Abstract

Daniel Simon Siemmeister

Graz, March 26, 2022

The aim of this work is to estimate based on data of many students, how many of this students will be studying actively in 3 years from now. The goal is only to predict the total number of actively studying students, in contrast to predicting if a student is actively studying individually.

To achieve this goal, a few approaches are tested, which try to solve the problem in different ways. Generally the problem is divided into two smaller subproblems. The first of the two deals with the prediction of the number of actively studying students who are already enrolled. The second subproblem is concerned with predicting the future of students who will enroll in the following two years. In most of the approaches, machine learning models of different architectures are applied and compared in terms of their predictive ability.

The key insights of the work are:

- True understanding of the problem. That means, only the number of acitvely studying students is required, and one does not need to classify each person exactly.
- Machine learning predictors for classification make a classification choice based on an estimated probability. The final choice is made based upon a predifined threshold value. However, one can use these probabilities without making an actual classification.
- Different machine learning algorithms produce similar results. The complexity of the algorithms is not decisive.
- Using the estimated probabilities, one can reasonably predict the expected number of active students.

Although the predictions give good results, it is important to note that more data is needed to adequately test all approaches. In addition, there is a risk that there will be fundamental changes in student behaviour over the 3 years and that the predictions will not come true.