**Classification Project:**

The dataset needs to be different than the ones that what was used in class and ideally will be 1,000 records or more. If you wish to use a smaller dataset, please get permission to do so.

The following section outlines the report expectations. Note that application of results may be hypothetical. The point of these components is to have you think through the issues associated with deploying a model in a real environment.

Report Requirements:

1. *Introduction*
   1. *Problem statement*

* The goal of this project is to create a model that classifies Majority/Minority Pell institutions.
  + - Does this school have a Majority population receiving a Pell grant?
    1. *Background*
* I have been researching into the CollegeScorecard dataset for the past year and have been using an indicator variable I created by categorizing institutions with either a majority or minority proportion of their students receiving a Pell grant. The original motivation for the indicator variable I created was to categorize institutions where lower-income students were attending. The proportion of Pell students was chosen to create an indicator variable because the original motivation of the Pell grant by U.S DoE was to aid low-income students. Today, a student may be eligible for a portion of the Pell grant if they or their parents (if the student is a dependent) household-income is less than $50,000. The portion the student receives increases as the amount of income decreases and the maximum amount of $6,495 is granted to the students whose household-income is less than or equal to $20,000. For these reasons, receiving a Pell grant is associated with low-income students.
  + 1. *What may be done if the classification problem is solved*
* I want to answer the question “Does this school have a majority student population receiving a Pell grant?”. By classifying these institutions, we can understand the features that are important in predicting whether a majority or minority of an institution’s student population is receiving a Pell grant.
  + 1. *Explicit Analytic goal*
* To predict whether an institution with certain features is likely to be either Majority Pell institution or Minority Pell institution.
  + 1. *Proposed application of the model results*
* I will use the information gained in this project going forward in my research of the CollegeScorecard dataset.
  1. *Data source*
* The CollegeScorecard dataset is released by the U.S. Department of Education through IPEDS (Integrated Post-secondary Education Data System). IPEDS surveys post-secondary institutions, the IRS, and FSA annually and collects the responses to 13 different surveys about the institutions. The data in the CollegeScorecard dataset comes from these 13 surveys and is released by IPEDS through the U.S. DoE. The unit of analysis for the CollegeScorecard dataset is post-secondary institutions and includes data on the institutions, the student demographics at the institutions, the admission rates, SAT/ACT scores, costs related to attending, academic outcomes, and loan outcomes.
  + 1. *Key “response” field/variable (what are you trying to predict)*
* **PELLCAT:** indicates whether an observation is a Majority Pell institution (0) or a Minority Pell institution (1). This

variable was created from the variable PCTPELL by categorizing institutions with greater than 50% of their student population receiving a Pell grant as Majority Pell institution and institutions with 50% or less of their student population receiving a Pell grant as Minority Pell institution.

* **PCTPELL:** the proportion of the institution’s student population receiving a Pell grant.
  + 1. *Data*
       1. *Key metrics (number of records, counts of response variable, number of features)*
* Originally, there were 2,384 variables and 6,806 observations and there were 3,439 Majority Pell institutions, 2,575 Minority Pell institutions and 792 institutions that were missing a value for the variable PCTPELL.
  + - 1. *Data concerns (e.g., missing values in key features)*
* Some institutions were missing values for the variable PCTPELL, and some institutions were in either a U.S state or Washington D.C.
* There were 16 variables that were related to either the institutions URL, ALIAS, FSA ID, or program description and 155 variables that were pooled and/or suppressed versions of other variables in the dataset.
* There were 1,419 columns that is missing a value for every observation.
* Of the remaining 783 variables, there was 523 variables that contained missing values. The method used to handle these missing values will be discussed in section 2A.
* Multicollinearity was found between some of the remaining variables

1. *Data Preparation*
   1. *Plans to address data concerns*

* I dropped the rows that were either missing PCTPELL or were not in a U.S. state or Washington D.C. This left 5,879 observations of which 3,435 institutions were labelled Majority Pell and 2,444 institutions were labelled Minority Pell.
* I dropped all columns that contained only missing values. In addition, I dropped columns that I think will not be useful in predicting such as the institutions URL, variables that are pooled or suppressed over time, and program descriptions. After dropping these columns, the dataset contained 783 variables.
* From the remaining 783 variables, there was 523 variables that contained missing values. I decided that a missing value tolerance of 42.8% would be the cutoff point for imputation and dropped the 337 variables that exceeded the cutoff point. The remaining 187 variables will have their missing values imputed with the median. The dataset now contains a total of 446 variables of which 3 are institutional IDs, 225 are categorical, and 221 are numeric.
* \*TALK ABOUT VIF AND CORRELATION\*
  1. Feature engineering
     1. Motivation for engineered feature
* There were 128 variables that correspond to program offerings at an institution. Of which 190 variables was a categorical variable that is either a 0 (if it does not offer the program), 1 (if the program is offered both in person and online), or a 2 (if the program is only offered online). The additional 38 variables corresponded to the proportion of degrees/certificates an institution awarded for a particular program they offer. Since there are 38 programs and each program relates to 6 variables, I decided to use Principal Component Analysis on these variables to reduce the number of columns.
  + 1. Engineering methodology
* ***Variable Creation:*** The CollegeScorecard dataset included the variable T4APPROVALDATE which is an institutions entry date into the Title\_4 Program. This variable was initially in one column in the format month/day/year. I split this column into three columns representing month, day, and year and then created a variable for the Season an institution entered the Title\_4 Program.
* ***One-Hot Encoding:*** I used One-Hot Encoding to create dummy variables for the variables ST\_FIPS, CURROPER, CONTROL, MAIN, HCM2, month, Season, OPEFLAG, PREDDEG, HIGHDEG, ICLEVEL, OPENADMP, ACCREDCODE, HSI, HBCU, ANNHI, PBI, TRIBAL, NANTI, and AANAPII since these variables are ordinal data.
* ***Target Encoding:*** My dataset included variables for the city the institution was in, the name of the institution, the ZIP code, and the name of the agency that accredits the institutions. I chose to target encode these variables, as well as the variable I created, season. In addition, I created variations of the ZIP code by taking either the first two, three, or four digits and creating a new column labelled ZIP2, ZIP3, and ZIP4, respectively.

1. Model Development
   1. Methods use (including motivation for using the methodology)
   2. Model assessment
   3. Results
   4. Winning model
2. Conclusions
   1. Model results implications
   2. Potential application of classification model
   3. Expected benefit
3. For this project select a dataset that includes a *binary*, *nominal class*, or *ordinal class* value that you wish to predict

(or a continuous value data element that could be used to create a label to predict)

1. Create a set of “classifiers” (prediction models) that predict the categorical output.
2. You must “engineer” (create) 5 new data columns/fields that are used in your models. Provide rationale as to the reason that you chose the creation of these columns.
   1. *Bonus (up to 5 points):* 1 point for each additional engineered column/field. Using PCA or Autoencoder to reduce

the data and create a classifier using it, will also net these 5 bonus points.

1. Determine which methodology/model is the “winner” and provide your reasoning (“I like this method because it’s really cool” is a valid answer, but for full credit I’m looking for something that’s more substantive).

* ***WHO MAY BE ELIGIBLE FOR THE PELL GRANT?***

The Federal Pell Grant is usually awarded to undergraduates who have a high degree of unmet financial need. Students whose families have a total income of up to $50,000 may be eligible for the need-based funding, though most Pell grant money goes to students with a total family income below $20,000. Your eligibility will be determined by the FAFSA. The U.S. Department of Education uses a standard formula, established by Congress, to evaluate the information you report when you apply for the FAFSA. After you’ve completed the financial aid application, you’ll receive a confirmation e-mail letting you know your application has been processed and providing you a link to your Student Aid Report, or SAR. You’ll also be able to view your Expected Family Contribution (EFC), the number that determines your eligibility for federal grants, and how much you’re eligible for. From there, your information is also sent to the school or schools you indicated on your FAFSA.

<https://www.scholarships.com/financial-aid/federal-aid/federal-pell-grants/>