Poverty Prediction in New York using Decision Trees

Adithya Potlapelli, Gayathri Chilukuri and Shiva Ram Reddy Rikka
Department of Electrical & Computer Engineering and Computer Science
University of New Haven
300 Boston Post Rd, West Haven, CT 06516
{apotl1, gchil4, srikk1}@unh.newhaven.edu

***Abstract*— Poverty is one of the topics that has been researched by a lot of economists and data scientists. This paper will make use of the decision tree model analysis to predict the poverty status in New York. The data used in this project is retrieved from the NYCgov poverty measure data. A decision tree is a non-parametric and supervised algorithm that goes through all the available attributes to make the best prediction. Variables that have a high correlation with the total income are dropped since income and poverty status shared the same characteristics. The decision tree analysis showed a person that earns less than $23627 a year would more likely to fall into poverty.**

***Index Terms*— Decision Tree, Random Forests, Correlation, Supervised Algorithm, Predictive Analysis**

I. 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.

This paper is structured as follows: Section II provides details on the literature survey. Section III explains the proposed method. In Section IV the details about experimental results have been provided; followed by the conclusion, future work, and references. The appendix provides the link to the GitHub repository.

II. LITERATURE SURVEY

In 2021 G. D. Singh et el. had a project on a data visualization approach for predicting the income class of the population [1]. The data for the research came from the 2011 Indian Census Data. This data relates to several states, among which Uttarakhand is considered, and included information on the population, sex ratio, literacy rate, religion, caste, etc., which are used as the input data. It employed Linear Regression and Data Visualization to quickly assess and estimate income classes based on factors impacting income in an area, such as population, gender ratio, type of opportunities available, age, and so on. All of these elements help predict an area's economic condition. A link between the two variables was discovered using linear regression. For example, the effect of an area's sex ratio on its income and possibilities. Based on the input value, it assisted in predicting the output value. The bar graph and pie chart visualization techniques also provided a better understanding of the linkages between various variables. The Gradient Boosting Classifier strategy was utilized to improve the technique's efficiency and accuracy in predicting numerous income groups. It had the highest accuracy rate of 85.78% on the test set. They mentioned that this model combines many techniques (data visualization, linear regression, coefficient of correlation) to create a more effective and better predictive model than all the previous work done in this field.

In 2021 Huang Zixi had done a paper on poverty prediction through machine learning [2]. The study looked at poverty as a result of multiple factors and proposed many feasible models for such predictions using machine learning, none of which accounted for the entire picture, while some attributes may outweigh others. As a result, an integrated prediction strategy based on data from the Poverty Probability Index and the Oxford Poverty & Human Development Initiative is required. The data that has been collected is derived from two sources: the Oxford Poverty & Human Development Initiative and the Poverty Possibility Index. The Poverty Possibility Index collects information from individuals. The data set contains 59 attributes for 12600 individuals, including information on education level, internet connection, age, number of financial activities in the previous year, and so on. The study analyzed the extent to which the parameters matter and the performance of each model by applying linear regression models, decision trees, random forest models, gradian boosting models, and neural networks on existing data. Cross-validation and grid research are used in the final advancement. The research concluded that, overall, gradient boosting has the highest accuracy of 78.53 percent on the test set for predicting poverty and education as the most influential factors.

In 2021 Huang Zixi had done a paper on poverty prediction through machine learning [3]. The study looked at poverty as a result of multiple factors and proposed many feasible models for such predictions using machine learning, none of which accounted for the entire picture, while some attributes may outweigh others. As a result, an integrated prediction strategy based on data from the Poverty Probability Index and the Oxford Poverty & Human Development Initiative is required. The data that has been collected is derived from two sources: the Oxford Poverty & Human Development Initiative and the Poverty Possibility Index. The Poverty Possibility Index collects information from individuals. The data set contains 59 attributes for 12600 individuals, including information on education level, internet connection, age, number of financial activities in the previous year, and so on. The study analyzed the extent to which the parameters matter and the performance of each model by applying linear regression models, decision trees, random forest models, gradian boosting models, and neural networks on existing data. Cross-validation and grid research are used in the final advancement. The research concluded that, overall, gradient boosting has the highest accuracy of 78.53 percent on the test set for predicting poverty and education as the most influential factors.

In 2021, Sheng B et el. had done a paper on a Feature-based Deep Neural Framework for Poverty Prediction [4]. The dataset contained anonymized data from poor families from 2013 to 2020. All of the training and test data utilized were subsampled from the Guangxi Province Data Platform of Poverty Reduction, which covered more than 7 million people and 1.6 million families. The research proposes a data-driven approach to capture the relation between poverty and related data, introducing Deep Poverty Forecast, a deep neural multi-channel model to encode multi-type features (DPF). The study used a star graph to represent family relationships and a data labeling method for supervised learning. The authors described extensive experiments done on five city datasets, and the findings showed that the suggested framework outperformed earlier methods utilizing two assessment metrics, Macro-F1, and Micro-F1. According to a local government survey, this method, implemented at the Data Platform of Poverty Reduction in Guangxi province, has covered 84 percent of new families falling into poverty while reducing the search space by more than 90 percent.

Navoneel Chakrabarty et el. in 2018 had done a research paper on a statistical approach to adult census income level prediction [5]. The data for the study was obtained from the UCI Machine Learning Repository at the University of California, Irvine. The data set contained information on 48,842 separate records and 14 attributes for 42 countries. The 14 attributes include information on age, education, nationality, marital status, relationship status, occupation, work categorization, gender, race, working hours per week, capital loss, and capital gain. The study proposed using the Ensemble Learning Algorithm, Gradient Boosting Classifier with extensive Hyper-Parameter Tuning, and Grid Search on Adult Census Data. The Validation Accuracy through this approach is 88.16 percent.

III. PROPOSED METHOD

A. Algorithm

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.

B. Preprocessing

Before we can create our decision tree, we must ensure that our dataset contains balanced target classes, the appropriate dimensions, and the most representative observations. In our investigation, we used feature selection and instance selection as preprocess approaches. The purpose of feature selection is to remove unnecessary or redundant characteristics in order to minimize the dimensionality of the dataset. Household Unit ID, Age Category, Poverty Gap, Tax Unit, and other factors are included. We also opted to exclude factors having a significant connection with total income because income and poverty status shared the same feature (low income directly decides whether the person is in poverty or not). If we add total income, the decision tree will select income as a criterion at each node, underestimating the influence of other factors.

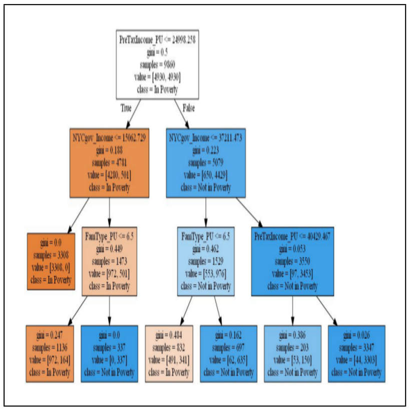
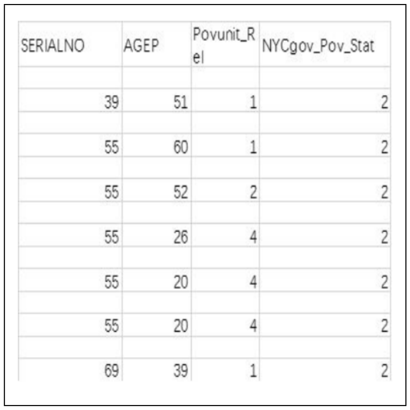


Fig. 1. Shows the model will pick Income 4 out of 6 times as the node

The instance selection will choose the most representative observations for the study. Our collection contains information on 70000 people and is organized by household unit. The table below displays the first seven observations from the dataset; observations with the same SERIAL NO indicate that the people belong to the same household. The Poverty Status of each individual is determined by the head of the family (Povunit Rel = 1); if the head of the family is poor, the rest of the family members are poor as well, regardless of other characteristics. In this scenario, we merely used the family head 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

Fig. 2. Table that shows 7 observations of the dataset

C. The Decision Tree

After preprocessing, we built a model with Sklearn and set the max depth to 3 to prevent overfitting. The decision tree is shown in the figure below. The accuracy(measured as the fraction of correctly classified samples) for the test set is 79%.

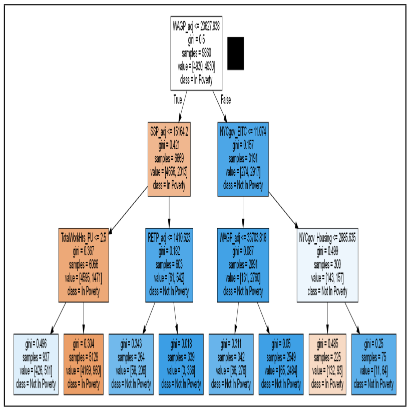


Fig. 3. Decision Tree on the dataset

It is understood that even if we preprocess the dataset before implementing the decision tree, the variance in the dataset might result in an unstable outcome. Only one tree is insufficient for our research. Instead of just one decision tree, create many decision trees, integrate all of the distinct trees, and base the final prediction on the majority vote to achieve a more robust and significant outcome.

D. Random Forests

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 4 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.

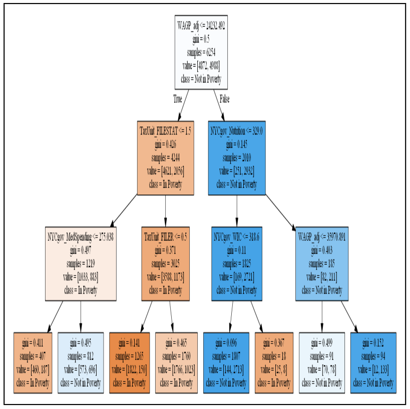


Fig. 4. Random Forest on the dataset

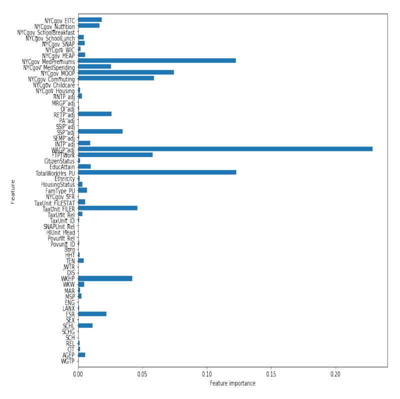


Fig. 5. Random Forest - Feature Importance

Figure 5 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.

IV. RESULT AND ANALYSIS

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.

V. CONCLUSION AND FUTURE WORK

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.

APPENDIX

Link to the project repository on GitHub
`https://tinyurl.com/CSCI6401DataMiningProject`

REFERENCES

[1] G. D. Singh, H. Vig and A. Kumar, "A data visualization approach for predicting the income class of the population," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2021, pp. 1042-1047, doi: 10.1109/ICECA52323.2021.9675850.

[2] H. Zixi, "Poverty Prediction Through Machine Learning," 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), 2021, pp. 314-324, doi: 10.1109/ECIT52743.2021.00073.

[3] Y. Xiao, "Predicting Poverty through Machine Learning and Satellite Images," 2021 International Conference on Digital Society and Intelligent Systems (DSInS), 2021, pp. 182-187, doi: 10.1109/DSInS54396.2021.9670555.

[4] S. B, S. Chen, H. Si, Y. Zhu, Z. Bai and S. Li, "A Feature-based Deep Neural Framework for Poverty Prediction," 2021 2nd International Conference on Computing and Data Science (CDS), 2021, pp. 568-573, doi: 10.1109/CDS52072.2021.00103.

[5] N. Chakrabarty and S. Biswas, "A Statistical Approach to Adult Census Income Level Prediction," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 207-212, doi: 10.1109/ICACCCN.2018.8748528.