Abstract

## Team Members

Anthony Harris, Emanuel Timbo, Samuel Williams

## Research Question

Can we predict financial success (income > 50k/yr) in the US using intrinsic and extrinsic attributes?

## Research Interest

As men of color in higher education in the United States, we are particularly aware of and determined to overcome societal boundaries to personal success. One of the main barometers for that success is financial health. Some of the boundaries that can impede that health are implicit bias, institutional discrimination, and overt inequity. These boundaries are determinant on a person’s intrinsic characteristics – race, gender, nationality, etc. We look to measure these immutable attributes against acquired attributes, such as education, to determine their significance in predicting financial success.

## Data Analysis

Using a dataset extracted from the US Census database, a model will be constructed that predicts whether a person’s yearly income is greater than $50,000 per year based on different attributes. Before this model can be constructed, the data will need to be prepared.

This preparation will be to ensure the most relevant outcome. Extraneous columns will be removed from consideration. Missing values will be handled through imputation. Remaining columns will be analyzed (and if necessary, adjusted) for outliers. Columns representing categorical data will have dummy columns created for algorithm input. Using discretization, categorical predictor alternatives will be reduced to a limited number of relevant options. Lastly, the predictors will be tested for collinearity to determine if further reduction is possible.

Following preparation, a series of analyses will be conducted to determine individual correlation between predictor variables and the outcome variable. After analyzing potential trends, different combinations of attributes will be tested and selected. The best attributes will be used for the algorithm. Because the output is binary and disqualifies other common models, logistic regression will be implemented for this dataset.