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

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. They formulate a dataset of the moodle log files of activities of about 350 students or participants in the course, which sums a total of one million entries. With all this data, the organisation now aims to get more statistical information about students' performance in the course.

Goals

The main goals are to identify common activity patterns of the participants to create 'typical' student profiles and to predict their final grades in the course. Additionally, a bonus goal would be if we can identify students who may be struggling based on their early moodle log activities.

Success criteria

The organisation desire to gain insight on students' performance issues which will help them to make decision on structuring the course better for future participants. If we successfully achieve our main goals then we would be able to clearly see which moodle log data are relevant as key indicators for future data points.

Predict student performance issues

a) Recognize if a student is struggling based on just a few early submission scores
b) Prepare a map of most common indicators of students that are struggling and need help

If we were to successfully train a model which can predict student outcome and performance as well as provide a map of most common indicators of student performance decrease, we would have succeeded in meeting our business success criteria.

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

Hardware: Dell XPS 13 (i5, 8GB), HP Omen
Software: Jupyter Notebook, Python libraries/packages

Requirements, assumptions, and constraints

Project deadline is fast approaching, within the next 2 weeks which puts a lot of strain on the developers and the fear of not producing excellent results or reaching the goals.

Given that the data (at least column names) is in Estonian, this require additional time for our team to get familiar and process the raw data. The dataset also contains students' personal information which cannot be shared publicly.

Risks and contingencies

Given that there are only two members on the team with two laptops, there is the hurdle 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, Mr Reimo.

Terminology

Common terms: Student, Grade, Exam, Test, Homework, Project
Typical student - ???
Technical terms: Dataset, Python, Model training, Data analysis

Costs and benefits

No financial cost or financial benefit.
This will cost us both several nights of sleep and deteriorate further our overall mental and physical well-being after an already exhausting breakfree non-stop grind semester.
In retrospect, we stand to gain a multitude of benefits: we learned basic data science and machine learning concepts and even applied them in a structured project setting. This is amazing and fantastic value.

Defining your data-mining goals

Data-mining goals

Understand data structure, hierarchy, file organization. In this phase, we have to extract a set of features we identified as key in training the most successful model possible.

Data-mining success criteria

If we evaluate the accuracy and AUC score of our trained model to be acceptable enough, we can also consider that our data mining has been successful.
Please, follow this given structure and cover all these aspects in your report. Consult this PDF-file with a chapter on Embracing the Data-Mining Process for more information on each of the deliverables. Keep the report concise and feel free to state that some aspect is not relevant in your project. If your project is not meant to benefit a 'business', then please specify who will benefit from the project and perform business understanding from their perspective. For instance, this could be either one or multiple individuals, organizations, or societies. Please focus on the goals that you plan to directly contribute to, not on the generic goals (like making the world a better place).
The report of task 2 should be 400-800 words.

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.

got acquainted with the moodle logging format as well as discussed in brief about how moodle themselves process log data and analyze it with the stock model they ship.

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.

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.

We also have a complete student gradebook file and we will use it in training and verification of our trained models.The semester started on 31.08.2020, the students were added to Moodle on 1.09.2020, the semester officially ended on 31.01.2021. Therefore we will limit the dataset to the first month of the semester in order to train a model that predicts just the student grade based on early activities on moodle.

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.

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: 7hrs] [Farnaz: 7hrs]**

List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

+ 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