David Amankwah
Individual Report
Data Mining Project

I. Introduction

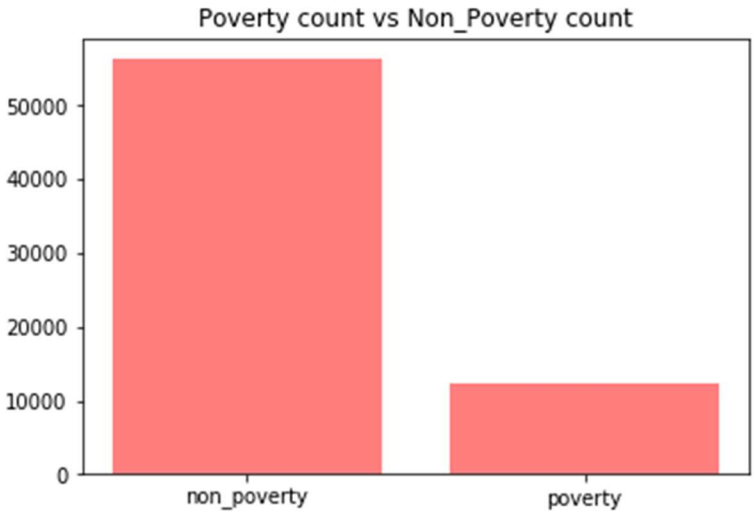
Poverty is one of the topics that has been researched by a lot of economists and data scientist. It is one of the economic problems most countries want to alleviate. As one of the countries that have the strongest economy, the U.S. also faces the challenge in domestic poverty. This project will focus on the poverty rate in New York City, 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 analysis to predict the poverty status in New York; finding the indicators that can make the best prediction. The rest of this report is organized as follows. Section II, the description of the data set provides the background and source of the data, alongside with some exploratory data analysis. Section III explains the decision tree algorithm used in this predictive analysis project while providing some background information on the development of the algorithm. Also, it describes the procedures used in our preprocessing stage of the data analysis and the implementation of the learning technique used in this project. Section IV provides the results from our experiments and finally, the conclusion reiterates our results and discusses has been learned and suggests improvements for future analysis.

II. Background of the dataset

The data used in this project is retrieved 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.

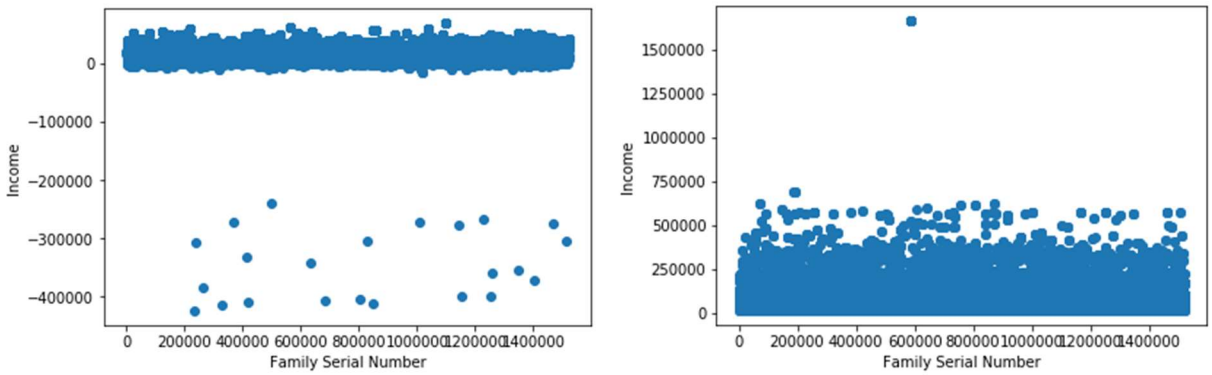
The dataset provides some features of families living in New York City (with unique identifiers) such as the educational attainment, employment status, annual income, etc. Based on these features, a decision is made to classify a given family to be in poverty or not. Running a summary statistic and some exploratory analysis on the data set, a significant observation noted is that the number of families in poverty is less than the number not in poverty. Below is a bar graph bar to illustrate this imbalance:



(Figure 1 shows the poverty versus non-poverty count)

Below are some addition graphs to illustrate the results from the exploratory data analysis.

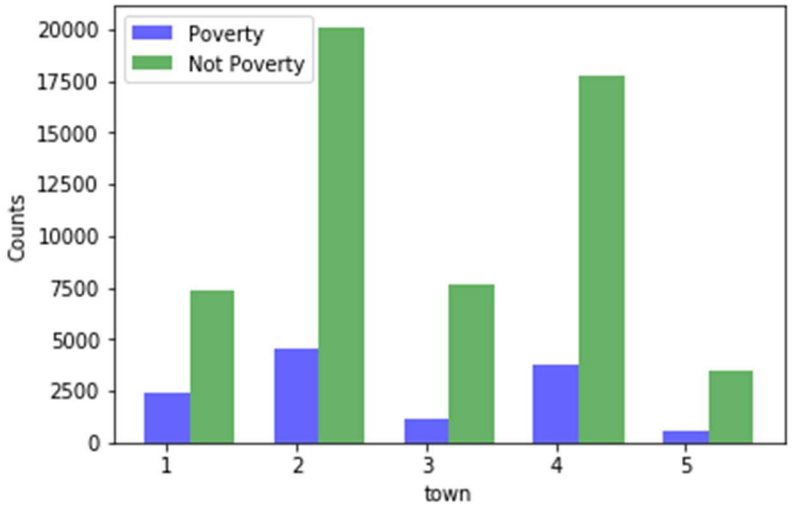
Graph below show the yearly Income in Poverty vs Non-Poverty Families;



(Figure 2 shows the income of poverty versus non-poverty families)

Graph presents the poverty and non-poverty count in different towns of New York City;

1: Bronx, 2: Brooklyn, 3: Manhattan, 4: Queens, 5: Staten Island



(Figure 3 shows the poverty versus non-poverty count across towns)