Examining Income Disparity in Connecticut Using ML

Victor Cazabal vmc3@williams.edu
Statistics and Computer Science

1. Introduction

Using machine learning to tackle issues of social justice is a new concept without much research done behind it. Factors that lead to this gap may include that there's not much profit in this field or because when not applied correctly, machine learning models create bias that can discriminate against individuals.

This project aims to tackle the pressing issue of social inequality, with a focus on income disparity in Connecticut. Income disparity is a critical social issue that affects economic stability, quality of life, and access to opportunities. In Connecticut, one of the wealthiest states in the U.S., the contrast in income levels across different demographics and regions is particularly stark. The state is ranked as having the 5th largest gap; only the District of Columbia, Wisconsin, Minnesota and Iowa had larger gaps in their populations. Non-Hispanic white Americans have a median household wealth of \$139,300, compared to \$12,780 for black households and \$19,990 for Hispanic households (Kavaler, 2022). Understanding and addressing these disparities is vital for policymakers, social workers, and communities striving for a more equitable society.

Accurate predictions of income levels can, for example, guide policymakers in designing targeted economic and social policies. By understanding which factors most strongly influence income, governments and organizations can allocate resources more effectively, create more equitable tax policies, and design social welfare programs that better address the needs of different population segments. I hypothesize that machine learning models can get very close to accurately predicting individuals' total income based on various factors; helping to create more focused and effective policies to lessen income inequality.

2. Preliminaries

I will be using a couple of machine learning algorithms to find a model that best predicts total income of individuals living in CT. I will be using packages such as sklearn to implement these.

I will first be implementing decision trees (Breiman et al., 1984). Decision trees are a non-linear, hierarchical model used in machine learning for both classification and regression tasks. A decision tree recursively partitions the input space to make predictions about the target variable Y. Each node in the tree represents a decision rule based on one of the input features $X = x_1, \ldots, x_d$, and each leaf node represents a prediction $\widehat{y}$.

For a given example with input features $x_1, \ldots, x_d$, the decision tree model routes the example through the tree, following the decision rules at each node until reaching a leaf node. The prediction $\widehat{y}$ is then the output associated with this leaf node.

The tree is constructed by selecting the splits that maximize the reduction in a loss function L, typically chosen to be mean squared error for regression tasks, defined as $L(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$. The decision of how to split at each node is based on finding the feature and threshold that produce the most significant improvement in L.

To prevent overfitting, techniques like pruning are used. Pruning involves cutting back the tree to a certain size or depth, which is controlled by a hyperparameter. The goal is to balance the model's complexity with its ability to generalize well to new data.

Then, I will compare the decision tree model to a random forest model (Breiman, 2001). Random Forests are an ensemble learning method which operates by constructing a multitude of decision trees at training time and outputting the average prediction of the individual trees for regression tasks. Given a set of input features $X = x_1, \ldots, x_d$, the Random Forest model aggregates predictions from multiple decision trees to produce a final prediction $\widehat{y}$.

Each decision tree in the Random Forest is built on a bootstrap sample, which is a random sample with replacement from the training data. The prediction for a given example by a Random Forest model is the average of the predictions from all the individual trees:

$$\widehat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x_1, x_2, \ldots, x_d),$$

where T_b is the prediction of the b-th tree, and B is the total number of trees in the forest.

Each tree in a Random Forest splits nodes based on a randomly selected subset of features and determines its splits to minimize the loss function, the mean squared error, as mentioned before.

Random Forests tackle overfitting effectively due to the averaging of multiple trees, each trained on different samples of data. This averaging process leads to improved generalization ability and robustness compared to individual decision trees. The hyperparameters, including the number of trees B and the maximum allowed depth of each tree, are tuned for optimal performance. The use of multiple trees and random feature subsets also provides a measure of feature importance based on how frequently a feature is used to split nodes across all trees.

In this project, random forests are utilized for their ability to capture complex, non-linear relationships in the data, making them suitable for the task of income prediction based on a diverse set of factors.

3. Data

This analysis leverages data from the American Community Survey (ACS) from 2018-2021 to predict income using a variety of features that capture personal characteristics. Using RStudio and the tidycensus package, I was able to download ACS data directly onto my

PUMA	Age	Sex	Employment	Race	Children	HoursWorked	Income
101	24	2	1	1	0	40	20598.56
101	45	2	1	1	0	55	200835.96
1101	48	2	6	1	0	0	0
1100	54	2	1	1	0	40	51496.4
1100	49	1	1	1	2	50	87543.88
500	37	2	1	1	1	40	20598.56
304	55	2	6	1	0	0	578.33
102	78	2	6	1	0	0	30897.84
906	62	2	6	2	0	0	76214.67
905	20	1	6	9	0	0	0

Table 1: First 10 rows of 2021 data on CT Individuals

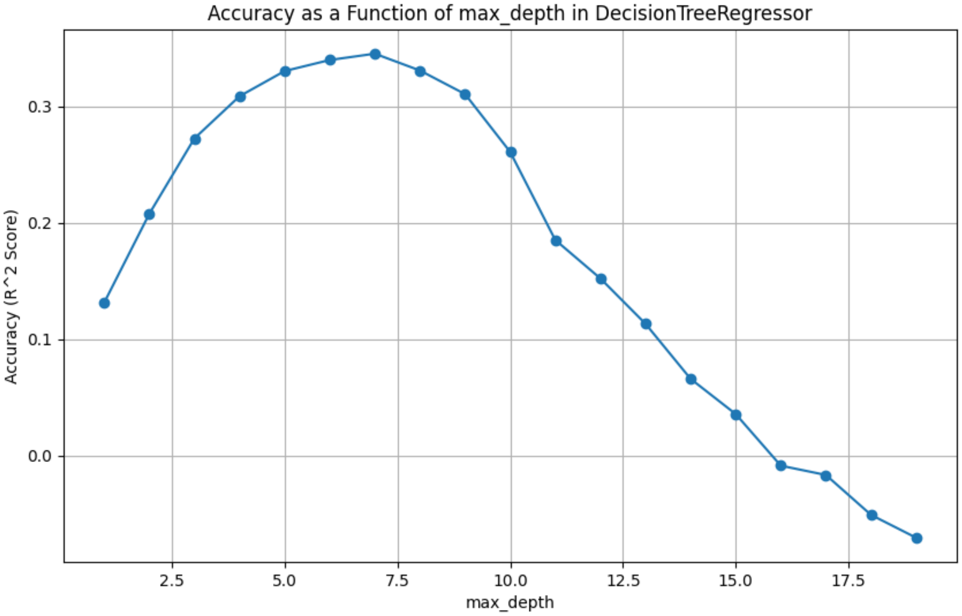
computer with the ability to choose any variables I want from individuals surveyed. Since the ACS is a national survey, I narrowed down my data to just CT individuals. The ACS is conducted and sponsored jointly by the U.S. Census Bureau. Extracted from IPUMS, Data from 2018 is used to train the models, 2019 data is used as a validation set and 2021 data is used as the final test set. Above, I have provided a sample of the data from 2021 that just includes 7 of the 32 variables used to predict total individual income (To be clear, a PUMA, or Public Use Microdata Area, is a statistical geographic unit used in the U.S for the analysis of ACS).

The ACS contains a plethora of information from each respondent and their household. Of these, I chose 32 variables that may have an influence on total income of an individual. After completing this study, I hypothesize that there are still more variables available in the ACS that hold strong predictive power for total income. Due to the limitations of this project being worked on my Mac and being the only person working on it, it is a challenging task to preprocess all the variables available in the ACS. Instead, I used my experience working with Census data at my internship this past summer to judge which variables might affect total income "the most", although I may have missed several. I have included 150 variables extracted capture a variety of characteristics for individuals from categories such as work, income, education, ethnicity, disability status, migration status, family interrelationships, welfare benefits, and veteran status. These original variables were recoded to create 142 unique variables that were used in the analysis. These new variables created include a large number of dummy variables created for categories such as state of residence, and occupation of the individual. The data used in the analysis contained 92, 368 variables and 142 recoded variables after cleaning.

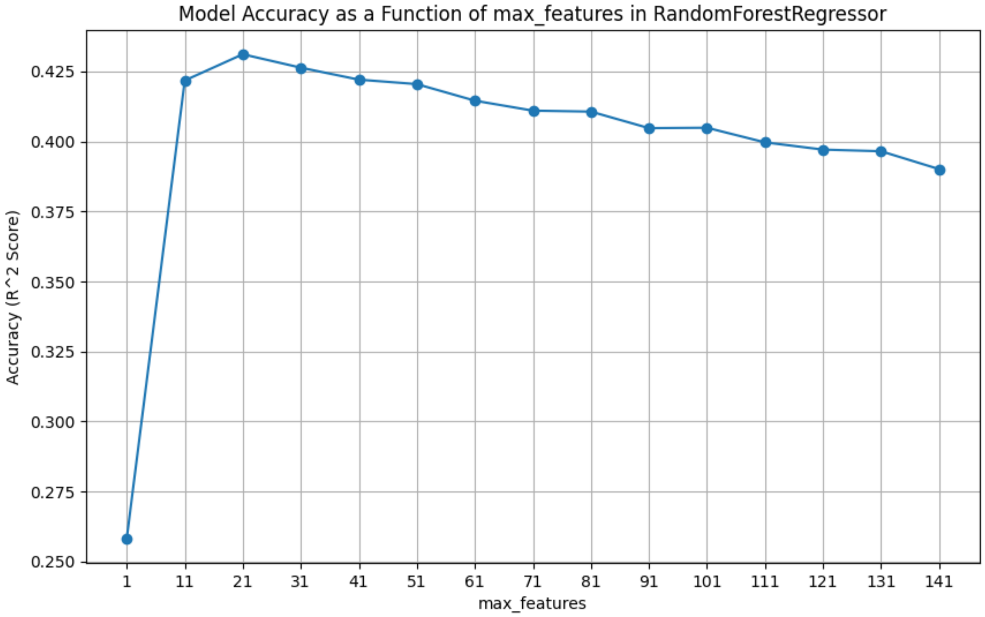
The response, or target, variable for the analysis is the real income for individuals. The information from the variable originally came from PINCP, which is the self-reported total personal income for an individual. According to the documentation, “PINCP indicates each respondent’s total pre-tax personal income or losses from all sources for the previous calendar year” (ACS IPUMS, 2020). While inflation may not have a large impact on the analysis since the data is over a period of four consecutive years, it is important to recognize that the incomes reported are in nominal terms and not real terms. Therefore, a deflator is used to account for inflation so that the incomes can be effectively compared. The deflator used was the ”Adjustment factor for income and earnings dollar amounts”, as provided by the PUMS Data Dictionary. Using this adjustment allows the dollar amounts to be compared across years of surveys. Only individuals with positive real incomes were then used in the analysis. This assisted with helping to increase interpretability of values for the metrics of evaluation.

4. Training And Validation Of Models

There isn't much literature on predicting total income based on population characteristics, much less on individuals from CT. Because of this, I have decided to train a simple linear regression model to use as baseline. The linear regression model reached an impressive R^2 of 0.338. We will be using this R^2 to compare accuracies between the baseline and other, more complex machine learning models.



The first machine learning model I will be training and then validating is the decision tree regressor using skicit-learn (Breiman et al., 1984). The graph above shows how well my decision tree model generalizes with different max depths. Based on the graph above, the best accuracy achieved is 0.345 with a max depth of 7. This algorithm is technically successful when compared against the baseline model, however it doesn't show the increase in generalization one might hope for.



The second machine learning model I will be training and then validating is the random forest regressor, again using skicit-learn (Breiman, 2001). The graph above shows how a default model of number of estimators set at 100, and max depths of each tree set at 20 does with various options of max features. This graph helped me reach an ideal balance of max features and max depth. The highest accuracy I achieved is with n estimators set at 500, max features set at 20 and max depth set at 20. This model achieved an R^2 of 0.435, thus being our most optimal model and will be the one that we're testing on 2021 data.

5. Results

Now that we have selected our most optimal model, we will be testing the random forest regressor on unseen data for the 2021 calendar year. These results are particularly interesting because we're using 2018 data to predict income from 2021, which is the year when Covid-19 was at an all-time high and unemployment was skyrocketing. Unsurprisingly, our model generalized a bit less successfully, achieving an R^2 of 0.415. This isn't a huge difference (it only being 2% less), however. This indicates the robustness to variability that the random forest regressor is capable of withstanding. Compared to our baseline of 0.338, we were able to achieve a 0.077 increase in accuracy.

To gain deeper insights into the model's performance, an analysis of feature importance was conducted. Random Forests provide a measure of how much each feature contributes to the accuracy of the model. The most influential features in predicting passive income were found to be the more hours that you worked, the older you are, if you worked in finance, whether you had a professional degree beyond a bachelor's degree, if you hadn't been married yet and if you lived in Fairfield County. This aligns with previous stereotypes of "rich" people, even exposing the disparity between Fairfield County and the rest of the counties in CT, thereby validating the model's ability to identify key predictors of income.

6. Ablation Study

In this ablation study, we focus on understanding the impact of two key features: industry columns and the 'Hours Worked' column versus the impact of a column that is known to affect income at a systematic level: sex. These features were chosen based on their perceived importance in predicting individual income. The study was conducted by removing these features one at a time from the dataset and then retraining the Random Forest model under the same conditions as the original model. The performance of each modified model was then compared to the original model's performance.

The first part of the study involved removing the industry columns from the dataset. These columns provide information about the sectors in which individuals are employed, which is intuitively a significant factor in determining income levels.

The original model, which included these columns, achieved an R^2 score of 0.435 on the testing set. After removing the industry columns, the model's accuracy decreased to an R^2 score of 0.394, a drop of 4%. This significant decrease in model performance underscores the importance of the industry data in predicting income levels. The industry in which an individual works is a strong indicator of their income bracket, and its absence in the feature set evidently hinders the model's predictive capability.

The second part of the study focused on the 'Hours Worked' column. This feature represents the number of hours an individual works per week, a direct and influential factor in determining their income.

Omitting the 'Hours Worked' column resulted in the model's accuracy decreasing to an R^2 score of 0.378, a decline of almost 6% compared to the original model. This result highlights the critical role that the amount of work, quantified in hours, plays in income determination. The 5% decrease in accuracy points to a model that becomes considerably less effective at predicting income when it lacks data on the work hours of individuals.

Finally, omitting the 'Sex' column result in the model's accuracy decreasing to an R^2 score of 0.402, a decline of over 3%! This is not too far off a decline an inherently influential variable such as 6% from hours worked.

The substantial impact of removing the 'Sex' column highlights the underlying gender disparities in income. This finding aligns with existing research on gender wage gaps and emphasizes the importance of considering demographic factors when analyzing income levels.

7. Discussion and Conclusion

This project, centered on using machine learning to examine income disparity in Connecticut, has underscored the potential of data-driven approaches in addressing social justice issues. One of the most significant findings was the impact of certain features - industry, hours worked, and gender - on income predictions, highlighting the critical role of careful feature selection in predictive modeling. The performance of our Random Forest model, particularly its robustness in comparison to the baseline linear regression model, was enlightening. However, the model's slightly reduced accuracy during the COVID-19 pandemic underscores the challenges faced by machine learning in rapidly changing societal contexts, emphasizing the need for models that can adapt to economic fluctuations.

Looking ahead, exploring deep learning techniques and incorporating time-series analysis could yield deeper insights, particularly in understanding how income levels evolve over time. Additionally, a more granular data approach, considering Connecticut's diversity, might offer more detailed insights. Despite the project's successes, it highlighted areas for future improvement, including broadening the scope of feature exploration and continuously addressing data privacy and ethical considerations. This exploration into income disparity not only provided technical insights but also touched on the broader social and ethical implications of using machine learning in the realm of social justice. The lessons learned here advocate for a responsible and ethically aware application of data science tools, aiming to contribute positively to societal equity and understanding.

References

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. *Classification and Regression Trees*. Routledge, 1984.

Bernard Kavaler. Connecticut has among nation's largest wealth gap by race and ethnicity, Jan 2022. URL `https://ctbythenumbers.news/ctnews/connecticut-has-among-nations-largest-wealth-gap-by-race-and-ethnicity`.