Statistic Model to Analyze Student’s Performance

Xuchuan Zheng, Sungki Park, Prashant Sharma

# Introduction

Academic success is important because it is strongly connected to the positive outcomes we value. Student who are academically successful and with high levels of education are more likely to get employed, have stable and better job, have more employment opportunities than those who with less education. Especially, academically successful adolescents have higher self-esteem, have lower level of depression and anxiety, and are less likely to abuse alcohol and engage in substance abuse.

In out final project for Data 603 - Statistical Modelling with Data, we have tried to develop a model to analyze the impact of various demographic and social factors on the performance of students. Academic performance, though it is not the only factor but is one of the crucial factors in shaping a student's future. To get into a good collage/university, student must score grades in school, a good college can lead a better future and economic stability. So, to secure good grades, getting into a great school is enough? Is there something more than a great school that can help a student to perform better? Do the social and demographic factors plays any role in student's performance? In our project we are trying to answer these questions.

# Methodology

### Data Source

We found our datasets for our regression analysis from UC Irvine Machine Learning Repository which is a website is providing a collection of databases, domain theories, and data generators for the analysis of machine learning algorithms. The data attributes include student grades, demographic, social and school related feature. The two datasets we downloaded are provided regarding the performance in two distinct subjects: Mathematics and Portuguese language. Since we didn’t have to analyze our data by the subject, we combined those two datasets into one dataset.

ADD MORE?~~~

### Variable Explanations and Data Assumptions

The dataset we are working with is collected during 2005-2006 at 2 Portuguese schools for Mathematics and Portuguese subject. In Portugal, the secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system. There are several courses that share core subjects as the Portuguese Language and Mathematics. A 20-point grading scales is used, where 0 is the lowest grade and 20 is the highest score. During the school year, students are evaluated in three periods and the last evaluation G3 corresponds to the final grade. There are closed questions related to several demographic (e.g. mother’s education, family income), social/emotional (e.g. alcohol consumption) and school related variables (e.g. number of past class failure) that were expected to affect student performance.

In our dataset, we have many categorical variables with ordered values. CONTINUES!~~~~~~

There are 649 rows instances and 30 features in the dataset. The following table is a complete list of variables used in our modeling process.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Scale** | **Type** |
| school | student's school | binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira | Qualitative |
| sex | student's sex | binary: 'F' - female or 'M' - male | Qualitative |
| age | student's age | numeric: from 15 to 22 |  |
| address | student's home address type | binary: 'U' - urban or 'R' - rural | Qualitative |
| famsize | family size | binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3 | Qualitative |
| Pstatus | parent's cohabitation status | binary: 'T' - living together or 'A' - apart | Qualitative |
| Medu | mother's education | numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education | Qualitative |
| Fedu | father's education | numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education | Qualitative |
| Mjob | mother's job | nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other' | Qualitative |
| Fjob | father's job | nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other' | Qualitative |
| reason | reason to choose this school | nominal: close to 'home', school 'reputation', 'course' preference or 'other' | Qualitative |
| guardian | student's guardian | nominal: 'mother', 'father' or 'other' | Qualitative |
| 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 | Qualitative |
| studytime | weekly study time | (Numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours | Qualitative |
| failures | number of past class failures | numeric: n if 1<=n<3, else 4 | Qualitative |
| schoolsup | extra educational support | binary: yes or no | Qualitative |
| famsup | family educational support | binary: yes or no | Qualitative |
| paid | extra paid classes within the course subject (Math or Portuguese) | binary: yes or no | Qualitative |
| activities | extra-curricular activities | binary: yes or no | Qualitative |
| nursery | attended nursery school | binary: yes or no | Qualitative |
| higher | wants to take higher education | binary: yes or no | Qualitative |
| internet | Internet access at home | binary: yes or no | Qualitative |
| romantic | with a romantic relationship | binary: yes or no | Qualitative |
| famrel | quality of family relationships | numeric: from 1 - very bad to 5 - excellent | Qualitative |
| freetime | free time after school | numeric: from 1 - very low to 5 - very high | Qualitative |
| goout | going out with friends | numeric: from 1 - very low to 5 - very high | Qualitative |
| Dalc | workday alcohol consumption | numeric: from 1 - very low to 5 - very high | Qualitative |
| Walc | weekend alcohol consumption | numeric: from 1 - very low to 5 - very high | Qualitative |
| health | current health status | numeric: from 1 - very bad to 5 - very good | Qualitative |
| absences | number of school absences | numeric: from 0 to 93 | Quantitative |
| G1 | first period grade | numeric: from 0 to 20 | Quantitative |
| G2 | second period grade | numeric: from 0 to 20 | Quantitative |
| G3 | final grade | numeric: from 0 to 20, output target | Quantitative |

### Modeling Plan

# Result

### Variable Selection Procedures

### Main Effects Individual T-tests:

### Hypothesis Statement for Individual T-tests

### Hypothesis Statement for Individual T-tests (Interaction Terms)

### Interaction Term T-tests:

### Hypothesis Statement for ANOVA Test:

### Multiple Regression Assumptions

1. Linearity Assumption
2. Independence Assumption
3. Normality Assumption
4. Equal Variance Assumption
5. Multicolinearity Tests
6. Influential Points and Outliers
7. Interpreting Coefficients
8. Prediction

# Conclusion

# Discussion