IDS2021 - Project B1

Student profiling based on Moodle log data

Team members:
Matevž Zorec
Farnaz Baksh

29th November 2021

Repository:
https://github.com/mbz4/IDS-PROJECT-2021-B01

This report modeled using CRISP-DM methodology, available here: Embracing the Data-Mining Process

Version History

Version	Date	Modifed By	Summary
1.0	29.11.2021	Matevž Zorec Farnaz Baksh	Create, structure and input the sections understood
2.0	30.11.2021	Farnaz Baksh Matevž Zorec	Followed the CRISP-DM format and input all sections
2.1	14.12.2021	Farnaz Baksh	Input ****Update* for a few sections

Table of Content

Business Understanding

Business goals

Background

The University of Tartu, Institute of Computer Science offered a computer programming course using Moodle for the Fall 2020 semester.

This course was designed to cater primarily to computer science students and students of other curricula, who need to complete basic programming. It emphasized teaching core programming concepts and did not require prior knowledge of programming.

Course participants (students) followed a carefully planned curriculum, where they:

a) *Attend lectures* (we assume not on Moodle)
b) Take Moodle based quizzes / tests
c) Submit Homeworks on Moodle
d) Submit Practice Session Exercises on Moodle
e) Submit team project files on Moodle
f) Attend final exam on Moodle
g) Received bonus points during exercise sessions for additional tasks

Moodle logged all of the students and staff activity during this time. This activity log dataset spans roughly 1 million entries from up to 350 students who were attending this course, as well as staff members, who were managing this course.

From this gathered activity information, the organisation now aims to get more in-depth information about students' performance in the course. Our contact person within this organization is Reimo Palm, who provided us with the data, descriptors and key goals.

Goals

Our main goals are:

1) to identify common activity patterns of the participants (students) in order to create 'typical student profiles'
2) predict course participants (students) final grades in this course
3) *(bonus)* identify course participants (students) who may be struggling based on their respective early generated Moodle activity logs.

Success criteria

The organisation is interested in gaining insight on students' performance indicators.
We assume their agenda is, amongst other things, to reinforce decision making in course planning. We have successfully met outlined goals, if we:

a) develop a method to extract all goal relevant data & indicators, including:
    i) recognize student performance issues
    ii) recognize if a student struggled using their earliest submissions
    iii) common student activity patterns
    iv) accurately identify student activity patterns as 'typical student profiles'
b) develop a model training framework for predicting student final grades in this course
c) develop a model training framework for identifying students struggling based on earliest submissions

Assessing the situation

Inventory of resources

Data: A total of 22 documents was given by the organisation:

- 20 spreadsheets of Moodle log data files
- 1 grade results dataset spreadsheet
- 1 description document

Data Science Students (developers): Matevz Zorec, Farnaz Baksh
Lecturers or TA (for guidance): Reimo Palm, Victor Pinheiro, Meelis Kull

Hardware: Dell XPS 13 (i5, 8GB), HP Omen 15 (R7 5800H, 16GB, RTX 3060)
Software: Jupyter Notebook, Python packages (pandas, sklearn, numpy)

Requirements, assumptions, and constraints

The project deadline is fast approaching and considered as a mission critical requirement. Within the 2 weeks, us, the developers, must produce results and report our findings. This puts immense stress, strain and fear of not meeting our own expectations.

Our initial assumption was that all data was to be provided in English, but the dataset is in Estonian, adding significant cognitive load, despite the enclosed translation in a separate description file. We understand the grade assessment process will take this fact into consideration.

We are encouraged not to disclose the data nor students personal information and will take steps to anonymise them in our final submission. This process will cost additional time and cognitive effort.

Risks and contingencies

Given that there are only two members on the team with two laptops, there is the risk of hardware failure and the need to backup the work via cloud storage.

A key risk is that the dataset analysis is too time-consuming or that perhaps we make misleading assumptions. A key contingency in this case is an abundance of sanity checks as well as communicating with the organisation contact person.

Terminology

Common terms: Student, Grade, Exam, Test, Homework, Project
Technical terms: Dataset, Python, Model training, Data analysis, Jupyter Notebook

Typical student profiles - a student who falls into one of the categories of the grading scheme (A, B, C, D, E, F) in the gradebook data.

Costs and benefits

No financial cost nor financial benefit.
In retrospect, we stand to gain a multitude of benefits where we will learn basic data science and machine learning concepts and apply them in a structured project setting.

Data Mining (DM)

DM Goals

Models:

a) model framework for predicting student final grades in this course
b) model framework for identifying students struggling based on earliest submissions

Reports:

a) understand and recognize student performance issues from moodle log data
b) recognize in data analysis if a student struggled
c) understand what the key features are in b)
d) report on how the data is structured overall
e) understand how the course structure is represented in data and student activity

Presentations:

a) develop and implement key graphical representations of experimental data analysis findings
b) find experimental data representations in data analysis to extract key information
c) implementation of algorithms that help extract key correlations and causal relationships

Processed datasets:

a) anonymity of all dataset
b) develop a framework for assessing student behaviour and classifying different students into categories of 'typical student profiles'
c) extract train and test data for each key goal
d) apply gained experience and adequately prepare extracted train and test datasets
e) apply further processing on datasets with elastic net regularization to further identify relevant features

DM Success criteria

We will consider our data-mining as successful if we:

a) evaluate trained models using AUC score and achieve a score of 0.80 or above from generated predictions,
b) have representations of findings from exploratory data analysis,
c) have an understanding of how the course structure impacted student performance.

Data Understanding

Gathering the data

We have acquired all relevant data for the project and have spent some time assessing its structure. All data files open as they should in Jupyter Notebook in the format we specify and we were able to run basic commands.

Data requirements

The data needs to be prepared such that key features are extracted. Parsing needs to be accurate and parsing functions may only be developed using copycat toy datasets.

****Update:* Our team spent almost two days writing a python program to translate all xlsx files from estonian to english. This helped us drastically by reducing the cognitive strain looking at a language we did not understand.

Data availability

Data is readily available using the organisation's Owncloud service and a backup was made on the team's google shared drive for easy access. Each team member also downloaded the data on their laptop to get started.

Selection criteria

Our current assumption (pre thorough analysis) is that we will likely limit the dataset to the first month and use that in training a model that predicts students final grade.

Describing the data

The log data is from the Computer Programming course in Fall 2020.
The Moodle log data files is split into 20 spreadsheets (xlsx format) according to lab groups. They total to a size of ~54MB and 1 million entries. Each log file contain 9 columns which is in Estonian.
We were also given 1 grade results dataset spreadsheet (xlsx format) of size 141KB with 342 entires and a description (txt format) document which houses the translation of the columns in English and a description of each column. This document also outlines the grading scale, and some additional summary of the data shared.

Exploring data

* Aeg (Time)
* Kasutaja t�isnimi (User full name)
* M�jutatud kasutaja (Affected user)
* S�ndmuse kontekst (Event context)
    - �lesanne: Assignment submission tool without autochecker
* Komponent (Component) and S�ndmuse nimi (Event name)
* Kirjeldus (Description)
* P�ritolu (Origin)
* IP-aadress (IP address)

It is likely most of these will not be key features in the final model training solution, however, our team is still exploring the data.
**** Update:* Our team did use most of these features to help us achieve our goals.

Data quality

The data we have acquired appears to be auto-generated and with little or no error.
We may intervene and rename some of the provided Estonian column names with translated English counterparts for better exploration and understanding.
Our goals only require students data thus we will not need the log entries related to the lecturers or TA of the course. Overall, it has enough data to get started with achieving our goals.

Project Plan

1) Student profiling using Moodle log data is not a new idea. It has been done before therefore we will spend time researching how others solved this problem before us. As a result we will try to collect as much inspiratory and exemplary material as we can, keeping it concise and clear to understand. **[Matevz: 6hrs] [Farnaz: 6hrs]**

2) In hoping to narrow down the key solutions we found relevant in our research from task 1, we will again examine the data we received to confirm what steps to take in task 2. It is highly likely that we will begin by creating an abstractable jupyter script that will parse thru a copied, subtracted test dataset. The scraped info will be passed thru pandas as a dataframe and examined. We hope to gather key insight from this task about how the data we have on student moodle log activities compares to findings from others in available research we covered. **[Matevz: 6hrs] [Farnaz: 6hrs]**

3) In task 2 we finished researching methods, and already started working with the dataset. In this task we will have to attempt to draw our first real assumptions about the data we have. Based on these assumptions and experience from research in tasks 1 & 2, we will setup a set of functions, each containing a script-standardised call for a different model training setup set. This setup will not be final as we have not yet decided on what the features and labels should be, but have already realized what the general outcome is (provided in project assignment instructions: predict student grade (0-100 ⇒ A-F))... The reason for putting the idea of using different models is that we want to perform hyper parameter tuning - we know already we will be relying almost solely on Random Forest Classifier, Decision Tree, TensorFlow or XGBoost for this project, based on research we have already conducted. **[Matevz: 7hrs] [Farnaz: 7hrs]**

4) At this point we have abstractified, parsed and setup several model functions to run the dataset thru. We must continue analysing the datasets and draw more conclusions. Additionally, for this task we should prepare the data for training, then train the models.

After training the models, we also need a function which outputs for each student the predicted outcoming grade score or perhaps even the associated final grade. **[Matevz: 7hrs] [Farnaz: 7hrs]**

5) We must analyze the results, confirm our initial assumptions and check with task 1 research if our method and results are alright (sanity check). Then we must prepare a concise report and presentation video and submit it. **[Matevz: 12hrs] [Farnaz: 12hrs]**

Methods and Tools

.

+ Research
+ Parsing
+ Data cleaning
+ Analysis
+ Multiple model training
+ Abstraction
+ Elastic net analysis
+ Hyperparameter tuning
+ Models: Random Forest, Decision Tree, XGBoost, TensorFlow
+ Dataset toy set extraction
+ Common sense
+ Sanity checking
+ Feedback from TAs