Data Science Project Proposal

Performance Analysis and Prediction of Secondary School Students

Contributors:

Oleg Rubenchik - +35796047152, oleg.rubenchik25@gmail.com, ID: U234N0588
Role: Programmer & Concept Dev.
Artem Boichuk - +35796075503, artemboychuk100204@gmail.com, ID: U234N1450
Role: Researcher & Model Dev.

Background and Motivation

What made us select this project is the search for success factors in adolescents --- What contribution does each factor bring to the academic performance of teenagers and young adults in the secondary education? We believe that adolescents have never been in such a need for guidance that they are in this day and age, with technology, worldwide conflicts, and other challenges overwhelming them. We hope that this modest project, no matter how insignificant on the bigger scale, contributes to the wellbeing of these young adults.

Project Objectives (Hypothesis)

Our dataset contains all kinds of observations: besides the obvious factors, like the health of students, we want to explore aspects that may be less obvious, such as the romantic interest of a student, or the duration of a student's commute to school. Is it possible to generate a synthetic dataset that will accurately represent real trends and behaviors to train a future model on it?

Data Sources

For our project, we will be using a Student Performance dataset collected from two Portuguese schools, containing student data regarding their performance in Maths and Portuguese. Our model will predict student achievements, with Math being a representative criterion for the exact sciences, such as Chemistry, Physics; Portuguese being a criterion for determining student success in humanitarian subjects, such as languages, History, etc. Despite the fact that the dataset contains only 396 mathematics/650 portuguese students records, with only 39 students matching both, our priority was to work with real data as a primary source. On top of that, our selected dataset has as much as 30 attributes being observed, which we were especially happy about.

Dataset can be found @ https://archive.ics.uci.edu/dataset/320/student+performance

Data Structure

We have two CSV tables: Math course and Portuguese language course.

It has the following 30 attributes:

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

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)

Data Preprocessing

In order to analyse the data we will need to clean it first. We will perform overall profiling of the dataset to see what else needs attention. We will need to analyse each attribute and determine how useful and important it for our purposes. Possibly, some data will not be in preferable data type,so we will need to change it without losing its value.

Data Analysis

We will use EDA to identify different insights regarding the dataset. Then, a synthetic dataset will be generated from dataset that we will have processed. This generated dataset will be used in our model, granted we won't face any major setbacks or unexpected challenges. In any case, we aim to produce solid work around this dataset that will greatly improve our skills.

Tools and Libraries

Tools:
Python
Excel
ChatGPT

Python Libraries:
NumPy
Pandas
Matplotlib
Scikit-learn
Pathlib

Expected Challenges

We expect to train a model on the synthetic dataset generated from the data we found. However, it might be challenging to finish such a task within the set 13 weeks, considering the workload from other courses and activities. Other than that, we may also encounter difficulties while generating the said synthetic dataset.