance, the following measures are tried. The r2 score from test data is improved from 0.139 (Random Forest 1) to 0.156 (Random Forest 1), which is still quite low for a predictive model.

- Dropping the less important features
- Applying transformation to some features
- Test various test_size

In the next step, more models are tried as shown in Table 4.1. By comparing the r2 score for training and test data, the ***Bayesian Ridge Regression*** is the best model. Subsequently, we implemented hyperparameter tuning with the assistance of GridSearchCV and RandomSearchCV. However, the best r2 score is still 0.144, indicating that the model performance can't be further improved by tuning hyperparameters of the Bayesian Ridge Regressor.

Table 4.1 Model comparison

Model	R2 score		Selected model
	Training data	Test data	
Linear Regression	0.361	0.077	
Random Forest 1	0.884	0.139	
Random Forest 2	0.880	0.100	
Logistic Regression	0.618	0.042	
Ridge Regression	0.370	0.057	
Bayesian Ridge Regression	0.332	0.144	√
Gradient Boosting Regression	0.796	0.033	
Support Vector Machine	0.439	0.136	
Stochastic Gradient Descent Regression	0.370	0.061	
Elastic Net Regression	0.111	0.045	
LGBM Regression	0.834	0.097	
CatBoost Regression	0.412	0.109	

5. Modeling

The model comparison recommends the Bayesian Ridge Regression for this project. We refit the model by using the whole dataset (no splitting). Theoretically, a higher r2 score is expected as we have more data used to train the model. However, the r2 score is surprisingly reduced to 0.12 from 0.144. Overall, no model can appropriately represent the student's performance dataset. Given such a low r2 score, it is meaningless to implement any prediction.

Alternatively, we tried using a few features ('G1', 'G2', 'absences', 'failures', and 'G3') to build a compromised Bayesian Ridge Regression model. It yields r2 scores of 0.942 and 0.936 for the training and test dataset. It is good enough for a predictive model. The predicted G3 and the observed G3 are shown in Figure 5.1. It also implies that poor data quality is responsible for bad model performance while the whole dataset is used in this project.

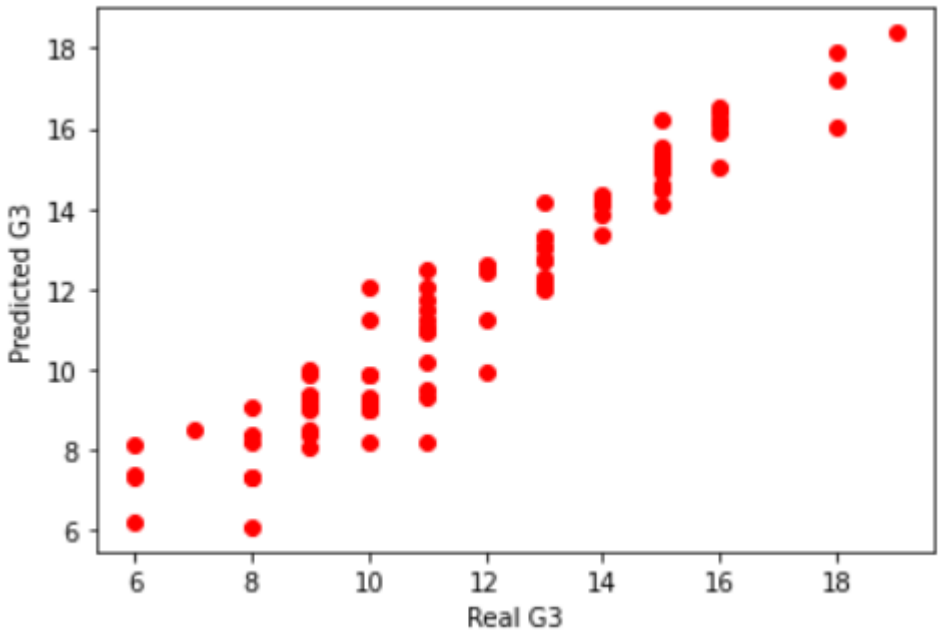

Figure 5.1 Predicted vs real G3 by a compromised model

Given the noninformative dataset, no model can yield an r2 score higher than 0.8 while using the whole dataset. Therefore, there is no way to identify the solutions to the problems below.

What measures can be taken in the secondary school to increase students' final year grades (G3) by 10% in the next period by introducing new policies?

6. Conclusions

We have to conclude that the dataset provided can't figure out a solution to improve students' performance because no predictive model can be developed properly. The poor model performance is mainly caused by the limited information included in the dataset. The following observations are summarized from this project.

- There is no missing or duplicated data;
- The rows with the target variable G3 of zero have been dropped.
- G3 is strongly dependent on G1 and G2
- All the other features are weakly related to the G3 (Pearson correlation < 0.3). One possible reason is the collection of data is less informative. Most of the numerical data are ordinal, which only exhibits limited correlation with the G3.
- The trained model is not reliable to predict G3 under various conditions as all the recommendations will be disputable.
- A compromised model only using G1, G2, and a few important features yield a satisfactory predictive model. However, it is not sufficient to deliver solutions for this project.

7. Recommendations & Future Work

To build a more robust predictive model for policymakers, the following changes are recommended.

- While the selected features are not closely related to the target variable, the ordinal features only deliver limited information. The format of data collection needs to be revised in the following surveys to collect more continuous data for regression, which may help solve the over-fitting problem and find a better predictive model.
- As the current features are not strongly related to G3, the experimental design needs to be revised to include more important features and remove negligible features.

Appendix

Table 1 Explanation of features in student performance dataset.

Feature	meaning	notes
school	students school	binary: GP - Gabriel Pereira or MS - Mousinho da Silveira
sex	students sex	binary: F - female or M - male
age	students age	numeric: from 15 to 22
address	students home address type	binary: U - urban or R - rural
famsize	family size	binary: LE3 - less or equal to 3 or GT3 - greater than 3
Pstatus	parents cohabitation status	binary: T - living together or A - apart
Medu	mothers education	numeric: 0 - none, 1 - primary education 4th grade, 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education
Fedu	fathers education	numeric: 0 - none, 1 - primary education 4th grade, 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education
Mjob	mothers job	nominal: teacher, health care related, civil services e.g. administrative or police, at_home or other
Fjob	fathers job	nominal: teacher, health care related, civil services e.g. administrative or police, at_home or other
reason	reason to choose this school	nominal: close to home, school reputation, course preference or other
guardian	students guardian	nominal: mother, father or other
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
studytime	weekly study time	numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
failures	number of past class failures	numeric: n if 1<=n<3, else 4
schoolsup	extra educational support	binary: yes or no
famsup	family educational support	binary: yes or no
paid	extra paid classes within the course subject	Math or Portuguese binary: yes or no
activities	extra curricular activities	binary: yes or no
nursery	attended nursery school	binary: yes or no
higher	wants to take higher education	binary: yes or no
internet	Internet access at home	binary: yes or no
romantic	with a romantic relationship	binary: yes or no
famrel	quality of family relationships	numeric: from 1 - very bad to 5 - excellent

freetime	free time after school	numeric: from 1 - very low to 5 - very high
goout	going out with friends	numeric: from 1 - very low to 5 - very high
Dalc	workday alcohol consumption	numeric: from 1 - very low to 5 - very high
Walc	weekend alcohol consumption	numeric: from 1 - very low to 5 - very high
health	current health status	numeric: from 1 - very bad to 5 - very good
absences	number of school absences	numeric: from 0 to 93
G1	first period grade	numeric: from 0 to 20
G2	second period grade	numeric: from 0 to 20
G3	final grade	numeric: from 0 to 20, output target