CS 105 Project Report

Investigating social factors predicting academic performance in students

Team G1-T15

Bryan Lee (bryan.lee.2019@sis.smu.edu.sg)

Charlie Angriawan (cangriawan@sis.smu.edu.sg)

Emmanuel Oh (emmanueloh.2019@sis.smu.edu.sg)

Yuen Zhi An (zhian.yuen.2019@sis.smu.edu.sg)

1. Abstract

Our project aims to identify and explain social factors that significantly influence student academic performance. To achieve this, we first procured a dataset of student records capturing information about social factors as well as exam scores, which we referred to as a measure of academic performance. We then used exploratory data analysis (EDA) techniques to find significant trends and correlations between the factors and exam scores, and honed in further on those variables to remove correlated and dependent ones. Finally, we tested the factors by using them to build a Naïve Bayes classification model and testing for precision, accuracy, recall and F-score. We seek to examine and explain our observations in this report.

2. Introduction

Academic performance is often seen as an important marker for success in early life. In the absence of other metrics, it may be used to assess an individual's work ethic and personality. Furthermore, access to employment opportunities and higher education is heavily predicated on having a good academic record. While academic performance tends to be attributed to personal factors, many external factors play a vital role as well. Indeed, it is well-understood that students who come from families with a higher socioeconomic status tend to outperform those from the lower end of the spectrum. While this inequality is difficult to mitigate, there are many ways in which aid is granted to disadvantaged students, such as through scholarships, bursaries and counselling services.

In this project, we seek to identify the most crucial social factors that predict students' academic performance. This may be useful in helping to identify at-risk youth, or in constructing policies to support those who are the most disadvantaged within the community. Finally, for the layperson, it can help to broaden perspectives by offering different perspectives in understanding differences in student performance.

3. Dataset

The dataset we acquired was a CSV file containing the records of Portuguese students enrolled in secondary education (Cortez & Silva, 2008). It contained information about social factors such as parents' education, family size, quality of family relationships, health and school travel time, along with their academic performance in Mathematics. Most of the social factor metrics were in the form of categorical data while the dependent variable of exam scores was a numerical data field.

During the pre-processing stage, there was no need to impute any missing entries, as all students had their details completely filled. However, we did convert some categorical data into numerical representation. There were many categorical data fields which took only 2 possible values, such as sex, parental status (together or apart) and family size (greater or lesser than 3). In those instances, we created a binary numerical equivalent (0 or 1) to represent those fields instead. This was done to allow us to include these variables in analytical and data visualisation techniques for numerical data.

4. Approach

The first part of our approach was to perform exploratory data analysis on every entry of the dataset in order to ultimately identify the factors which had a significant impact on exam scores. We started off with a general five-number summary and box plot to get an idea of the general distribution of the exam scores G1, G2 and G3. Following which, we performed a univariate analysis by plotting out histograms for all categorical and numerical variables. This was able to help us visualise the distribution of students according to each of these features.

We then moved on to multivariate methods of data analysis, first by plotting bar graphs of mean G3 scores of students for distinct values of all categorical and numerical features. This allowed us to find any marked difference in mean scores across different categories or values of each feature, and hence glean whether those factors did or did not make a difference in G3 scores for students. However, it was also important at this juncture to make reference to the previous histogram distributions of student count. Certain feature values with very little students, for example, would be more prone to outliers changing the value of the mean significantly. From this analysis, we were able to shortlist a list of significant factors amongst all the features. Finally, we plotted a heatmap showing the correlation between each of these factors in order to identify those more closely correlated. We then used deductive reasoning to determine if those correlated factors were dependent on one another, and came up with ways to simplify our model and remove redundancies. Over the course of the EDA process, we made use of the Plotly graphing library (Plotly Inc., 2015) to plot and visualise our data.

In the second portion, we wanted to determine if the factors we identified had an impact on G3 scores. We chose to train a Naïve Bayes classification model according to the factors we identified, in order to predict whether a student would pass or fail the final exam. We used Scikit-learn (Pedregosa et al., 2011) predictive data analysis library in order to accomplish this. We trained the model using 75% of the student entries from the dataset and used the remaining 25% to test our model. Overall, the model performed well, with accuracy = 0.707, precision = 0.702, recall = 0.937 and F-score = 0.803.

5. Results and Discussion

Review

The evaluative results of the model indicate that the factors chosen indeed have an impact on academic performance. These factors have also been assessed in some depth in the EDA process. Broadly speaking, a student's academic performance can be predicted by their previous records, family and quality of parental relations, as well as socioeconomic mobility. However, it can also be affected by personal factors like intention and intrinsic motivation.

These findings are in line with common understanding and perception. Many scholarships and student aid efforts are aimed at those financially less well-off, and there are many programs directed at helping children from less stable families. However, one compelling finding is that the most marked difference in student performance comes from the distinction between those who intend to pursue further education and those who do not. What we can gather from this is that personal motivation is a strong factor in determining academic outcomes. This is significant because it implies that it is possible to make

education more beneficial for students if we are able to inculcate in them a willingness to learn and further their knowledge. It also suggests that it would be a worthwhile endeavour if educational institutions focus developing systems and measures that encourage passionate and invested learners.

Limitations

The primary limitation of our project has to do with the nature of the data collected. Much of the data was qualitative in nature, and relied more on self-reported values rather than quantitative measurements. When comparing between multiple individuals, there is no way to fully standardise such qualitative responses. The data would have been better suited for statistical data analysis if, for example, family relationships were measured by time spent with family, or health was quantified by certain biological metrics. Apart from this, the fact that the study was conducted in the context of Portuguese schools might make our findings less generalisable to student populations as a whole, as there could be more significant cultural factors at play that were unaccounted for in this study.

Further studies

There are many ways to build on the findings identified from this study. For one, we could extend the scope of the study in several dimensions. It would be interesting to see if these findings hold true for students in different age groups, or those hailing from different cultures and nationalities. Different studies controlling for such factors could integrate these existing findings into a broader contextual understanding.

Another possibility would be to change the nature of the data collected. Since an individual's social factors tend to be more fixed and immanent, it might be a better idea to conduct a longitudinal study between groups of people based on differences amongst the variables identified. Such a study would also be more equipped to explain the causal relationship between these social factors and academic performance.

6. Conclusion

Overall, we were able to achieve our intended goal of applying data analysis techniques to identify social factors that had an impact on academic performance in students. In the process, we were careful to avoid logical fallacies in mistaking correlation for causation and understood the limitations of the data we were working with. We made an effort to minimise redundancies and tried our best ensure that the factors we identified were independent. On the whole, the project was able to develop our technical expertise in data analysis and provide insights on the important ways in which an individual's environment and social context can influence their academic performance.

7. Contributions

Bryan Lee

Searched for useful plotting libraries and functions for exploratory data analysis, as well as visual design and formatting of plotted graphs. Looked for statistically significant trends in EDA process.

Deliverables

- Codes: Dataset pre-processing, Exploratory Data Analysis
- Report: Dataset, Approach

Charlie Angriawan

Came up with the proposal for a Naïve Bayes classifier as a predictive model and tested out performance of the model in predicting different metrics.

Deliverables

- Codes: Introduction, Dataset pre-processing, Classification model
- Report: Introduction, Dataset, References

Emmanuel Oh

Explored potential causal explanations for observed results from EDA and offered solutions to minimise redundancy between observed factors. Edited overall writing and finalised structure of the report.

Deliverables

- Codes: Exploratory Data Analysis, Insights, Classification model
- Report: Introduction, Results and Discussion

Yuen Zhi An

Experimented with possible methods of EDA and data visualisation. Refactored code to be clean and modular. Ensured that project progress was up to date with timeline.

Deliverables

- Codes: Introduction, Insights, Conclusion
- Report: Results and Discussion, Conclusion

7. References

Cortez, P., & Silva, A.(2008). *Using Data Mining to Predict Secondary School Performance In A. Brito and J. Teixeira Eds.* Porto, Portugal: EUROSIS.

Inc., P. T. (2015). *Collaborative data science*. Retrieved from https://plot.ly

Pedregosa, F., Varoquax, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E.(2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.