Unveiling the Crystal Ball of Education

Team Java

Julia Chu, Victoria Shi, Yiru Zhang, and Yuyan Zhang

2/27/23

Using data from two Portuguese secondary schools, this project examines the connection between individual factors and educational outcomes. Focusing on inference rather than prediction, this study creates three classification models to investigate the explanatory correlation between student demographics, social and school-related characteristics, and the final year grade, G3. The findings indicate that past evaluations have a significant impact on student achievement, and that education attendance plays a significant role in determining educational success regardless of socioeconomic and demographic factors. In addition, the study suggests that policymakers and educators prioritize improving attendance rates and provide targeted support for students who are struggling. Parents can also play an important role in their children's education. The study concludes that additional research is required to identify additional strategies for enhancing academic achievement.

1 Background / Motivation

Education is fundamental to one's own growth and that of society in its entirety as it equips individuals with the information, abilities, and morals they'll need to thrive in modern life. Certainly not always, but quite frequently. Gains in income, decreased poverty, and increased productivity all point to it as a major factor in the expansion of the economy and the improvement of social conditions.

Our research will look into what role personal attributes play in explaining academic success or failure.

2 Problem statement

The problem statement involves examining the correlation between individual factors and educational outcomes with the aim of identifying the factors that have the greatest influence on student performance. The project's specific objective is to determine which characteristics are associated with a greater likelihood of passing or failing. This is an inference problem.

This is also a classification problem because the final grade is not continuous and will be encoded as two values (pass and fail). To evaluate model accuracy, multiple metrics, including accuracy, precision, and recall, will be considered. Our primary objective is to identify variables that correlate with failure in order to provide additional resources to students who are struggling. Therefore, we will reduce the number of false positives, allowing us to identify more students who require assistance.

Priority in model development is variable selection that identifies statistically significant relationships with the final grade. Additionally, we are interested in identifying relationships about predictors that are associated with progress (improving over the course of a school semester.)

The overall project objective is to gain a deeper understanding of the complex factors that impact educational success and student performance. By identifying the variables most strongly associated with passing or failing, we can provide targeted resources and support to students who are struggling.

3 Data sources

The information was obtained from Kaggle. In reality, though, the data originates from the UCL Machine Learning Repository and the University of Minho in Portugal. The data set contains student demographic and performance statistics from two Portuguese secondary schools. And we intend to identify the specific characteristics (which will serve as our predictors) that influence the final G3 grade for the year (outcome variable).

4 Stakeholders

A wide variety of stakeholders, such as parents, students, professionals working in the education business, economists, and policymakers, could perhaps have an interest in our projects. It is possible that through examining the dataset and gaining insights, we can discover ways in which school curricula can be improved as well as methods by which policymakers and economists can provide better support for future generations.

5 Data cleaning and quality check

There are no missing values in the dataset. During the data quality check, some outliers were identified in each predictor variable, but after conducting statistical tests, it was determined that these outliers were not influential points and would not significantly impact the model. However, data preparation and transformation were further conducted to enhance model development

```python
import pandas as pd
import warnings
warnings.filterwarnings('ignore')

df = pd.read_csv("data/student-mat.csv")

num_col = list(df.select_dtypes(include=['int64', 'float64']))
cat_col = list(df.select_dtypes(include=['object']))
print(f"numerical columns \n{num_col}\n")
print(f"categorical columns: \n{cat_col}")

#Numerical Predictors Distribution
display(df[num_col].describe())

cat_table = pd.DataFrame(columns=['Column Name', 'Missing Values', 'Unique Values', 'Value

for col in cat_col:
    missing_values = df[col].isnull().sum()
    unique_values = df[col].nunique()
    value_counts = df[col].value_counts().to_dict()
    cat_table = cat_table.append({'Column Name': col,
                                  'Missing Values': missing_values,
                                  'Unique Values': unique_values,
                                  'Value Counts': value_counts}, ignore_index=True)

display(cat_table)
```

```
numerical columns
['age', 'Medu', 'Fedu', 'traveltime', 'studytime', 'failures', 'famrel', 'freetime', 'goout'

categorical columns:
['school', 'sex', 'address', 'famsize', 'Pstatus', 'Mjob', 'Fjob', 'reason', 'guardian', 'scl
```

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime
count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.0000
mean	16.696203	2.749367	2.521519	1.448101	2.035443	0.334177	3.944304	3.235443

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime
std	1.276043	1.094735	1.088201	0.697505	0.839240	0.743651	0.896659	0.998862
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000
25%	16.000000	2.000000	2.000000	1.000000	1.000000	0.000000	4.000000	3.000000
50%	17.000000	3.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000

	Column Name	Missing Values	Unique Values	Value Counts
0	school	0	2	{'GP': 349, 'MS': 46}
1	sex	0	2	{'F': 208, 'M': 187}
2	address	0	2	{'U': 307, 'R': 88}
3	famsize	0	2	{'GT3': 281, 'LE3': 114}
4	Pstatus	0	2	{'T': 354, 'A': 41}
5	Mjob	0	5	{'other': 141, 'services': 103, 'at_home': 59,...
6	Fjob	0	5	{'other': 217, 'services': 111, 'teacher': 29,...
7	reason	0	4	{'course': 145, 'home': 109, 'reputation': 105...
8	guardian	0	3	{'mother': 273, 'father': 90, 'other': 32}
9	schoolsup	0	2	{'no': 344, 'yes': 51}
10	famsup	0	2	{'yes': 242, 'no': 153}
11	paid	0	2	{'no': 214, 'yes': 181}
12	activities	0	2	{'yes': 201, 'no': 194}
13	nursery	0	2	{'yes': 314, 'no': 81}
14	higher	0	2	{'yes': 375, 'no': 20}
15	internet	0	2	{'yes': 329, 'no': 66}
16	romantic	0	2	{'no': 263, 'yes': 132}

6 Data preparation and transformation

In the data preparation phase, several steps were taken to transform and prepare the data for analysis. The following is a summary of the steps taken:

- Identification of categorical variables: All columns in the data frame were reviewed to identify the categorical variables. A loop comprehension was used to find the number of unique values corresponding to each column, and this information was used to decide on the data processing approach.
- Conversion of yes-no variables to binary variables: The code created a dictionary for binary mapping and applied it to all columns in the data frame that contained ‘yes’ or ‘no’ as the response.

- Transformation of predictors with 2 unique values: The code transformed predictors with two unique values into binary variables by mapping them to 0 or 1. This was done for variables such as `school`, `sex`, `famsize`, `address`.
- Creation of new predictors: To handle variables with more than two unique values, the code created dummy variables using the "get_dummies" function in pandas. The dummy variables were created for the `Mjob`, `Fjob`, `reason`, and `guardian` columns. The original columns were then dropped, and the dummy variables were concatenated with the original data frame.
- Combination of correlated predictors: The code combined the `Dalc` and `Walc` columns into a single "`Alc`" column to reduce correlation between the two variables, and combined the `Fedu` and `Medu` columns into a single `famEdu` column to capture the combined education of both parents. This helped to reduce data redundancy and the noise in the data set, and removed the dependency among predictors.
- Conversion of data types: The original data frame, which consisted of both categorical and numerical values, was converted into one that only consisted of numerical data types or uint8. This made it easier and more convenient for later variable selection and model development.

In conclusion, the data cleaning and preparation phase transformed and prepared the data for analysis by converting categorical variables into binary or numerical values, combining correlated predictors, and converting the data types into a more convenient format for analysis. These steps helped to ensure the validity of the results and facilitated later variable selection and model development.

6.1 EDA for Base Model Development

1. During the base model development, we transformed the response variable `G3` grade into a binary variable using the common pass/fail boundary as the standard. In an exam worth a total of 20 points, a grade of 12 or higher is considered passing (1), while anything below is deemed failing (0). We used this standard to classify if a student passes or fails consistently throughout the analysis. We plotted the grade distribution in its raw format (out of 20) and in its binary form, respectively, to visualize the general distribution of student grades.
2. Period 1 (`G1`) and Period 2 (`G2`) grades are strongly correlated with each other, as evidenced by the darker shade on the pairwise correlation plot.
3. Period 1 (`G1`) and Period 2 (`G2`) grades are highly indicative of the final grade (`G3`), as evidenced by the scatterplot displaying the relationship between `G1` and `G2`, with data points colored based on `G3`.
4. Both the bar plot based on importance score of the decision tree and the line plot demonstrate that `absences` and `failures` are crucial for predicting the final grades, with grades decreasing as absences and failures increase.

5. Medu (mother's education) and Fedu (father's education) are strongly correlated, so we combined Medu and Fedu into famEdu (family education).
6. Dalc (weekday alcohol consumption) and Walc (weekend alcohol consumption) are strongly correlated, so we combined Dalc and Walc into Alc (alcohol consumption).
7. A new correlation plot after removing and combining predictors shows that major dependencies among predictors have been resolved.
8. Categorical predictors like parents' jobs (Mjob and Fjob) also affect a student's grades. However, due to the difficulty in generalizing their impact across various categories, we did not focus on this factor during the base model development. Nevertheless, it remains an important aspect to consider in future analyses or more specialized models.

6.2 EDA for Progress Model Development

For the model that predicts progress between different tests, the insights used are different form the insights used for the base model because we have created new categorical variables as our responses: G1_G2 and G2_G3, which will be 1 if a student improves from the previous test and 0 if a student gets the same or lower score. I visualize the relationship between the predictors and G1_G2 and G2_G3 using barplots because a lot of the predictors are binary and other types of plots make it hard to observe the trend. Here are some findings:

1. Observing the barplots of the predictors versus G1_G2, I see that the predictors age, reason_other, Mjob other, romantic, and Fjob_services have some kind of relationship with our response variables: as their values change, the mean of the response variable will change relatively more than the plots for other predictors. Note: I did not include all the graphs in the code file because we have 43 predictors in the dataset.

2. Similarly, based on the barplots of predictors versus G2_G3, the variables Mjob_teacher, famrel, and reason_course seem to have the most effect on the repose variable as the predictors' values change.

7 Approach

During our data exploration, we recognized that predicting a continuous score from mostly binary predictor variables would be challenging. This is because binary variables typically have a non-linear relationship with the outcome variable, which can lead to inaccurate predictions and a lack of precision in capturing the nuances of the data. Thus, we decided to approach the research question with a classification method, which allowed us to develop the following three different models for different use cases:

- Logistic regression model: This model aims identify classify whether a student passes or fails based on relevant predictors.

- Base model: The base aimed to identify the most important quantitative and categorical predictors using the decision-tree and chi-square variable selection methods as well as identify the important interactions between key variables, if any.
- Progress model: This model examines the factors associated with a grade increase between the first and second grading periods.

In our classification models, we prioritize minimizing the false positive rate (FPR) as it is crucial to identify and assist students who genuinely require help with their grades. Misclassified a student who needs help as not needing it may not be as harmful as the opposite scenario, where a student who genuinely needs help is not identified and subsequently fails. Thus, our focus is on reducing FPR to minimize the chances of misclassified students in need.

8 Developing the model

8.1 Logistic Regression Model

Variables for the logistic regression model are chosen with Lasso. Since we want to keep the logistic regression interpretable, we found the 5 predictors that had the biggest absolute value and used those as predictors in a logistic regression. All the predictors used were found to be significant and the overall model is significant as indicated by low coefficient p-values and a low LLR p-value, respectively.

We then tried to see if the model would be appropriate for prediction by checking the accuracy and other metrics. We want to minimize FPR (students who are not doing well incorrectly classified as doing well), but it was difficult to find a threshold that kept FPR low, but maintained decent accuracy. Moreover, there we big differences between training and testing data ROC_AUC, as well as some other metrics, that suggest that the model might be overfitting some predictors.

This led us to explore other models that might be more useful for prediction.

Final logistic regression model:

$$p(\textbf{passing}) = \frac{1}{1 + e^{-(-1.3128\textbf{failures}-1.6826\textbf{schoolsup}-1.3522\textbf{Mjob\_teacher}+0.4451\textbf{Medu}+0.8009\textbf{sex})}}$$

8.2 Base Model Development

In this study, we utilized a manual categorization approach to convert the continuous response variable into a categorical response. We set a threshold and categorized the values greater than the threshold as “pass” and values less than or equal to the threshold as “fail”. This valid method allowed us to effectively predict student performance through classification.

Several techniques were used to develop and evaluate the model for the initial and variable selection phase and base model development. The following is a summary of the techniques used:

- K-Fold cross-validation and train-test split: K-Fold cross-validation and train-test split were used to assess the model's accuracy. In K-Fold cross-validation, the data was divided into K equal parts, and the model was trained and tested K times, with each part being used as the test set once. In the train-test split, the data was divided into a training set and a validation set, and the model was trained on the training set and tested on the validation set.

Two different feature selection methods were employed to identify the most important quantitative and categorical predictors.

For identifying the most important categorical predictors, the SelectKBest method with the chi-square test was used. First, only the categorical columns were extracted from the dataset. Then, the SelectKBest method was applied to choose the top 5 features based on their chi-square scores. The scores were stored in a pandas Series object and sorted in descending order to display the most significant categorical predictors.

- Decision tree search: To select the most important quantitative predictors, a Decision Tree Regressor model was fitted to the training data. We used `DecisionTreeRegressor` as opposed to `DecisionTreeClassifier` as we were treating the problem as a prediction problem on the first hand, followed by manual classification based on the pass/fail threshold. Feature importances were calculated, and used a horizontal bar plot was created to visualize the top 10 features with the largest importance values. Variables `failures` and `absences` were identified as the two most important features.
- Chi-square test variable selection: The chi-square test was used to select the most important categorical predictors by assessing the dependence between the categorical predictors and the target variable. `schoolsup` (extra educational support) and parents' job (mother's job `Mjob` and father's job `Fjob`) were important as relatively important predictors.

Given the limitations of a non-linear regressor (`DecisionTreeRegressor`) in identifying important features for linear regression, more than one decision tree was built to analyze and identify the important features. While the selected features often vary, the most important factors kept at each tree's top branch - features consistently selected as the most important ones regardless of trees built - were `failures` and `absences`. With this information, the base model was developed using only these 2 most important features, along with their interaction terms.

Four models was fitted using all possible combinations of predictors `failures` and `absences`: 1. $G3 \sim absences$ 2. $G3 \sim failures$ 3. $G3 \sim absences + failures$ 4. $G3 \sim absences * failures$

All achieved a 100% accuracy on the training and testing data. Accuracy was used as the only metric at this stage, as it provides a direct measure of the model's basic performance without giving weight to the cost of different errors. Later, more refined metrics will be used to weigh the priorities and relative costs of different types of errors.

We chose $G3 \sim absences * failures$ as the final base model as it captures the meaningful relationship of these two features and their combined effect on predicting student grades (`G3`). The term `absences:failures` has a P-value as small as 1.473151e-03. The extremely small P-value means that the interaction is statistically significant.The other models, although achieving an accuracy of 1.0 on the test set and the full dataset, do not account for the interaction between absences and failures. This could lead to a less accurate prediction of student grades in real-world scenarios where the effect of absences on student grades is not constant across students who have experienced different levels of failures. By including the interaction term, the last model better accounts for the combined influence of absences and failures on student grades, making it the best choice among the four models presented.

Final base model: $G3 \sim absences * failures$

$$G3 = 11.1824 + -0.0144(absences) + (-2.9952)(failures) + 0.1382(absences : failures)$$

8.3 Progress Model Development

The progress model seeks to identify attributes that indicate whether a student has improved between G1 and G2 and between G2 and G3.

We considered building on top of previous findings. However, using the same set of predictor from previous models does not yield convincing results:

1. Models are not significant: best LLR p-value achieved for the model for G1_G2 is around 0.01, which is far less than other models built-in this project.
2. Most of the variables are insignificant as the p-value for individual coefficient is high.
3. Measurements of classification accuracy is largely affected by the chosen threshold: even if we change the threshold by 0.01, there might be huge change in the observations that belong to FN and FP, so we need to balance between accuracy and recall.
4. After we balance the measurements and choose a fixed threshold, the highest accuracy/recall is less than 65% and the lowest FPR/FNR is larger than 35%
5. We also tried to calculate ROC-AUC, which is independent of the threshold, and the value is only around 0.65.

This part of the exploration is not shown in the code file, but based on this attempt, the direction of the prgress model is completely different from the base model: we will do new variable selection and new models.

To select the best predictors, Lasso is performed on the train dataset twice, because we need two new models for G1_G2 and G2_G3. We definitely don't want too many predictors in our models because giving too many advices to the students and families might make them feel overwhelmed. Therefore, I selected five predictors based on the absolute values of the coefficients after Lasso. It turns out that they match with our EDA but have small differences.

For the model built for G1_G2, there is one selected variable from Lasso that is different from the variables selected in EDA, so I build two models and compare them. Since our main concern is to help students and we don't want to leave out any student who actually needs help. In other words, predicting a student as able to improve when they cannot (False positives) will be harmful because the teacher will put effort in this student. Thus, the most important metric in the classification of improvement is FP. The model based on Lasso has lower FPR and other measurements are reasonable, so we will use progress_model_1_1 to predict whether a student improves from the first test to the second test.

We will have similar consideration for the two models builts for G2_G3. However, in this scenario, the model built using the predictors from Lasso has much higher recall (43.8) than model using predictors from EDA (25%) while the FPR of the Lasso model (41.2%) is much higher than the FPR of the model using EDA (28.8%). Since both models have an obvious disadvantage, we will select the final model based on what we concern more–FPR, so we will use progress_model_2_2 to predict whether a student improves from the second test to the final test.

Final Progress model: 1. $\mathbf{G1\_G2} = 6.3710 - 0.4348 * age + 0.9210 * reason\_other + 0.5279 * Mjob\_other - 0.4735 * romantic - 0.1581 * famrel$

2. $\mathbf{G2\_G3} = -2.4021 + 0.6112 * Mjob\_teacher - 0.4904 * reason\_course + 0.3824 * famrel$

9 Conclusions and Recommendations to stakeholder(s)

Despite the limitations of the current dataset, the following conclusions and recommendations can be beneficial to stakeholders:

Findings:

1. Absences and failures are important quantitative predictors in determining a student's grades (Base Model).
2. School support and parents' jobs have emerged as important quantitative predictors of students' grades in our analysis (Base Model).
3. The mother's job is a significant predictor of academic improvement, with mothers working in teaching and other professions showing a positive relationship with the improvement of grades from G1 to G2 and G2 to G3. On the other hand, mothers working in health care and civil service may have limited time and resources to support their children's education, while housewives may lack the necessary knowledge to help their kids

study, both leading to a stagnant predictor in students' academic improvement (Progress Model).

4. Students' reasons for choosing a particular school are related to their academic progress. The variable used to measure this consists of values such as proximity to home, school reputation, course preference, and others. Interestingly, the "other" category showed a positive coefficient in the model from G1 to G2. Conversely, the "course" category had a negative coefficient in the model for G2 to G3, suggesting that students who chose the school based solely on their interest in the courses were less likely to care about their grades. Due to data limitation, we were unable to determine what motivational factors play into the 'other' category. Potential reasons may include personal growth, knowledge acquisition, or other contextual factors.
5. The number of failures a student has experienced, whether they receive school support, mother's job (particularly if they are a teacher), mother's level of education, and sex are all significant when it comes to classifying whether a student passes or fails (low p-value, model itself has high LLR p-value).
6. Based on the logistic regression model, with each unit increase in failure, the odds of them passing this class are multiplied by 0.269. In other words, the odds of passing the current class decreases by 80% for each additional class they have failed in the past. While there are likely factors that play into both past class failures and current class performance, this suggests that this cycling of failing is hard to break.

Recommendation:

1. Develop and implement an early warning system that takes into account students' attendance and performance as well as the influence of school support and parents' jobs on academic success. This system should allow for close monitoring of students who may be at risk and provide tailored interventions to address their specific needs, ultimately enhancing their chances of success.
2. Encourage parental involvement in their child's education. Provide targeted support and resources to students whose mothers work in health care and civil service (or similar fields) and those whose mothers are housewives. This support could include after-school tutoring programs, access to educational resources, and involvement in school-based activities to enhance parents' understanding of how to support their children's education.
3. Schools should prioritize personal growth and knowledge acquisition opportunities, in addition to offering interesting courses. This can include non-academic challenges and curriculum planning, and can be highlighted in marketing materials, open house events, and other outreach efforts.

10 Model Limitations

The objective of the research project was to classify the performance of secondary school students in two Portuguese secondary schools. Unfortunately, the study contains a number of inference limitations. First, the dataset is from 2008, therefore it may not reflect the current educational scene. In addition, the study had 395 observations in total, which increases the danger of overfitting and makes it challenging to apply the results to a larger population. It is important to note that the data used in this analysis has no missing values. However, this may not reflect the typical cleanliness and completeness of real-world data sets, which often contain missing or inconsistent values and require more extensive cleaning and preparation. In future work, it will be important to take these considerations into account and apply appropriate data cleaning and preparation techniques to ensure the validity of the results. In addition, the study was done in Portugal, which may limit its applicability to other nations or areas.

Future study could benefit from increasing the dataset to include a larger sample size and a broader range of demographic and socioeconomic variables, potentially encompassing a variety of nations. This would allow for more rigorous studies and findings that may be applied to a wider range of population types. It may aid in the identification of gaps in educational results and provide insight into potential interventions that may be implemented to assist underrepresented populations. Inclusion of a larger range of indicators, such as cultural influences, mental health, and family environment, could further illuminate the intricate relationship between individual determinants and educational results.

11 GitHub and individual contribution

Github link

Individual contribution

Team member

Contributed aspects

Details

Number of GitHub commits

Julia Chu

Data Quality Check, Outliers, EDA, Github & Report Management

conducted initial exploratory data analysis along with visualization that provided insights for model development and variable transformation.

26

Victoria Shi

Data quality check & preperation; EDA; base model development; Github

Data quality check, preprocessing and preperation; visualization and EDA of base model; feature selection and variable interactions; report overview.

55

Yiru Zhang

EDA and progress model development, Github

explanatory data analysis for the progress model; visualization and variables selection; progress model development and assessment of prediction.

12

Yuyan Zhang

Logistic regression model and nested model(abandonned), Github

Basic logistic regression model, as well as exploration of a nested model, but could not find a non-trivial model that was accurate enough to then subdivide and make more models.

6

12 References

[1] 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, ISBN 978-9077381-39-7.

[2] Dipam7. (2021). Student Grade Prediction [Data file]. Kaggle. https://www.kaggle.com/dipam7/student-grade-prediction