Data science project

Student profiling based on moodle log data

Team A3
Mariliis Malahhov, Jane Õunaid

Task 2. Business understanding

Background.

Reimo Palm, a lecturer of Tartu University, seeks to get information about students' success and possibly identify struggling students in the Programming course. The course is generally taken by first year informatics students, but it is also open to students from other specialities in the Faculty of Science and Technology. The course teaches Python and gives a general idea of programming constructs and ideas. It doesn't require any previous programming experience.

The course requires taking moodle quizzes, submitting homeworks, submitting practice session exercises and a team project. In the project students make a functioning Python program. There are also two compulsory tests that require solving two programming exercises. The tests also contain a quiz which has to be passed in order to pass the test in general. There is also a final exam, which has a similar structure to the tests. During the course it is also possible to earn extra points by solving extra exercises.

At the beginning of the course, students who felt secure in their Python skills had the chance to take a pre-exam during the second week.

Goals.

Our goals are to identify students who might be struggling based on their early activities and possibly predict the final grades.

Success criteria.

We will consider the outcome successful if we manage to achieve previously defined goals. We hope that our project helps to assess how well the course is structured and possibly make the course better.

Inventory of resources.

Human resources: two students, Reimo Palm for guidance

Hardware: two university issued standard laptops

Software: Python libraries

Data: Moodle log files of about 350 students, which contain viewing materials, submitting homeworks, solving quizzes; about 1 million entries.

Requirements, assumptions, and constraints.

Project deadline is 13th of December. Raw data contains personal information (names and personal identification codes) which will be replaced so no one can be identified. Requirement for acceptable finished work is achieving at minimum the first goal.

Risks and contingencies.
Currently not relevant, we hope it will stay that way.
Terminology.
We consider the terms relevant to our project such as “student”, “grade”, “test”, “homework” etc commonly understood. For our first goal, we define “struggling student” as someone who got low test scores and homework scores during the first 6 weeks but finished the course with any grade.
Costs and benefits. No financial costs, probably no financial benefits we can currently foresee.

Data mining goals and success criteria.
We already have the data so we don’t need to mine it.

Task 3. Data understanding

Gathering data.
Data already exists. It is in the .xlsx format. The files open without problems so data is accessible.
Data requirements. The data is in the .xlsx format, for analysis it must be converted into .csv format.
Selection criteria. Data contains a file with grades of every student. It also contains files for each practice group. These files contain student activities, such as answering tests and presenting homeworks, viewing forums etc. Data we select: relevant data is the grades, for goals 1 and 2.

Describing data.
Data consists of moodle log files of 342 students; it contains grades and student activities. Source: moodle, Reimo Palm. Number of cases: 342 students.

Number and description of fields.
Grades file: 111 columns. They contain students’ names, ID codes, e-mail addresses, test scores for each week, total result of test scores, homework scores for each week, total result of homework scores, project description’s assessment, score for first version of project, score for second version of project, total score for the project, scores for first test quiz part, scores for test exercises, scores for retaken tests and their parts, scores for second test quiz part etc, scores for exam parts, total exam scores, points for extra credit, final course score with extra credits, final score.
Practice group file with moodle activities
Columns are: time, full name of student, affected user, event context, component, event name, description, source, IP-address.

Data suitability: for our project, the grades file seems to be suitable. The log files contain a whole lot of information, part of which seems irrelevant; at this time we think we are not going to use them.

Exploring the data.
Our first goal is to identify students who might be struggling based on their early activities. The grades file contains some fields that are not relevant or contain personal information. We deleted the personal information columns and gave each student an ID, starting from 1. The missing values were denoted by "-", we replaced them with "NaN". We also deleted columns with only empty values or only "NaN" values. Some students chose to take the pre-exam and a successful result meant they didn't participate in the rest of the course. Therefore we deleted the rows that contained these students. We were left with 325 students. From a cursory look there seemed to be a few students who dropped out of the course after the first few weeks. Per our definition of struggling student, those who dropped out are irrelevant and their data might interfere with the results.

Verifying data quality.
Irrelevant, our data seems to have the necessary quality for achieving our goals.

Task 4. Planning the project

Detailed plan for the project.

1. Creating a plan for the project, think about the methods and models to use. (10 hours, both members)
2. Do the preliminary data cleaning in xlsx format described in task 3, change the .xlsx format into .csv format. (about 2 hours, Jane)
3. Create a blank Jupyter Notebook.
4. Write down the project description and goals into the notebook. (0.5 hours, Jane)
5. Import data-make sure to get column names properly, change commas in floats to dots. (0.5 hours, Jane)
6. Data cleaning: remove students who dropped the course after the first few weeks. (1 hour, Jane)
7. Select relevant data for the first goal. It should contain IDs, test scores and homework scores from the first 6 weeks, final grades.
8. Create a plot for final grades to get an initial overview. (0.5 hours for the previous and this task combined, Jane)
9. Make a new column with average of first 6 tests, another column with average of first 6 homeworks, third one with the average scores of both tests and homeworks. (15 min, Jane)

10. Create a plot: x-axis is averages of test scores, y-axis is final grades. Create a second plot for homeworks. Create a third plot for averages of both test and homework scores. (2 hours, Jane)
11. Analyse the plots, see if there is correlation between initial scores and final grades. (0.5 hours, Jane)
12. Write analysis for first goal results. (1 hour, Jane)
13. Select data for second goal: sum of all quiz scores, sum of all homework scores, final grades. (0.5 hours, Mariliis)
14. Convert numerical grades into categories: A,B,C,D,E,F. - 0,5 hours, Mariliis
15. Convert grades into binary features with one-hot encoding. (0.5 hours, Mariliis)
16. Split the dataset into training set (80%) and test set (20%).
17. Train a decision tree classifier and/or random forest classifier on the training set. (2 hours hours for the previous and this task combined, Mariliis)
18. Use the model(s) on the test set. (1 hour, Mariliis)
19. Compare the accuracy of the models with real grades (assess the models). (1 hour, Mariliis)
20. Possibly make plots for outcomes for visual overview. (0.5 hours, Mariliis)
21. Write analysis of outcome. (1 hour, Mariliis)
22. Write a final report. (1 hour, both members.)
23. Make sure all the code is uploaded to Github repo.
24. Make a video for the project presentation. (3 hours, Mariliis)
25. Make a poster for the presentation. (3 hours, Jane)