US College Tuition Prediction

By: Miho Hunter

ABSTRACT: The goal of this project is to use regression models to predict the US college tuition.

DESIGN: The average US college tuition has risen significantly over the last two decades. I am interested in finding out which factors contribute to the high cost of college education.

DATA: The data used for this model are Forbes' America's Top College List, Cost of College List from the Stephen Phillips Memorial Scholarship Fund, College Data from kaggle, and Global 2000 List by the Center for World University Rankings (2021-2022 edition) from Center for World University Rankings.

ALGORITHMS: The first step was to perform EDA and find relationships between variables. The baseline Linear Regression model was built with all features. In order to improve the model, I dropped high p value features and also scaled the features to standard normal distribution. I used an 80/20 train/test split and also K-Fold cross validation to validate models. I inspected the model parameters and confirmed that the assumptions hold good.

TOOLS: Selenium was used for scraping data from Forbes and the Stephen Phillips Memorial Scholarship Fund site. Beautiful Soup was used for scraping data from the Center for World University Rankings. Pandas was used for exploratory data analysis, and matplotlib and plotly express were used for visualization. Linear models were built using sklearn and statsmodels. Fuzzywuzzy was useful for string matching when merging the different data frames.

COMMUNICATION: All slides, visuals, codes, and data are available on my github site.

CONCLUSIONS: The baseline model with all features had R^2 = 0.727. R^2 was improved to 0.754 for the final model with selected and scaled features.

The Estimated Regression Equation:
Tuition = -0.0062 + 0.3305*private - 0.3262* Apps + 0.6109*Accept -0.3093*Enroll + 0.1086* Top10perc +0.1569 *Terminal +0.1557*Perc_alum_donate +0.3454 * Expend + 0.1338 *Grad_Rate