Rahul Jain

Professor Molontay

Data Science

3 April 2023

Data Science HW3: Exercise 2

Using predictive analytics methods, we would like to identify first-year college students at risk of dropping out and academic failure. Write a one-page long essay covering the following points! (50%) If you would prefer, you can write an essay about a data science research question of your choice.

1. Design the analysis, think through the stages of the analysis following CRISP-DM methodology! How can you state the problem as a data science problem? What machine learning approaches could be used here? (See Lecture 02!)
   1. Business understanding – What does the business need?
   2. Data understanding – What data do we have / need? Is it clean?
   3. Data preparation – How do we organize the data for modeling?
   4. Modeling – What modeling techniques should we apply?
   5. Evaluation – Which model best meets the business objectives?
   6. Deployment – How do stakeholders access the results?

b) Do you think that the requirements of a successful data science projects are met? Go through the 7 requirements that we have covered in class! (See Lecture 02!)

c) What ethical questions are raised in this project? Mention utilizations that you think are useful and ethical and give examples of bad applications as well!

1. This could be seen as a binary predictive modelling problem with a question as such: what factors dictate a first-year college student’s decision to drop out of college?

There are many factors at play when a student, whether a first-year or not, makes the decision to drop out of college. According to Stanford Professor of Education and the longest serving President of California's State Board of Education, Michael W. Kirst, some reasons include a lack of staff or financial support, a poor fit with the college, homesickness, and uncertainty about their course or career (Klochkova). Each of these feelings depend on extremely different factors, from an individual’s engagement in class to distance from home, and income to career choice. To be able to effectively identify at-risk college students, we would need a lot of information on each student, of which each element is likely not clean. We would need to perform quite a bit of data transformation such as log transformation for income and discretization of distance from home.

We can organize the data such that we have a row for each student in our dataset and columns for each attribute. The final column is a binary value for whether or not the student dropped out – 1 for yes and 0 for no. To ensure we aren’t training on the target variable we would need to remove this final column and store it in a separate array. I believe the best model for this dataset is a K-Nearest-Neighbors classifier. A decision tree classifies precise cutoffs on the axes, but I’m not convinced that the decision to drop out is based on a few precise factors but rather a combination of them.

The stakeholders can access the results by predicting the likelihood of each enrolled student dropping out by calculating the number of nearest neighbors that did drop out of college and dividing by the K-value selected weekly. Students that have high likelihoods can be further examined as to what attributes placed them in the high-risk group and can be provided administrative support.

1. I believe this project meets the 7 requirements for a successful data science project. By doing research and conducting interviews with college administrators we can ensure the project is headed in the right direction from the very beginning.

We also have access to lots of data since there are many students enrolled in college and a sizable proportion of them drop out – over the years, this lends itself to many observations available for us to see.

We have observed that there are many reasons for a student to drop out of college and each reason has a number of different attributes that may contribute to the decision. We must collect information on dozens of attributes for each student to predict which of these reasons is most likely to cause a student to drop out.

Hopefully the data we collect is clean – I’m not sure exactly where or how this kind of data is stored but if it’s clean then the project meets all 7 requirements for a successful data science project.

The collected data can be unbiased if we avoid self-reporting or hold students liable for any misinformation if they do self-report. This data could be pulled from either government databases or college applications to ensure reliability of the data.

If we diagnose a student as at risk for dropping out early on, we can remedy the underlying issues to ensure they continue their education. This means that there is a high capacity to act, since we are working to make changes at the individual level as opposed to the systemic level.

Finally, we are able to measure the ROI of a project by comparing dropout rates before and after implementing this predictive modelling system.

1. One of the biggest ethical questions that arise from this is the use of personal data. Some students may find it to be an infringement of their privacy if universities a student’s personal information to a) find similar cases where a student may be struggling or b) predict if a student is struggling at all.

# References

Klochkova, Victoria. *Why College Students Drop Out and How to Prevent It*. 19 August 2016. Web. 3 April 2023. <https://collegepuzzle.stanford.edu/why-college-students-drop-out-and-how-to-prevent-it/>.