

As a graduating senior, I have just completed my application season for graduate school. I received a few acceptance letters, but more rejections. Most of the results seemed reasonable to me, but that didn't preclude a few of the safety schools I applied to from rejecting me. In my opinion, graduate school admissions is a very metaphysical thing, so it is my keen interest to explore the logic behind the admissions office's decisions.

I will be using a dataset called "Graduate Admission 2" that I found on Kaggle. According to the source, this dataset is inspired by the UCLA Graduate Data Set. <https://www.kaggle.com/datasets/mohansacharya/graduate-admissions>. The citation would be "Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019". It includes 400 observations, 7 predictors and 1 response variable. Among those variables, 1 out of 7 is a categorical variable and the rest of them are quantitative variables. This dataset is well-preserved and well-managed, so I didn't find any missing data in it. If it happened to be some missing data in the future, I would reckon that I will learn and initiate the data cleaning process like the one did in the sample final project.

The variable I am interested in predicting is the "Chance of Admit". In other words, the objective of this project is to explore the most important factors for a student to get into graduate school and to select the most accurate model to predict a student's chances of gaining admission into Graduate School (this is my response variable) based on the several predictors such as "GRE Score", "TOEFL Score", "Target University's Rating", "Statement of Purpose Strength", "Letter of recommendation Strength" "Cumulative GPA", "Whether or not has research experience before", respectively. Since my outcome, aka response variable, is continuous, I believe my question would be best answered in regression approach. Based on my intuition and my application experience that three predictors, GPA, SOP and Research experience will be especially useful. The goal of my model is inferential because I aim to find the possible correlation and causality, and state relationship between outcome and predictor.

Since we have covered most of content and knowledge about the data loading and EDA, I plan to have my data set loaded before next week, and do the EDA before week 5. However, since there is a lot of content we haven't learned and covered yet, I will try to synchronize my projects according to the progress of our studies later.

I think there will be a lot of running errors in the data modeling section, because it involves a lot of interleaved dataset. The question I have for the dear prof. Coburn and instructional team is that "Will the curriculum cover more than just theoretical content to ensure we have enough tech and practical knowledge to complete our final project? Because now I feel like we are exploring a uncharted field, but we can't move forward because of the insufficient knowledge reserves.