Group proposal

David Amankwah
Xinchun Chen
Abhishek Nimmakayala

Poverty is one of the topics that has been researched by a lot of economists and data scientist. It is one of the economic problems most countries want to alleviate. As one of the countries that have the strongest economy, U.S. also faces the challenge in domestic poverty. This project will focus on the poverty rate in New York, since it is one of the most representative cities in the United State, also with the highest levels of income inequality in the country (Abadi,2018).

This project will use the decision tree model to predict the poverty status in New York; finding the indicators that can make the best prediction. The data is from the NYCgov poverty measure data, which is generated annually by the poverty research unit of the Mayor's Office of Economic Opportunity (NYC Opportunity). The number of observations for this dataset is 68,644 with 79 unique variables. The majority parts of the data cleaning work will be reducing the size of the dataset, which includes reducing the features and also the observations. As for the final outcome of our model, we will use confusion matrix to measure the accuracy of our model and also compared our output with and without preprocessing.

Rough Schedule:

Mar 30 — April 5: Finding the dataset, specify the problems, preprocessing the data

April 6 — April 13: Data Mining

April 14 — April 22: Evaluation and Result Exploitation

References:

Abadi Mark. (2018). "Income inequality is growing across the US — here's how bad it is in every state" Busniess Insider. https://www.businessinsider.com/income-inequality-in-us-states-ranked-2018-3