Team FitBit - Optimization

Rikka, Shiva Ram Reddy (00759744)
srikk1@unh.newhaven.edu

Chilukuri, Gayathri (00755590)
gchil4@unh.newhaven.edu

Potlapelli, Adithya (00761532)
apotl1@unh.newhaven.edu

November 12, 2022

GitHub Repository

1 Introduction

Poverty is one of the topics studied by many economists and data scientists. It is one of the economic issues that most governments wish to address. As one of the world's most powerful economies, the United States faces the challenge of domestic poverty. This project will concentrate on the poverty rate in New York City because it is one of the most representative cities in the United States, as well as having the country's highest income inequality. The project will make use of decision tree model analysis to predict poverty in New York and identify the features that can generate the most accurate predictions.

2 Dataset

The data utilized in this project is derived from the NYCgov poverty measure data, which is generated annually by the poverty research unit of the Mayor's Office of Economic Opportunity. This dataset has 68,644 observations and 79 distinct variables. The dataset includes various characteristics of New York City households (with unique identities), such as education levels, work status, annual income, and so on. Based on these characteristics, a decision is reached on whether or not a particular family is poor.

3 Data Mining - Modeling Technique

A decision tree is a supervised, non-parametric method that uses all available attributes to create the best prediction based on information gain. By computing the impurity of the parent and child nodes, the information gained can tell us the relevance of a specific property. Because we had a relatively large dataset (68644*79), we decided to use the Gini as our cost function for our analysis.

4 Parameters and Hyperparameters

- *criterion:* The function to measure the quality of a split.

- *splitter:* The method used to select the split at each node.
- *max depth:* The name of the hyperparameter max depth refers to the maximum depth to which we allow the tree to grow. The deeper you go, the more complex our model becomes; If we increase the max depth value, the training error will always decrease, but consider the case of testing error; if we set the max depth too high, the decision tree may simply overfit the training data, causing the testing error to increase, but if we set the value of it too low, the decision tree may have little flexibility to capture the patterns and interaction.
- *min samples split:* The minimum number of samples required to split an internal node.
- *min samples leaf:* The minimum number of samples necessary at a leaf node. A split point will be considered at any depth if it leaves at least "min samples leaf" training samples in each of the left and right branches. This has the potential to smooth the model, particularly in regression.
- *min weight fraction leaf:* The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Samples have equal weight when 'sample weight' is not provided.
- *max features:* The number of features to consider when looking for the best split.
- *max leaf nodes:* Grow a tree with "max leaf nodes" in the best-first fashion. Best nodes are defined as relative reduction in impurity. If None then an unlimited number of leaf nodes.
- *min impurity decrease:* A node will be split if this split induces a decrease of the impurity greater than or equal to this value.
- *class weight:* This actually means that when the algorithm calculates impurity to split at each node, the resulting child node is weighted by class weight by giving the child sample weight, distribution of our classes has been started then the weight of class and then depending on where our tree leans, we can try to increase the weight of the other class so that the algorithm penalizes the sample of one class relative to the other.
- *ccp alpha:* Complexity parameter used for Minimal Cost-Complexity Pruning. The subtree with the largest cost complexity that is smaller than "ccp alpha" will be chosen.

5 Optimization

Instead of just one decision tree, Random Forest builds several decision trees, combines all of the distinct trees, and bases the final prediction on the majority vote to get a more robust and significant outcome. For this study, we adopted the bootstrap sampling technique to generate 1000 distinct trees. Figure 1 depicts one of the 1000 decision trees. As the graph indicates, it selects different groups of traits to create a new unique tree. Another significant aspect of the random forest is the feature importance, which indicates which attribute has more information gain and is closer to the decision tree's root.

6 Outcomes and Visualization Techniques Used

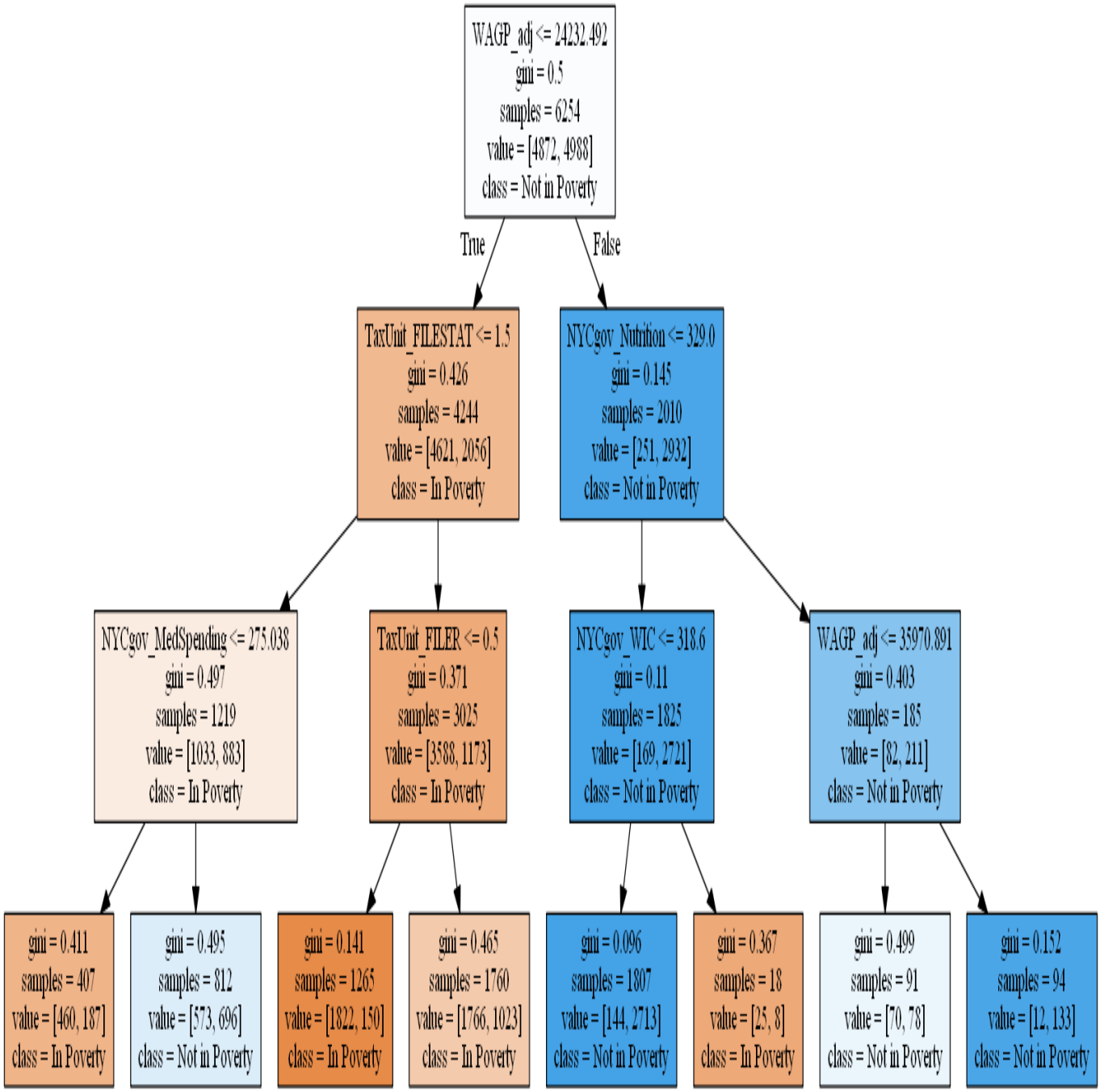

Figure 1: A Random Forest Tree

Figure 2 depicts the random forest's feature importance. According to the plot, the top five attributes in terms of relevance are wage, medical insurance cost, working hours, total medical spending, and commuting cost.

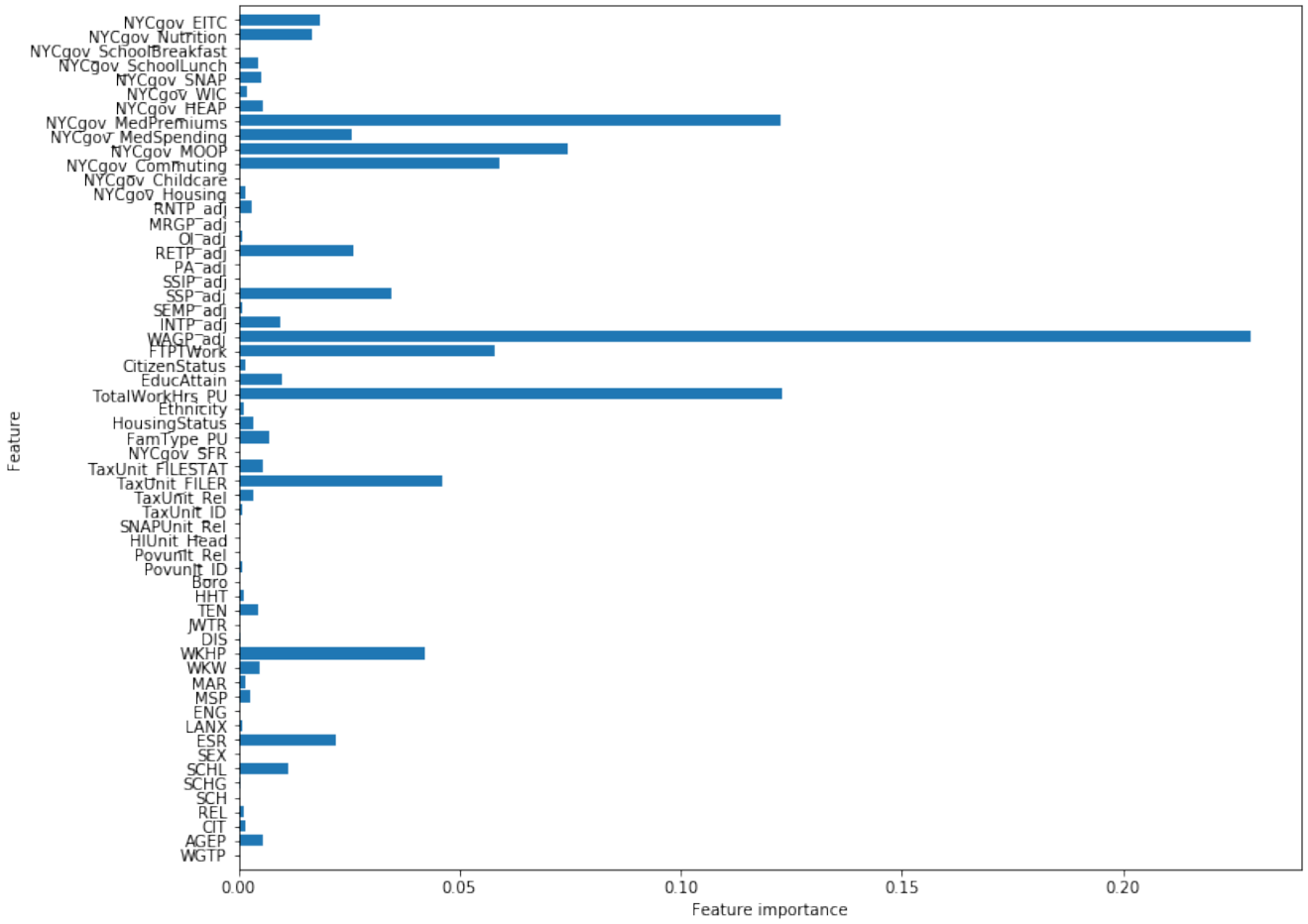

Figure 2: Random Forest - Feature Importance

7 Conclusion

The end result demonstrates that wages have the most informational value in determining people's poverty status, followed by medical insurance costs and working hours. According to the official website of New York State (ny.gov), at the end of the year, the minimum wage in NYC will increase to $15, which is approximately $28,000 per year. Our decision tree reveals that a person earning less than $23,627 per year is more likely to fall into poverty, implying that earning the minimum wage can keep a person out of poverty in the majority of scenarios.

One limitation of our study is that the data may not be independent; for example, a person with a lower wage may work fewer hours and spend less money on medical premiums. Because those three traits are possibly highly correlated, the model will understate the impact of other independent variables. One solution is to run the correlation map for all variables to identify collinearity in the dataset. However, because the majority of the variables in this dataset are categorical, this would be challenging.