Lab5 - Simulating students careers

Martina Martini s306163
Politecnico di Torino

Assignment Develop a simulator of the student career in the university, that should be an enhanced version of LABG4.

Q1. Explain the purpose of the simulator, i.e.. the main questions to address
Q2. Explain in details the random elements in the simulated system
Q3. Explain the main assumptions of the adopted simulation model
Q4. Explain all the input parameters
Q5. Describe the (eventual) open data sources adopted in the simulation
Q6. Explain all the output metrics
Q7. Describe the main adopted data structures
Consider a realistic/interesting simulation scenario:
R1. Motivate the choice of all the input parameters
R2. Comment in details the numerical results, focusing on the questions described above.

1 Random elements, inputs and main assumptions

1.1 Introduction

The simulation of the academic careers of a set of students attending the Master's Degree in Data Science and Engineering at the Institute of Politecnico di Torino brings us to pose some questions. Some of them may be the following. How are the graduation grades distributed? How are the number of sessions to graduate distributed? What are the averages? How does the choice of the internship influence the students' careers? Does it influence also the graduation grade? Does the career change if a student is fully dedicated to the writing of the thesis? What are the most difficult courses to pass, according to the number of times taken to pass the exam? And finally, what are the characteristics of the most talented students?

To answer all of these questions, a simulation model is implemented as explained in the following sections.

1.2 Stochastic elements

To simulate such students' academic careers, some stochastic elements must be considered. Firstly, the number of exams to take for each session is given by a uniform distribution between 0 and 5. Consequently, the choice of which exams to take for each session is a random element of the simulation. Secondly, since the choice of taking more than three exams in a session can influence the preparation, a certain percentage uniformly distributed between 0.05 and 0.15 is subtracted from the probability of passing each exam. Concerning this, the probability of passing or failing an exam is constructed to follow the Bernoulli distribution, where each attempt is treated as an independent Bernoulli experiment with the same probability of success p and of failure $p - 1$. Similarly, given iid trials, the probability of accepting or rejecting a grade follows the Bernoulli distribution. The probability of passing the exam on the first attempt or the following ones is given by the open data from the website of the Politecnico di Torino. From the same online page, the grades distribution of all the exams is taken and stored in order to associate a certain grade to an exam according to the number of people that passed such an exam in 2022 with a certain grade.

The syllabus of the MSc course presents also some tables containing optional courses. The choice of which ones to take is given by the number of students that in 2022 chose a certain course over the ones in a specific table. The same methodology is used for the choices of the free ECTS. Instead, if the student decides to attend an internship or a challenge, the number of sessions in which the student is blocked in this is treated as a uniform variable: in case of the internship the r.v. is $\sim U(1, 2)$, in case of challenge the r.v. is $\sim U(2, 5)$. Similarly, the number of sessions that the student takes for writing the thesis is treated as a r.v. $\sim U(2, 5)$. Finally, some bonus points are added to the final grade based on the quality of the thesis (from 0 to 4), the goodness of the presentation (from 0 to 2) and other metrics like the speed of graduation or the number of lauds (from 0 to 2): all of these three variables are uniformly distributed. To conclude, a random seed is introduced for reproducibility purposes.

1.3 Input parameters

For what concerns the input parameters, I assume to set the following fixed variables: the total number of considered students is 100, the total number of courses for the MSc is equal to the number of courses in the syllabus (15, considering the thesis and the *table A*), the number of sessions per year is 4 and the minimum number of credits achieved to insert the thesis or the internship in the syllabus is equal to 48. Moreover, the grade probability distribution of each exam, the list of all the courses in the MSc, including the ones to choose for each table and their ECTS are given as inputs. Also, the probability of passing an exam and the probability of being chosen among the ones in the same table (interpreted as the number of enrolled students in an exam divided by the sum of all the enrolled students in each exam in the table) is used as reported in the Polito website. Finally, notice that the maximum grade to get is 30 and the graduation grade is computed as the average grade divided by 30 and multiplied by 110, plus the bonus points. If the graduation grade is greater than 112.5, the student graduates with "110 cum laude". To conclude, a confidence level is fixed at 0.90 and the accuracy acceptance at 0.97, while the number of batches used for the confidence intervals' computations is set at 4.

Notice that the given parameters are found and taken from the official syllabus of the Polito website and from the AlmaLaurea one.

1.4 Main assumptions

Regarding the assumption I made, some hypotheses are considered in the simulation. First of all, the time is expressed in terms of years: each year is interpreted as a complete cycle of 4 sessions. Then, some data were not encountered on the official website, so they were assumed: firstly, since the percentage of students that choose to do an internship is equal to 44.4% - according to the Polito website - I assumed that 10% of people chose to attend the challenge and the remaining 25.6% the free ECTS. Speaking of, since these free ECTS are courses taken in common with the students of MSc in Computer Science, their aggregated statistics are not usable, so their probability of being chosen is assumed. The same discussion is done for the course Object Oriented Programming. Instead, if the student does not achieve at least 48 credits, the probability of choosing the challenge, the applied data science project or two free credits is also assumed. Moreover, since there are no statistics about the dropping students in MSc in Data Science at Polito, I assumed that every student concludes their career.

Finally, I assumed that if the number of remaining exams to take is one, then the probability of accepting the grade is 100%, while if it is less than 4, the higher the grade, the higher the probability. Also, if the student passed at least half of the total exams and the mean grade is greater than 27, then the probability of rejecting low grades for the next exams is higher.

2 Output metrics, data structures and further analyses

The output metrics used to evaluate the results of the simulation consist of what shown in Figure 1. It is also displayed how many students found a certain course the most difficult counting how many of them re-took the exam more than one time. The main data structures are very naive and include several lists, used for the final analyses, a few dictionaries and the class Student, whose attributes are the identification number, the number of remaining exams (they are initially set to the total number of courses), the grades he gets, the number of sessions and years he takes to graduate, the final grade and the list of the exams he does not pass each session.

Further analyses are added at the end of the simulation, such as some statistics about the graduation grades, the average grades, the average times to retake an exam and the time taken to pass the last exam. As you can notice, different simulations running on batches fo 100 students can bring pretty similar results: the time to graduate is a little bit higher than expected, but the mean graduation grade is correctly around 100. Also, I tried to filter the data-frame overcited to evaluate the most talented students, who got the higher grades by re-taking on average the exams at most 1.5 times, trying at most 2 times the last exams and graduating in just the minimum number of sessions. Moreover, it is displayed how the internship and the full dedication to the thesis influence the student's career: in the first case, (as expected) the graduation grades of the people taking the internship are quite often slightly lower than the mean graduation grade, while in the second case, it is higher (since I assumed that if the last exam to take is the thesis, the chance to get higher bonus points is higher). For what concerns the most difficult exams, mathematics in machine learning and data ethics and data protection seem to be the most selected ones.

In conclusion, another interesting measure is given by the correlation between the average graduation grade and the number of years taken to graduate: this mean remains stable when considering one or two years to graduate, but it increases when the years become more than two. All of these considerations are taken after achieving very high accuracies (0.997 for the grades and 0.976 for the sessions).

For clearer evaluations, please refer to the code provided.

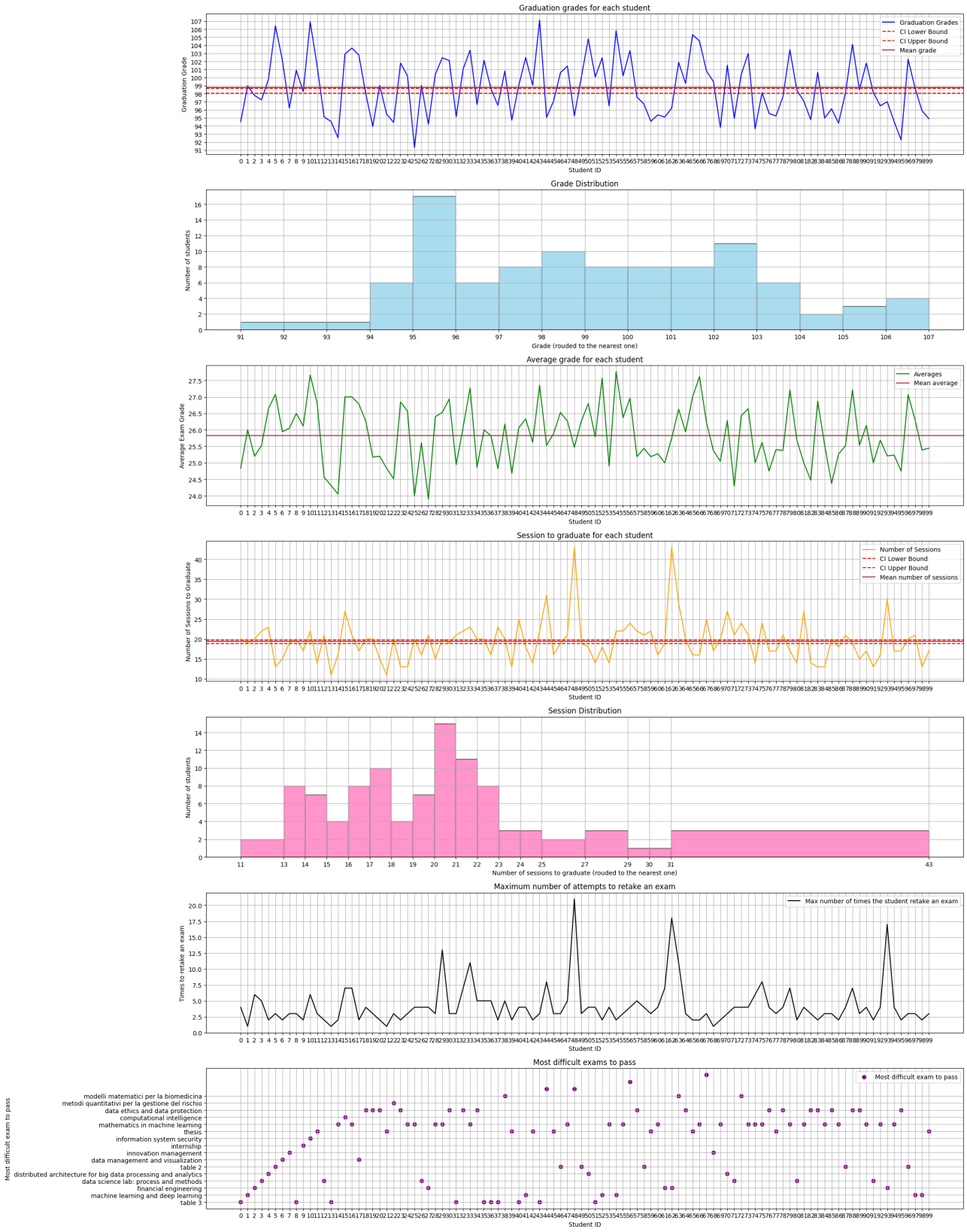

Figure 1: Main output metrics