Socio-economical and Educational Impacts on Student Performance

By: Suzanne Chung

CKME 136 – Winter 2017

Github: https://github.com/emsuzq/capstoneproject

Introduction

Students academic performance can be affected by multiple variables, from family, health, lack of interest, social activities etc. the list can go on. Are we able to narrow down what actually affects the students' performance in school?

A survey was conducted on students who are enrolled in Math and Portuguese class who were questioned on numerous social and educational factors. The dataset used included the students' grades throughout their enrollment in these courses. Based on the data given, what are the significant social and educational impacts that affect the students' final performance?

I will be evaluating all attributes to see the significance importance it has on the students' grades by using explanatory analysis to perform preliminary analysis. I will create a Bayesian network that will in turn provide causalities of the students' final grades.

Literature Review

Is Alcohol Consumption Associated with Poor Academic Achievement in University Students?

This report is observing students at the University of Gloucestershire, UK and alcohol consumption specifically. The data collected had five alcohol consumption measures: length of time and amount consumed during recent drinking occasion, frequency of alcohol consumption, heavy episodic drinking and drinking problems. Three educational measures were used: the importance of achieving good grades, students' appraisal on their performance against their peers and students' actual mark. The research was broken down to demographics specifically sex and age and to see which measure was strongly associated with academic outcomes and how academic achievements are associated with alcohol consumption. The results from regression analysis demonstrated that males were positively associated with all alcohol consumption measures. The students' actual mark was not associated with any alcohol consumption measure. Overall, Alcohol consumption showed negative associations with academic performance.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3843305/

Prediction Accuracy of Academic Performance of Students using Different Datasets with High Influencing Factors by Jai Ruby and Dr. K. David

This research focuses on two different datasets (Arts College and UCI) that analyzes the prediction accuracy of academic performance of students while using Multi Layer Perceptron (MLP) classification algorithm. The main attributes used in both datasets were fathers' job, mothers' job, travel time, first period grade, extracurricular activities, second period grade and previous class performance. The datasets were split into two-thirds training sets and one-third testing set and found that Arts College had a prediction accuracy of 64.5% while UCI was 91.42%. It was concluded that the attributes chosen under the UCI dataset are high influences on students' performance.

http://www.ijarcce.com/upload/2016/february-16/IJARCCE%2017.pdf

The Effects of Socioeconomic Characteristics of Students on Their Academic Achievement in Higher Education by Ekber Tomul and Gokhan Polat

A study was conducted with 691 undergraduate students to see if their academic performance was heavily influenced by school-related and socioeconomically factors. The focus was on families socioeconomic status such as: parents' educational status, family income, the settlement where the family lives, the status of the fathers' workplace, number of siblings and the educational background of the student. The study uses correlation and regression analysis in order to determine whether socioeconomic characteristics are related and have an effect on academic achievement. The study found that there were no significant relationship between family's socioeconomic status despite numerous research and articles that state otherwise. The results show that the type of high school that the student was enrolled in has a strong relationship with the students' performance.

http://pubs.sciepub.com/education/1/10/7/

Bayesian Network with Examples in R by Marco Scutari and Jean-Baptiste Denis

Information and reference book used to help build my Bayesian network and understanding causality inferences within the network.

Dataset

In this research project, I will be using two datasets provided by UCI Machine Learning Repository, which can be found here: https://archive.ics.uci.edu/ml/datasets/Student+Performance. Both datasets represent the students' performance in two different classes, Math and Portuguese. The data collected has 32 attributes that reflect family socioeconomic status, demographics and school-related features.

The attributes are listed below, provided by UCI Machine Learning Repository:

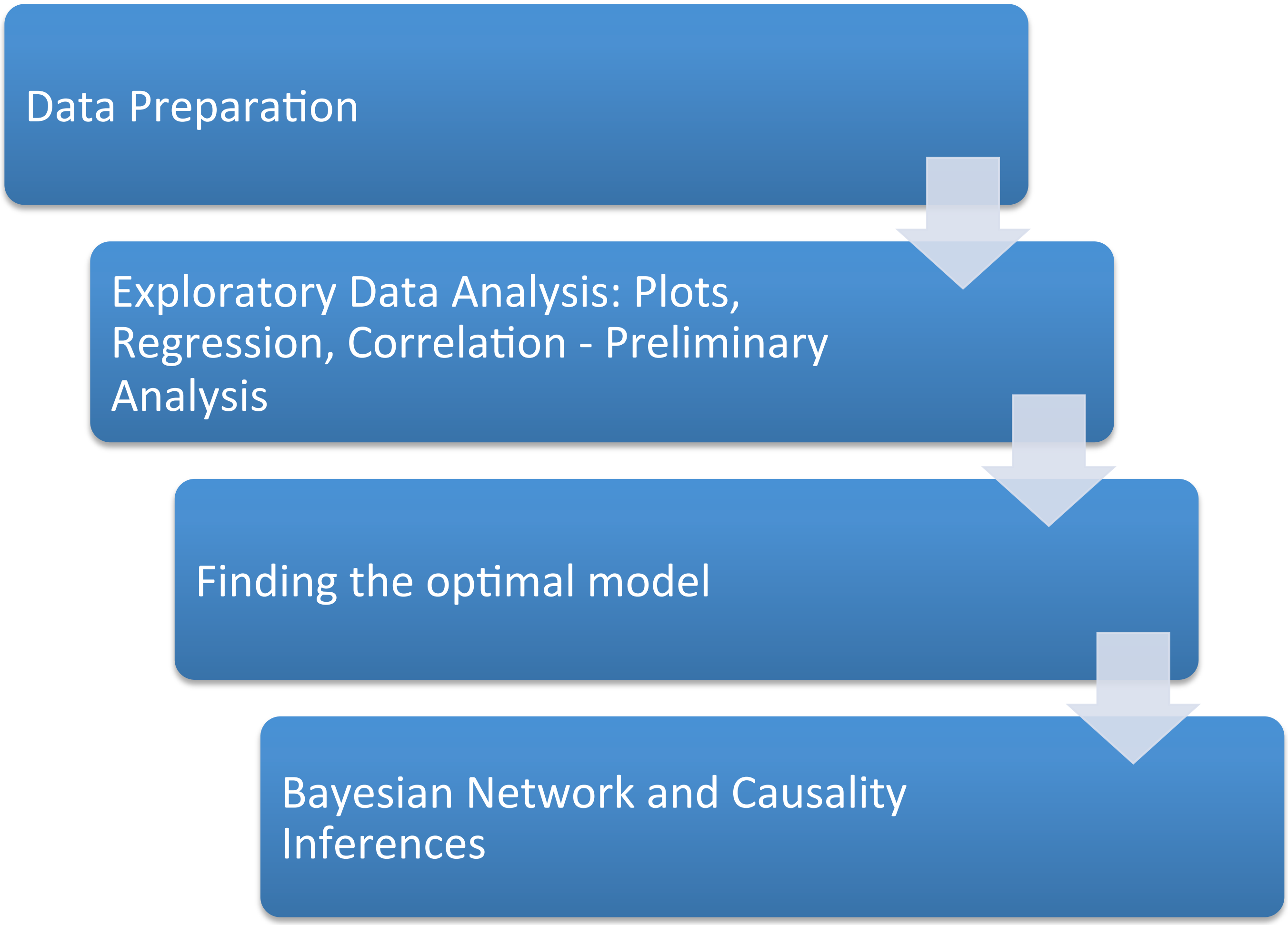
- school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
- sex - student's sex (binary: "F" - female or "M" - male)
- age - student's age (numeric: from 15 to 22)

- address - student's 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 - parent's cohabitation status (binary: "T" - living together or "A" - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- Fjob - father's 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 - student's 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)

All attributes will be used in this research project.

Based on the literature reviews, I found that my focus would be on family socioeconomics and attributes that are believed to be the causes and significant impacts on a student’s grade. G3, also known as the students’ last grading in the class will be referred to as the students final overall grade in the course.

Approach



Data Preparation

Exploratory Data Analysis: Plots, Regression, Correlation - Preliminary Analysis

Finding the optimal model

Bayesian Network and Causality Inferences

Step 1: Data Preparation

There are two datasets that represents students enrolled in Math and /or Portuguese courses. I will be filtering data to students that are enrolled in both courses. To ensure that that I've captured the students corrected, we merge the datasets based on these attributes below:

- School
- Sex
- Age
- Address
- Family size
- Parents cohabitation status
- Mother's education
- Father's education
- Mother's job
- Father's job
- Reason to choosing specific school

- Nursery attendance
- Internet availability

All attribute values in the new combined dataframe have been converted to all numeric values.
Github link:
https://github.com/emsuzq/capstoneproject/blob/master/R%20Code/Code/Data%20Prep.R

Step 2: Exploratory Data Analysis: Plots, Regression, Correlation - Preliminary Analysis

I've used various bar graphs and box plots to show a simplistic relationship between the some of the attributes. Based on some articles I've read (listed in Literature Review section), I was curious if some of the relationships would stand out in this dataset, some included:

- How many of the students' parents attended higher education?
  - Did this contribute to their family support towards student involvement in school?
- Are there differences between the parents behavior based on the student enrollment in a course (supporting math class over Portuguese class)?
- Relationship between study time and the students' past failure rate.

At this point, decided to only pursue students enrolled in the math class based on the results observed. I've updated the dataframe for all attributes to be numeric.

In addition to the exploratory data analysis, I performed the regression analysis and correlation matrix:

- Applied linear regression to the math class dataframe
- Analyzed the regression to evaluate the significant attributes
- Performing correlation matrix to remove the most correlated attribute in the data.

Github link:
https://github.com/emsuzq/capstoneproject/blob/master/R%20Code/Code/Exploratory%20Data%20Analysis.R

Step 3: Finding the optimal model

In order to find the best fitted attributes to build an optimal model, I used the follow method:

- Ranking Feature by importance
  - Train the math class data where the attributes that were found highly correlated was removed
  - Used kknn method for training
  - VarImp function is used to provide an estimate of the attribute importance

In addition to the method above, I used a feature selection method to compare my results with the Ranking Feature approach:

- Train the same math class data used in Ranking Feature method
- Run the Recursive Feature Elimination function on the dataset, requesting only the top 20 attributes that are required to build an accurate model
- Results produced listing top 20 attributes that I will be using in Step 4.

Github:
https://github.com/emsuzq/capstoneproject/blob/master/R%20Code/Code/Finding%20Optimal%20Model.R

Step 4: Bayesian Network and Causality Inferences

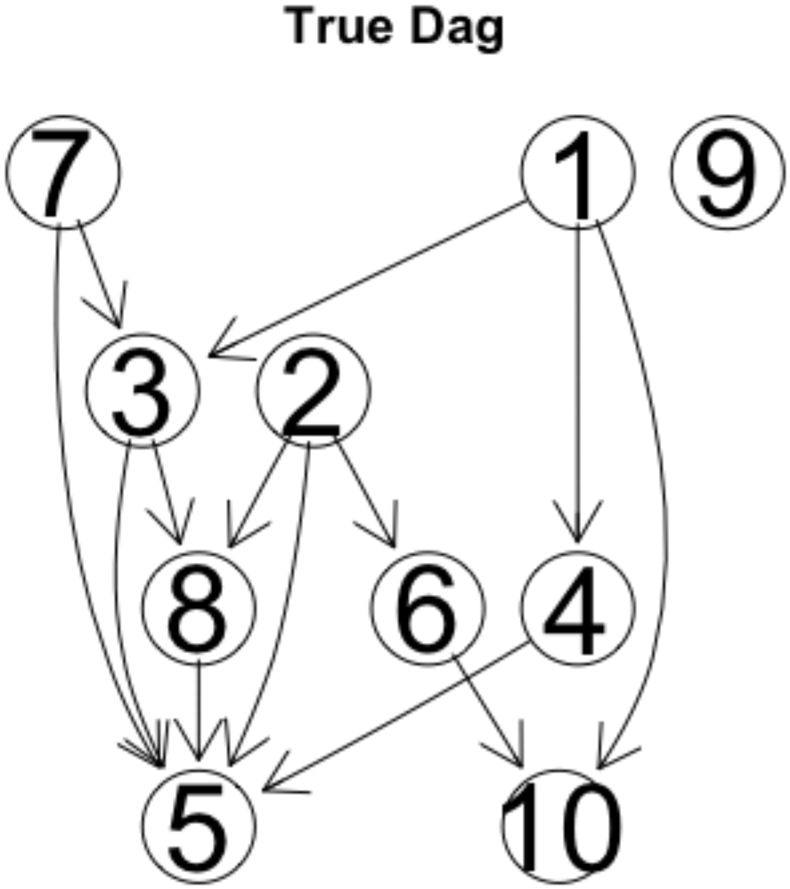
Based on the attributes that I found in step 3, I was able to create a Bayesian network using two different methods to see if there were any differences in the plot.

Secondly, to answer my research question, which attributes effects the students' grades, ie what causes them to perform poorly or better?

The last method I used is the causality inferences on G3 (final grades).

Github:
https://github.com/emsuzq/capstoneproject/blob/master/R%20Code/Code/Bayesian%20Network%20-%20Math%20Class.R

Results



	Attribute
1	G1
2	Absences
3	Failures
4	School Support
5	Age
6	Romantic
7	Higher
8	Guardian
9	Activities
10	G3

In the network diagram above, #10 that is labeled G3 for students' final grades, we can see that the students' initial grades and their romantic involvement have an effect on the students' final grade. Viewing this from a socio-economical standpoint and articles that were reviewed, students who are involved in romantic tend to have lower grades.

To find the causality in this model, I used the Gaussian CI Test and set my alpha to 0.05 based on the dataset and the regression model I ran in my preliminary analysis.

The network identifies that the students' initial grades also affect the final grades. We can assume that G1 is used either to motivate or reduce the interest in performing in the class. After the results from the Bayesian Network, I looked into Exploratory Analysis which can be found:

Final Grades vs Initial Grades:
https://github.com/emsuzq/capstoneproject/blob/master/R%20Code/Code/EDA%20between%20Final%20Grades%20and%20Initial%20Grades%20in%20Math%20Class.pdf

Final Grades vs Romantic:
https://github.com/emsuzq/capstoneproject/blob/master/R%20Code/Code/EDA%20relationship%20between%20final%20grades%20and%20romantic%20in%20Math%20Class.pdf

The relationship between Final Grades and Initial Grades based on the EDA performed after our results tells us that the students who perform lower aren't motivated to improve their performance, therefore continue to perform the same. This shows how students' are affected negatively.

The relationship between Final Grades and Romantic based on the EDA performed after our results tells us that majority of the students who receive a grading of 12 or higher, have no relationship involvement and move to be high performers. Students who score below average said yes to having a relationship, between 50%-60% of the students.

Conclusions

The analysis demonstrated the cause of the students' performance. Although, I expected more attributes to be involved, this is still a fair evaluation based on the data provided (about 400 students surveyed). We are able to find the most optimal attributes using a Ranking Feature that would help build an estimated optimal model. Form the optimal model; we were able to create a network to identify the causality of the students' final grades.

Some improvements would be to focus on specific attributes. There 33 different attributes that showed no real relationship with one another. Survey a larger quantity to have a more accurate result.