Project Report

Poverty Prediction in New York

Group 8

Xinchun Chen, Abhishek Nimmakayala, and David Amankwah

I. Introduction

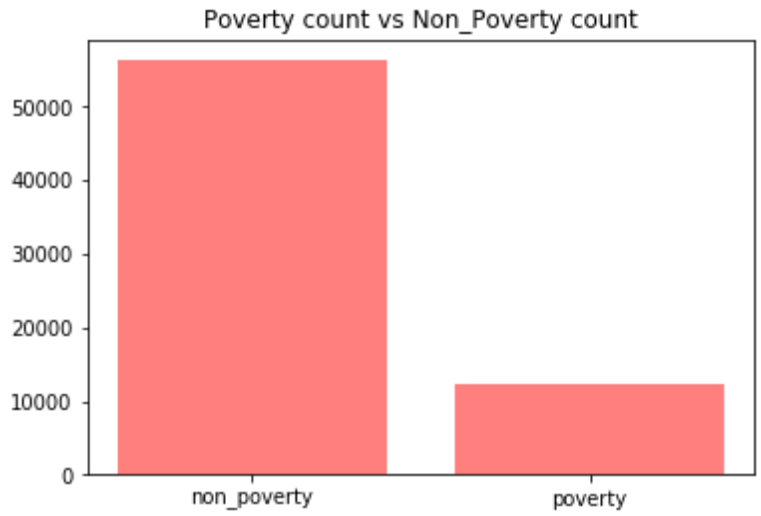
Poverty is one of the topics that has been researched by a lot of economists and data scientist. It is one of the economic problems most countries want to alleviate. As one of the countries that have the strongest economy, the U.S. also faces the challenge in domestic poverty. This project will focus on the poverty rate in New York, 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 provide 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 (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 statistics 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 nonpoverty count)

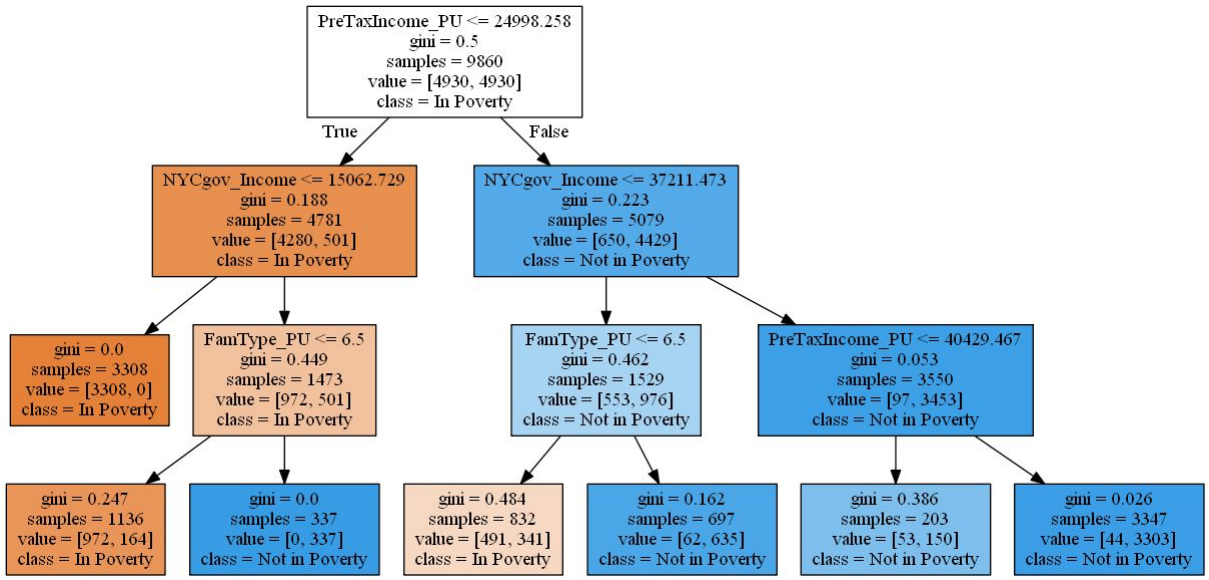
III. Algorithm and Preprocessing Methods

A Decision tree is a non-parametric and supervised algorism that go through all the available attributes to make the best prediction base on information gain. The information gain can tell us the importance of a certain attribute by calculating the impurity of the parent and child nodes. For our study, we use the Gini as our cost function because it would calculate the result faster, as we have a relatively big dataset 68644*79. The function of the Gini Index is $Gini = 1 - \sum(pi)$.

***Preprocessing*:**

Before we can produce our decision tree, we need to make sure our dataset has balanced target classes, right dimensionality, and most representative observations. The preprocess techniques we implemented in our study are feature selection and instance selection. For the feature selection, The goal is to drop irrelevant or redundant features to reduce the dimensionality of the dataset. Some of the variables are *Household Unit ID*, *Age Category, Poverty Gap, Tax Unit, etc.* We also decided to drop the variables that have a high correlation with the total income, since income and

poverty status shared the same characteristic (low income directly decides whether the person is in poverty or not). If we include total income, the decision tree would pick income as criteria in every node, which will underestimate the impact of other variables (as figure 2 shows below).



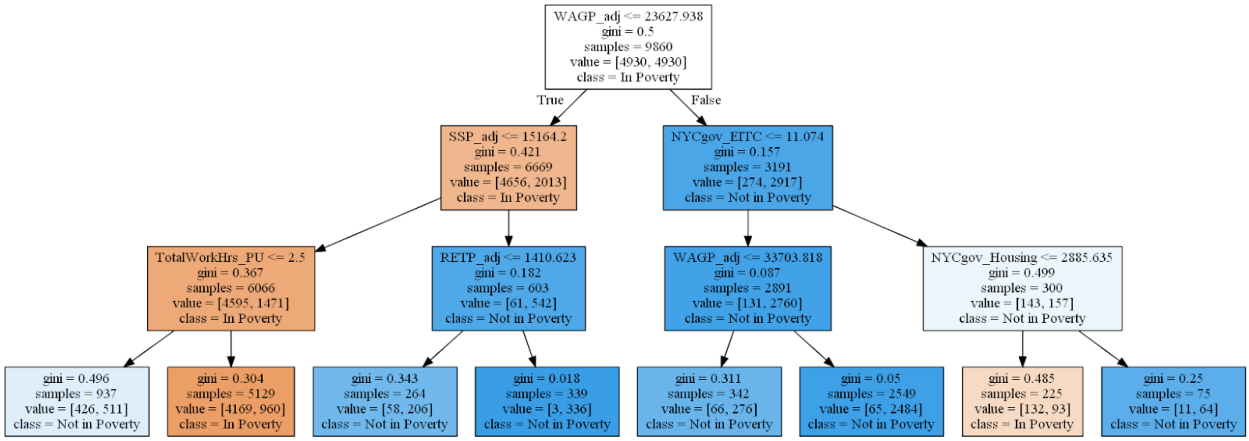
(Figure 2 shows the model will pick Income 4 out of 6 times as the node)

The instance selection will select the observations that are most representative for the study. Our dataset records 70000 individuals and is grouped by the household unit. The table shows the first 7 observations of the dataset, the observations that have the same SERIALNO indicates the individuals belong in the same household. The Poverty Status of each individual is decided by the head of the family (*Povunit_Rel = 1*), if the head of the family is in poverty then rest of the members are in poverty too regardless of other features. In the case, we only kept the head of the family as our unit of measurement.

SERIALNO	AGEP	Povunit_Rel	NYCgov_Pov_Stat
39	51	1	2
55	60	1	2
55	52	2	2
55	26	4	2
55	20	4	2
55	20	4	2
69	39	1	2

The Decision Tree:

After the preprocessing, we using Sklearn for the model building, I set the max_depth to 3 in order to avoid overfitting. Figure 3 below is our first decision tree. The accuracy of the training set is 83% and 79 % for the testing set.

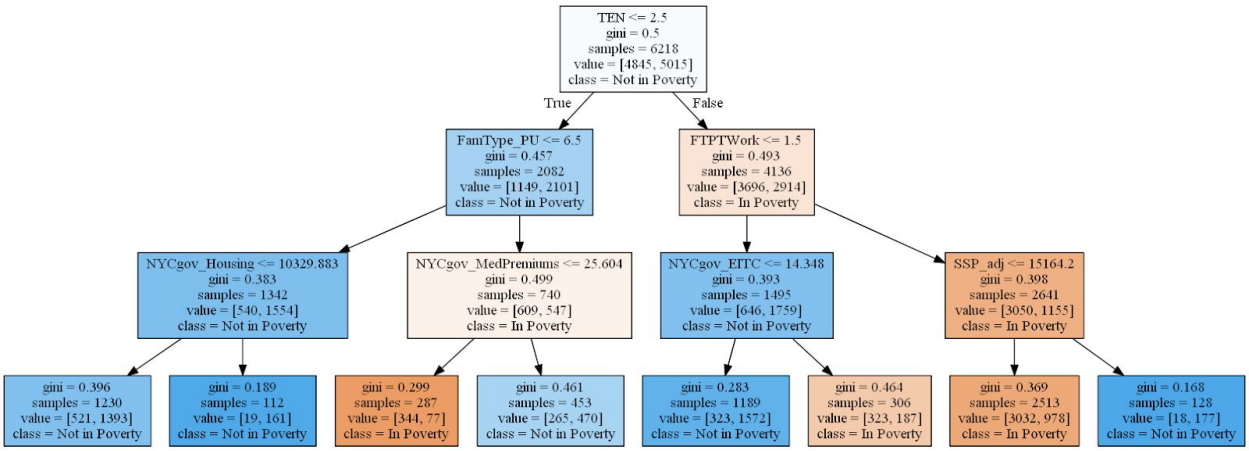


(Figure 3)

Even though we preprocess the dataset before we implement the decision tree, it can still provide an unstable result because of the variation of the dataset. Only one tree is not representative enough for our study.

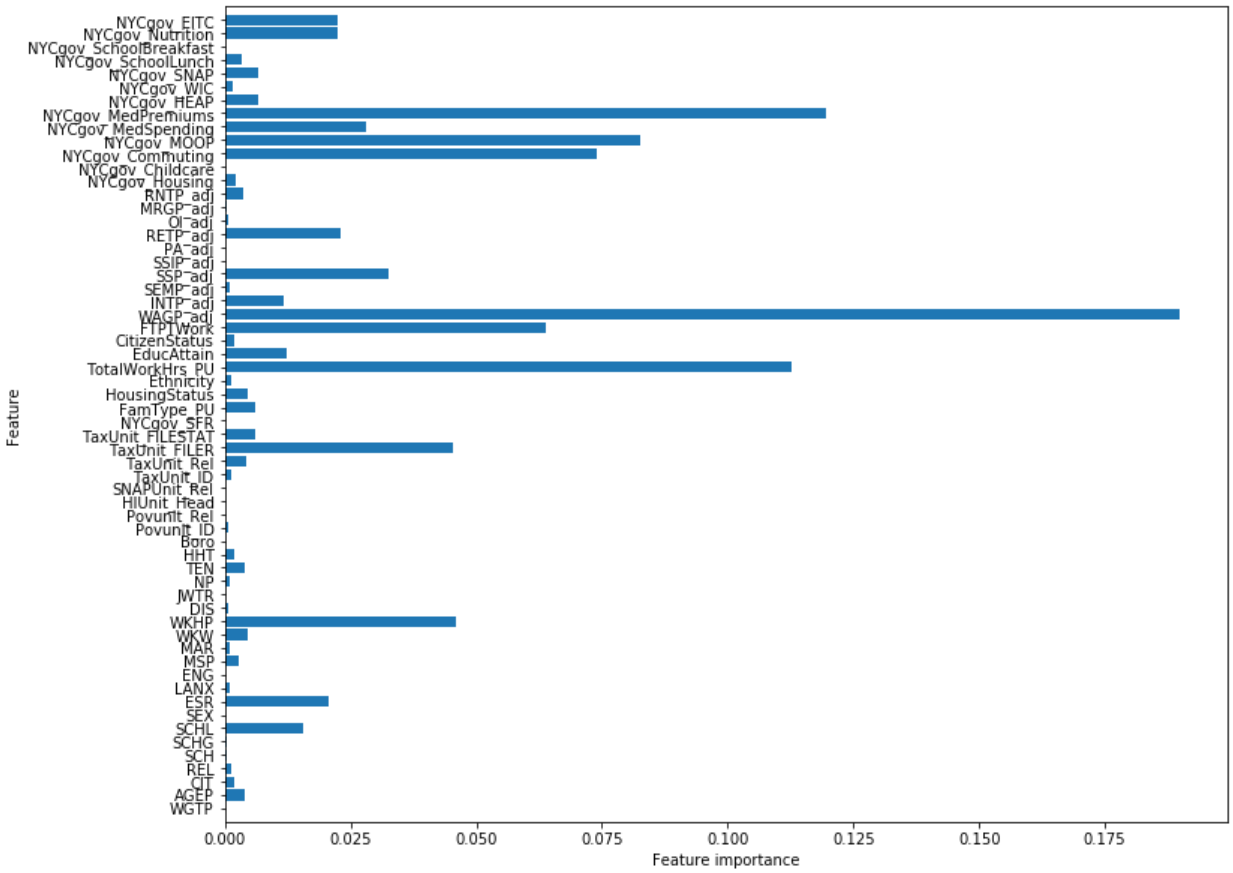
Random Forest:

Instead of only one decision tree, Random forest creates many decision trees and combine all the unique trees to get a more robust result. For this study, we used the bootstrap sampling technique created 1000 unique trees. Figure 4 shows one of the 1000 decision trees. As the graph shows, it picks different sets of attributes produced a



(Figure 4)

new unique tree. Another important information of the random forest is the feature importance, it tells which attribute has more information gain that is closer to the root of the decision tree. Figure 5 shows the feature importance of the random forest. As the plot shows, the top 5 attributes ranked by the importances are *wage, cost for medical insurance, working hours, total medical spending and commuting cost.*



(Figure 5)

IV. Result and Conclusion

The final result shows wage has the most information gain on deciding people's poverty status,followed by medical insurance cost and working hours. *WAGP_adj* measures individual's wage in past 12 months, base on the official website of New York State (ny.gov), at the end of 2019, the minimum wage in NYC will increase to $15,

which is roughly $28000 a year. Our decision tree shows a person that earns less than $23627 a year would more likely to falls into poverty, that means earning a minimum wage can guarantee a person stay out of poverty in most of the scenario.

One of the limitation of our study is the data may not be independent, for example a person with less wage income may also has less working hour and less money spend on the medica premium. Those three attributes potentially have high correlation with each other, this problem will cause the model underestimate the impact of other independent variables. One way to fix this is run the correlation map for all the variables and to identify the collinearity in the dataset. But this would be difficult for this dataset as it has majority of categorical variables.

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

"How to Handle Imbalanced Classes in Machine Learning"(July 5,2017), EliteDataScience. Retrieved from: https://elitedatascience.com/imbalanced-classes

Mayor's Office For Economic Opportunity. NYCgov Poverty Measure Data(2016). Retrieved from: https://data.cityofnewyork.us/City-Government/NYCgov-Poverty-Measure-Data-2016-/y9gu-cxxw