CS 573 Final Project Proposal

October 31, 2016

1 Basic Info

A. Project Title: Student Performance Visualization

B. Names, e-mail addresses, GitHub id as below

- Chong Zhou, czhou2@wpi.edu, zc8340311
- Han Jang, hjiang@wpi.edu, MissHanJ
- Haitao Liu, hliu5@wpi.edu, lht1949

C. The link to the project repository Click project link here or go to the url: `https://github.com/zc8340311/CS573FinalProject`

2 Background and Motivation

Education is a key factor in affecting people career and achievement in the later-on life. So it's interesting to understand how the students' background and personal effort impact the learning process. We have a real-world dataset that include such factors, such as student grades, demographic, social and school related features. For the learning measurement, the dataset consists of the grade of Math and Portuguese, where are the targets that we could predict from the student background, personal efforts and other social and school related features. The reason that we choose this dataset is that educational data mining is very interesting by exploring the factors that contributes to the success of learning. If such factors are well understand and discovered, we can help the students learn more effective and efficient. Another reason is from our personal experiences, Both learning a foreign language and learning the math needs some "talents". But the talent needed for language and math seems different, since math and language are processed in different area of our brain . Based on the above motivation, we would like to explore the relation between students' background and their grades in math and Portuguese and also to verify our assumption that the "talents" for math and Portuguese are different.

3 Project Objectives

The high level goal of this project is to build informative and elegant visualization that helps us understand the student performance and related features.

A. Discover the trends and relationships

- The trend between our target and student factors
- The features with high correlation with each other
- Whether the students' performance is consistant

B. Discover the outliers

- Any student stands out? or departure from the general trend? For example, for certain kind of student's background, he/she should achieve some good or bad grade, but his/her performance depart from our expectation. Why this outliers happen? We expect our visualization will offer some evidence or hints to discover and help explain outliers.

C. Discover the feature Importance We could split the data according to some

4 Data

We obtain the student performance data set from the UCI machine learning data repo: Click here or go to the url: `http://archive.ics.uci.edu/ml/datasets/Student+Performance`

This data incudes student achievement in secondary education of two Portuguese schools. The data features consist of student grades, demographic, social and school related features. There are Two datasets are students performance in two distinct subjects: Mathematics (mat) and Portuguese language (por).

Attribute Information as below:

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2 sex - student's sex (binary: 'F' - female or 'M' - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: 'U' - urban or 'R' - rural)

5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')

10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')

11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13 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)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if $1 \leq n < 3$, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

31 G1 - first period grade (numeric: from 0 to 20)

31 G2 - second period grade (numeric: from 0 to 20)

32 G3 - final grade (numeric: from 0 to 20, output target)

5 Data Processing

The data contains many ordinary features, we need to translate these ordinary features to dummy variables. The data is stored in CSV files. There are no missing values. So there is no extra data cleanup process.

6 Visualization Design

6.1 Trends discovery

1. parallel plot to show the trend relationships between features. 2. stream plot to show the trend of grads for different class of students.

6.2 Outlier discovery

3. Apply multiple projection method to discover the best projection method and according to the projection method to discovery the

6.3 Feature Importance Discovery

7 Must-Have Features

1. Allow zoom-in and zoom-out for the trends' and outliers' detail. 2. Allow customized manipulation of data. 3. Multi-dimension visualization plot which could involve multiple features while offering as much information as possible. 3. Apply distinguishable color map to represent the categories.

8 Optional Features

9 Project Schedule

10 Reference

UCI data repo