Interactive Application To Explore Impact of Lifestyle and Family Background on Grades

Tang Haozheng, Timothy Lim, Wu Yufeng
Singapore Management Institute, Singapore

ABSTRACT

With the constant emphasis of education and the impact it has on a person, there is hence important to know what affects a student's grades. In this paper, with the focus on lifestyle and family background of a student, we aim to identify which factors have an impact on a student's grades. This will be done through developing an application on R Shiny that will serve as a platform to perform exploratory data analysis, clustering analysis and predictive analysis which will have elements of interactivity for users to explore the data set at a deeper level. These tools will then be used to determine the factors that have a significant impact on grades.

1. INTRODUCTION

In the past many years, there has been an emphasis on education around the world because of the impact it has on a person, be it in terms of employment opportunities and quality of life. It is hence important to know what are factors that affect one's academic performance. While there are many factors that can impact a person's academic performance, family background and one's lifestyle are two of the larger factors.

Since there are many sub-factors in family background and lifestyle choices, the motivation of this study is to look deeper at these sub-factors to see which are the factors that have a greater correlation in the impact on a student's grades. More specifically, this study aims to study the correlation between each factor and a student's grades, as well as aiming to build a model that can accurately determine the academic performance of a student. From the findings, targeted help may be administered to students in these specific areas attributing to poor grades in school, therein helping them have a higher chance of a better future.

2. MOTIVATION AND OBJECTIVES

Despite similar studies on the impact of lifestyle and family demographics on grades being done in past works, there is a general lack firstly a visualization of the data set where the data is mostly discussed through words and numbers.

Our work was hence motivated by the absence of data visualization and interactivity of variables identified for analysis in the impact it has on student grades. This would be important in allowing users to explore the data to gain insights on it. To fill this gap, our goal is to create an application that allows users to be able to not just see the final results of which variable has the highest impact on grades, but to allow the exploration of the data set in an interactive way, to have a better sense of the data set, as well as to compare the impact between each variable.

The application will include the following features in order to fulfill the above objectives:

1) To provide data exploration through bar charts and histograms to study the distributions and proportions of each variable
2) To provide clustering analysis of the data points to see which students are similar in terms of the surveyed variables
3) To perform predict analysis to identify which variables or a combination of variables have a significant impact on a student's academic performance
4) To provide interactive functions to perform data exploration

3. LITERATURE REVIEW

Through performing a literature review, we found various studies that explore how one's lifestyle affects a student's academic performance. However, as lifestyle being such a generic term, many variables can be considered lifestyle. As such, majority of these studies focus on physical activity and eating habits which are not variables in our study.

However, from these papers went through, it was observed that these either only provided tabular data on the data set, or basic charts for data visualization. Most studies also do not provide much exploration of the data set. For example in a study by Li and Qiu (2018), the focus was largely on finding out what impacts a student's academic performance, but provided only an overview of the data set used. No data

visualization or any interactivity was available for users to have a closer look at the data.

4. DESIGN FRAMEWORK

The application will allow users to review three main types of analysis - exploratory and confirmatory data analysis, clustering analysis and predictive modeling. These being the main analysis types will be split by a separate tab in the application. Within each of these analysis types will involve different charts, which will include some functions of interactivity to allow the user to explore the data further.

In designing the individual charts, we first look at the analysis to be performed, then choose the appropriate chart type for visualization.

In exploratory data analysis, three charts will be utilized. Firstly, for the purpose of reviewing the distribution and proportions of each variable, a bar chart was used where the percentage proportions were included in the charts for clarity in cases where the proportions are similar. Secondly, to compare inputs between each variable against the average grade, the violin chart was used, with an overlay of the box plot. Third, a correlation plot was utilized to show the correlation amongst the variables, including the target variable, average grade. With many variables, an interactive correlation plot will also be available for selection should the user wish to view specific correlation scores. This is because with the many variables in the data set, putting in numbers into each cell might cause the chart to become messy. In each of these three charts, the user is able to decide the data set and the variable to be explored.

For the clustering analysis,

Finally for the predictive model analysis, the decision tree modeling will be used. Here, three charts will be presented - the decision tree diagram, a bar chart to show the level of importance of the variables, and a confusion matrix. The decision tree will the decision making process based on the inputs of each variable to reach the conclusion of the grades. However, knowing this flow may not be enough. Hence, a bar chart to show the importance of each variable was included, to rank the variables according to the impact it has on a student's grades. Besides this, it is also important to evaluate the performance of the model. Hence, the confusion matrix was included to allow the user to review the accuracy of the prediction of the grades. Here, the user is able to firstly select the data set, the number of bins for the grades, as well as the variables to be studied further.

The demonstration of using each chart will be discussed further below.

5. DATA PREPARATION

The data set used in this study was taken from Kaggle, which contains two csv files, student-mat.csv and student-por.csv. The data was obtained from a survey of Portuguese students aged 15 to 22 from two schools on family demographics, lifestyle factors, along with their grades in mathematics and Portuguese language. The data set contains other variables that can be broadly classified into family demographics and lifestyle factors.

Below are the variables included in the data set:

##	Variable name	Description
## 1	school	Student's school
## 2	sex	Student's sex
## 3	age	Student's age
## 4	address	Student's home address type
## 5	famsize	Family size
## 6	Pstatus	Parent's cohabitation status
## 7	Medu	Mother's education
## 8	Fedu	Father's education
## 9	Mjob	Mother's job
## 10	Fjob	Father's job
## 11	reason	Reason in choosing this school
## 12	guardian	Student's guardian
## 13	traveltime	Home to school travel time
## 14	studytime	Weekly study time
## 15	failures	Number of past class failures
## 16	schoolsup	Extra education support
## 17	famsup	Family educational support
## 18	paid	Extra paid classes within course subject
## 19	activities	Extra-curricular activities
## 20	nursery	Attended nursery school
## 21	higher	Wants to take higher education
## 22	internet	Internet access at home
## 23	romantic	In a romantic relationship
## 24	famrel	Quality of family relationships
## 25	freetime	Free time after school
## 26	goout	Going out with friends
## 27	Dalc	Workday alchohol consumption
## 28	Walc	Weekend alcohol consumption
## 29	health	Current health status
## 30	absences	Number of school absences
## 31	G1	First period grade
## 32	G2	Second period grade
## 33	G3	Final grade

##	Variable name	Description
## 1	school	Student's school
## 2	sex	Student's sex
## 3	age	Student's age
## 4	address	Student's home address type
## 5	famsize	Family size
## 6	Pstatus	Parent's cohabitation status
## 7	Medu	Mother's education
## 8	Fedu	Father's education
## 9	Mjob	Mother's job
## 10	Fjob	Father's job
## 11	reason	Reason in choosing this school
## 12	guardian	Student's guardian
## 13	traveltime	Home to school travel time
## 14	studytime	Weekly study time
## 15	failures	Number of past class failures
## 16	schoolsup	Extra education support
## 17	famsup	Family educational support
## 18	paid	Extra paid classes within course subject
## 19	activities	Extra-curricular activities
## 20	nursery	Attended nursery school
## 21	higher	Wants to take higher education
## 22	internet	Internet access at home
## 23	romantic	In a romantic relationship
## 24	famrel	Quality of family relationships

## 25	freetime	Free time after school
## 26	goout	Going out with friends
## 27	Dalc	Workday alchohol consumption
## 28	Walc	Weekend alcohol consumption
## 29	health	Current health status
## 30	absences	Number of school absences
## 31	G1	First period grade
## 32	G2	Second period grade
## 33	G3	Final grade

- Student's school School
- Student's sex
- Student's age
- Student's home address type Address
- Family size
- Parent's cohabitation status
- Mother's education
- Father's education
- Mother's job
- Father's job
- Reason in choosing this school
- Student's guardian
- Home to school travel time
- Weekly study time
- Number of past class failures
- Extra education support
- Family educational support
- Extra paid classes within course subject
- Extra-curricular activities
- Attended nursery school
- Wants to take higher education
- Internet access at home
- In a romantic relationship
- Quality of family relationships
- Free time after school
- Going out with friends
- Workday alchohol consumption
- Weekend alcohol consumption
- Current health status
- Number of school absences
- First period grade
- Second period grade
- Final grade

Further data cleaning and manipulation was done, such as creating dummy variables for non-binary categorical variables, as well as having a new variable average grade, which will be used as the target variable. The data set was also split into four smaller data sets based on school and subject taken. This is reduce any variability due to the school the the student attends or the subject the student takes. Part of the interactivity will include the selection of the data set, to allow the user to explore the data sets of each combination of school and subject.

6. THE APPLICATION

6.1 Exploratory Data Analysis

6.1.1 Distribution and Proportions

The bar chart and histogram will be used for users to see the distribution of proportions within each variable. This gives a clear breakdown and proportion of the inputs in each variable. Nominal data variables will utilize this bar chart, where the percentage proportion will be indicated so that comparisons can be done not out of estimation, but with clear data. Here, the user is able to firstly select the data set of interest, then the variable to review can be selected next.

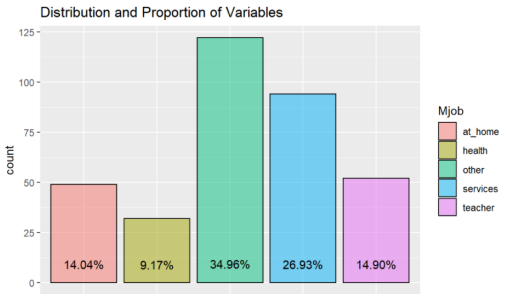

Figure 1: Bar Chart

As for the histogram, it will be used for the variables with a larger range. Due to larger range, the user is able to change the number of bins on top of selecting the data set and the variable. The histogram will also be interactive in a way where the user is able to hover the cursor over the data point to see the count.

INSERT PICTURE

6.1.2 Comparison of Results

The violin graph is the chart the user is able to use for comparing the difference between each variable with respect to the average grade. From the chart, besides having the boxplot where the user is able to see the upper quartiles, lower quartiles and mean, the user is able to see the distribution of grades through the width of the violin graph. The chart also provides statistical indications, where if any two inputs are significantly different in the results, the p-value will be shown. The user is able to select the data set to review as well as the specific variable to study compare.

INSERT PICTURE

6.1.3 Correlation

For reviewing the correlation, the user is able to select one of two charts. The first chart is static where no interactivity is implemented. This chart is to give the user a broad overview of the correlations between each variable, with average score being at the bottom left. The correlations are presented through coloured ellipses, where red indicates a negative correlation while blue indicates a positive correlation, with the degree of correlation indicated by the gradient of colour.

The other plot that the user can look at is one with interactivity. This allows the user to hover the cursor over each box, where a tool tip will appear, indicating the two participating variables as well as the correlation value. This allows the user to see specific correlation values between variables. Similar to the static correlation chart, the correlation value is indicated by the colour and gradient.

6.2 Cluster Analysis

6.2.1 Dendrogram and Heatmap

In cluster analysis, the aim is to group all the data points into smaller groups based on their traits, which are given by the individual inputs for each variable. From here, the user is able to explore and see which students are similar in terms of their demographics. In this section, a dendrogram along with a heatmap can be studied. Each row in the heatmap represents one data point, or in this case one student, and they are arranged based on the inputs of each variables by clustering them together. The clusters created can be observed from the right of the chart, indicated by the different colours.

In terms of interactivity, the user is able to select the data set for reviewing, then from the heatmap and dendrogram generated, hovering the cursor over each cell will show the details of each cell, and zooming in can be done by dragging out a rectangle.

INSERT TWO PICTURES

6.2.2 Comparison of Clusters

Below the dendrogram and heatmap, a series of violin charts with boxplots will be presented. This chart allows the user to see the distribution of grade scores for each variable, sorted by the inputs from each cluster formed previously. From the chart, the user will be able to compare the grades obtained by students based on which cluster they are in for each input. The user will be able to select the data set to be studied as well as the variable to explore.

6.3 Predictive Modeling

For predictive modeling, the decision tree method was used. Three charts will be shown here - the decision tree, a bar chart based on the level of importance for each variable selected, and finally a confusion matrix. Here, the user is first able to select the number of bins to be created for the grades obtained by the students. Then, the user can select the variables to be studied further. The decision tree will automatically update and show how each variable represented by the circles and the logic indicated on the lines, and finally the grade bin shown by a box.

Figure 2: Decision Tree

Below the decision tree is a simple bar chart to show the importance of each of the variables selected, sorted by descending order. Here, the user is able to immediate know which variable impacts grades the most out of the variables selected.

Finally, a confusion matrix is presented, with the X-axis being the actual grade bins, and the Y-axis being the the predicted grade bin. This chart shows the user how well the model is able to predict the grade bin with the selected variables.

Figure 3: Importance Bar Chart and Confusion Matrix

7. CONCLUSION AND FUTURE WORK

Being able to demonstrate the data visualization and implementation of interactivity within the application allows users to further explore data sets in a deeper way. Beyond just knowing what impacts a student's grades, the user is able to explore the demographics of the students as well as to study certain groups of students.

Although our application was able to pick out some of the more significant variables, the results are not as ideal as we wanted as the level of accuracy is still not optimal. Hence, further work could be done to firstly implement and compare other methods of analysis and modeling. Secondly, to have a larger data set to obtain more accurate results.

Implementing another function where users can input their own data sets would be useful, to allow this application to be applicable to any user who has a data set that they would like to explore and perform analysis on.

Acknowledgements

The authors would like to thank Prof. Kam Tin Seong for his guidance and support in the completion of this project.

References

[1] Li, Z. and Qiu, Z. 2018. How does family background affect children's educational achievement? Evidence from Contemporary China. The Journal of Chinese Sociology. (Oct. 2018).