Project title: Open University Project
Team members: Karl Martin Teras, Dmitry Vyakhirev

Business Understanding

Identifying your business goals

Background:
Open University is a public research university in the UK focused on distance learning. The university had been collecting data about students' interactions with the virtual learning environment during the period of 2013-2014. The data have then been made available as the Open University Learning Analytics Dataset (OULAD).

Business goals:
We hope our findings will inform UT students on what strategies to use in order to succeed in an online course.

Business success criteria:
We will consider our project as successful as long as we can conclude that peer students find it useful based on the feedback we receive during the poster session.

Assessing your situation

Inventory of resources: The dataset consists of seven csv files and has information about 22 courses, 32593 students, their performance results, and logs of their interactions with the course materials (10,655,280 entries) [1]. We are going to use Python, Jupyter Notebook, Google Colab, and Git to work on this dataset. We can use the forum on Piazza to get help from instructors and peer students in achieving the goals for this project.

Requirements, assumptions, and constraints: The dataset is available under CC-BY 4.0 license which requires us to cite it. A PDF version of the poster is due December 16 at noon. The poster session will be held on December 19 in Paabel.

Risks and contingencies: To achieve our goals, we are going to look for patterns in students' interaction with the virtual learning environment and use that to train a model. However, this data may not be enough to get a model with good accuracy. In this case, we will try to include other factors such as demographics in the analysis.

Terminology:
VLE = virtual learning environment
module = course
module presentation = course offering

B presentation = presentation starting in February
J presentation = presentation starting in October

Costs and benefits: The project requires 30 hours of work per student and can potentially benefit the students coming to the poster session.

Defining your data-mining goals

Data-mining goals:
First, we are going to determine if there is a correlation between the time students started working on their homework and the grade. Our second goal is to find patterns in the way students interact with the course materials and their performance. Finally, we will try to make a model that predicts a student's grade based on the time they start working on homework and what materials they are going to use.

Data-mining success criteria:
We can tell if the analysis was successful if the first two goals are accomplished and we were able to find a model that gives better results than a random choice.

Data Understanding

The data has already been collected by the Open University and the whole analysis requirements of the data was built around the data that was provided.

Gathering data

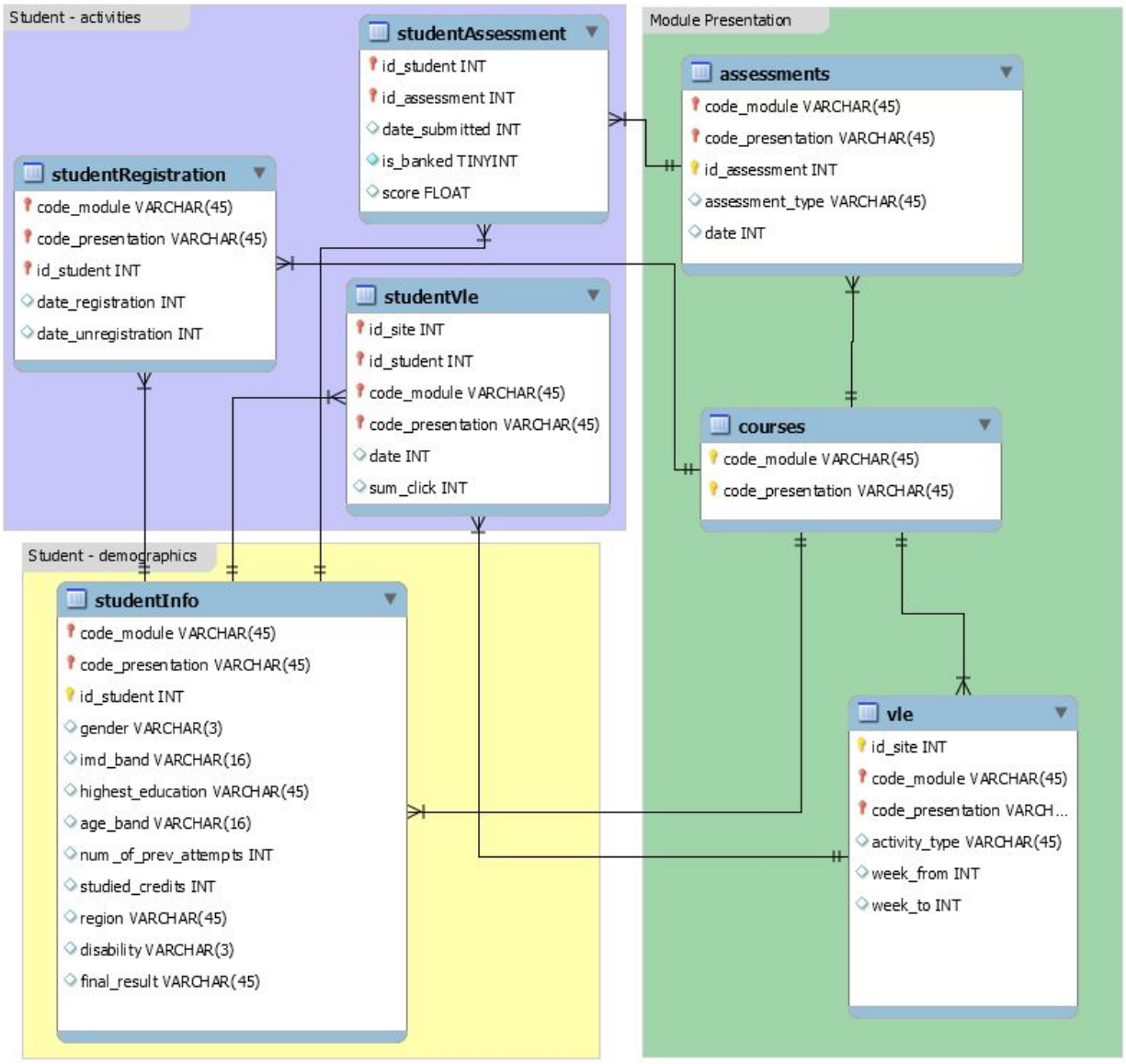
Outline data requirements:
The data needs to contain timestamped usage of the visual learning environment for students. This data needs to be linked to a student and have information about the learning environment resource type. If data is about different courses and the student is registered to different courses then the VLE usage data should be linked to a specific course. To assess the results of the students, the points for assessments are required. The timeframe of the data should be at least one course of data but a year (two courses) or multiple years is preferred.

Verify data availability:
The data exists in the format that was described in the requirements.

Define selection criteria:
We will be using the data collected by the Open University about their virtual learning environment. It is a database containing the information described in the requirements with

tables for courses, vle collection data, students, assessments with grade data and the links between the tables..

Describing data



Source: https://analyse.kmi.open.ac.uk/open_dataset#description

studentInfo (32 593 rows) - Contains info about the student

- code_module (AAA, BBB, GGG, etc): module code
- code_presentation (2013J, 2014B, etc): presentation code (year and month)
- id_student: internal student id
- gender (M or F): gender of the student
- region (East Anglian Region, Scotland, etc): region the student is based from
- highest_education (HE Qualification, A Level or Equivalent, etc): The education level

- imd_band (Multiple deprivation index)
- Age_band
- Num_of_prev_attemps
- Studied_credits
- Disability
- final_result

studentRegistration (32 593 rows) - links studentInfo and courses with a date:

- code_module (AAA, BBB, GGG, etc): module code
- code_presentation (2013J, 2014B, etc): year and month
- id_student (11391, etc): internal student id
- date_registration (-1, -40, 5, etc): The number of days since the start of the course
- date_registration (135, 12, etc): The number of days since the start of the course before unregistration

studentAssessment (173 912 rows):

- id_assessment: internal assessment id
- id_student: internal student id
- date_submitted (18, 22, etc): the number of days after the start of the course
- is_banked (0 or 1): 1 if the student has done the assessment in a previous presentation (course)
- score (0-100): the score of the assessment

studentVle (10 655 280 rows):

- code_module (AAA, BBB, GGG, etc): module code
- code_presentation (2013J, 2014B, etc): presentation code (year and month)
- id_student: internal student id
- id_site: internal site id
- date: number of days since the start of the course
- sum_click: number of clicks on that day

assessments (206 rows):

- code_module (AAA, BBB, GGG, etc): module code
- code_presentation (2013J, 2014B, etc): presentation code (year and month)
- id_assessment: internal assessment id
- assessment_type (TMA, Exam or CMA): The type of the assessment. TMA is tutor marked assessment and CMA is computer marked assessment.
- date: the deadline date, described as the number of days since the start of the course
- Weight (1-100): the weight of the score on the final grade

courses (22 rows):

- code_module (AAA - GGG): module code
- code_presentation (2013J, 2014B, etc): Presentation code. The year and month. J is for October and B is for February.
- module_presentation_length: the length of the course is days

vle (6 264 rows):

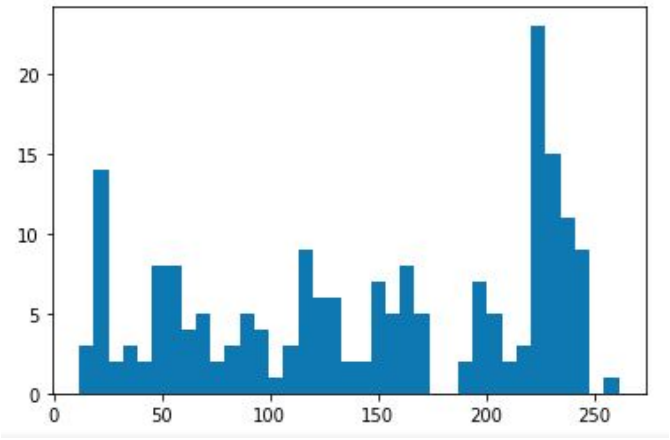
- id_site: internal site id
- code_module: module code

- code_presentation: presentation code (year and month)
- activity_type (resource, oucontent, url, subpage, etc): the type of the visual learning environment site content
- week_from: the week the material was planned to be used from. *Not present for most.*
- week_to: the week the material was planned to be used to. *Not present for most.*

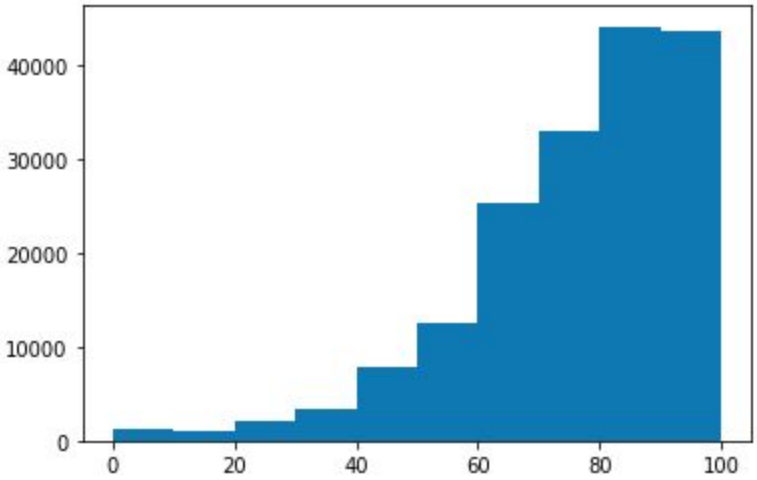
Exploring data

The dataset contains 17875 male students and 14718 female students.

The assignment deadline times seem to follow a weekly/monthly rhythm until the end of the course where all the rest of the deadlines and the exam is.

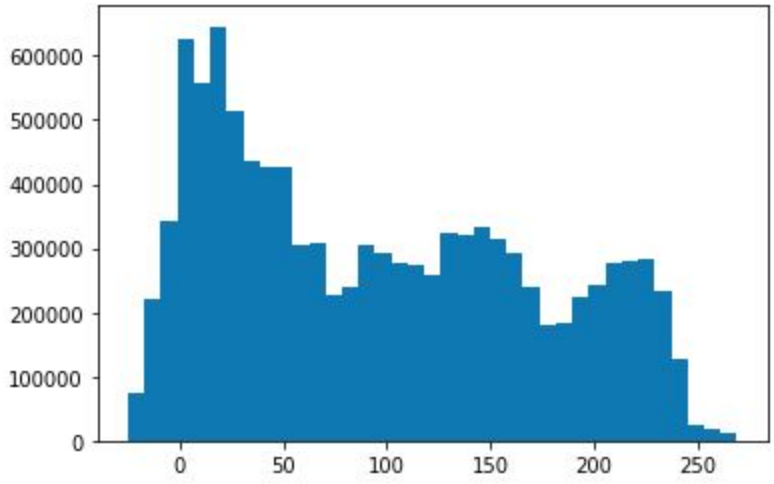


The scores for the assessments seem to be given on the high side.



Most of the VLE resource types are “resources”. There are also a lot of subpages, oucontent, url and forumng rows. There is also a type for quizzes and questionnaires. In total there are 20 different types.

Unlike the deadlines for the assessments the dates the students use the VLE are concentrated at the beginning of the course with a rise for the middle of the semester and the end of it.



Verifying data quality

The dataset was specifically selected for our purpose. Everything we wanted from the data is there and because the data was collected automatically from one source there is nothing absent that we need.

Planning your project

Project plan

Task	**Estimated time per person**
Understanding and visualizing the data	6h
Extracting features out of the timeline data	5h
Using the data to create a predictive model	10h
Collecting data for the project poster	3h
Creating graphs for the project poster	3h
Creating the project poster	5h

Clarification: both people are estimated to spend that much data on the tasks.

Methods and tools

Tools: Python 3, scikit-learn, numpy, matplotlib
Methods: Try different learning models and use the one that works

References

1. Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017).