Title: Undergraduate Students Enrollment Quantity Prediction
Name: Bo Li

Over time, we can perceive the explosive growth of students in higher education. This trend is most important for us as undergraduate students, which can help us understand our competitiveness in society. We can make decisions on pursuing future education and planning careers.

The data used in this project came from two main sources with 5 features: (1) total undergraduate enrollment, (2) percentage for school age student with the total student population, (3) nominal Gross Domestic Product (GDP), and (4) total population for the United States.

In National Center for Education Statistics, I selected the tables for 'Total undergraduate fall enrollment 1970 through 2018' and ' Estimates of the resident population, by age group: 1970 through 2019' from https://nces.ed.gov/programs/digest/current_tables.asp. Moreover, the Federal Reserve Economic Data (FRED), https://fred.stlouisfed.org/, publish the data for GDP starting from 1947 and population start form 1959.

The initial format is poorly labeled and organized after loading, so I slicing the data from original tables and relabel them. In these datasets, I also delete some unnecessary information dividend by attendance status, sex of student, and control and level of the institution. And I add a new 'school age/resident population %' column derives from the population from age 18-19 divide by the total resident population. After reorganizing data and concatenating tables together, the resulting table has four columns with yearly data from 1970 to 2018.

After cleaning the data, I complete a plot line graph that shows yearly enrollment data in **Figure 1** directly. We can see a recession period during the 1990s and a peak in 2010 followed by a drop in the number of undergraduate enrollment. From the graph, undergrad enrollment is not continuously increasing, but with fluctuation from 7 million in 1970 to 19 million in 2018. Moreover, we can see the number of enrollment for female students exceeds male students after the year 1978.

Next, I perform a principal component analysis over my three feature columns. **Figure 2** shows the results: the first column capture almost 99.9% accuracy in the blue line without scaling. However, this is large because of unit variation among features. For example, the value represented in GDP is in billions and the population is in thousands which are hard to compare. After scaling (orange line), the variance ratio of the first component now becomes 0.86 and the second component with 0.99. Although through PCA analysis, the features can be reduced without losing information, it is critical for linear regression prediction. More features can produce higher accuracy indicate all components are necessary for regression.

For my last analysis, I create a sklearn pipeline consisting of a StandardScaler transformer and a LinearRegression model. The dataset is also split by default method of train_test_split and the model achieves a score of 0.884 for linear regression. Moreover, **Figure 3** shows the coefficient weights for each of the features used by the model. The most important factor for predicting future undergraduate enrollment is the num of population and increasing GDP has a counter effect on the growth of undergraduate students.

To conclude, my analysis shows that undergraduate enrollment has a positive relation with population size. Also, more people tend to pursue bachelor degrees during the economic depression for improving their skills and com competitiveness strength.

Figure 1: Undergraduate enrollment, by sex

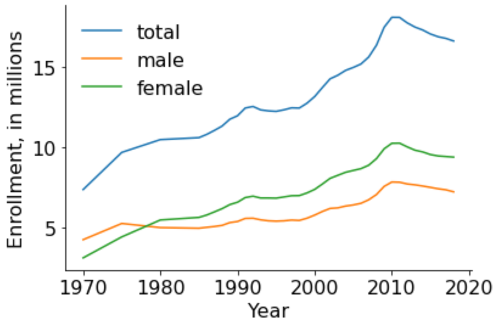


Figure 2: Principal Components Analysis

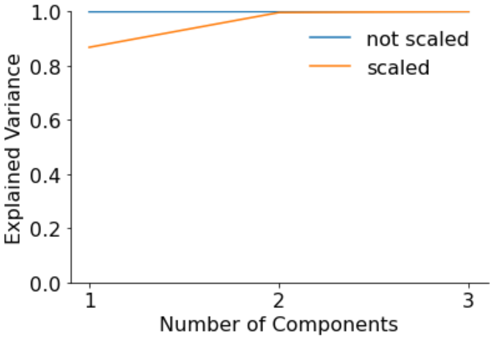
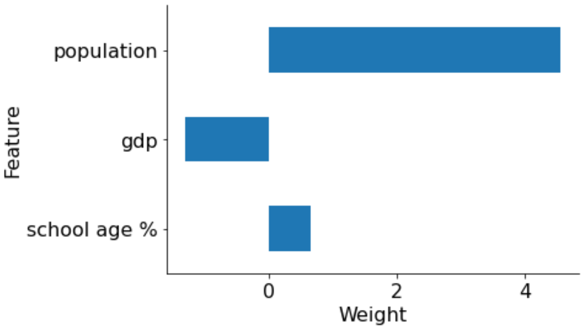


Figure 3: Linear Regression Coefficients


Feature
population
gdp
school age %
Weight