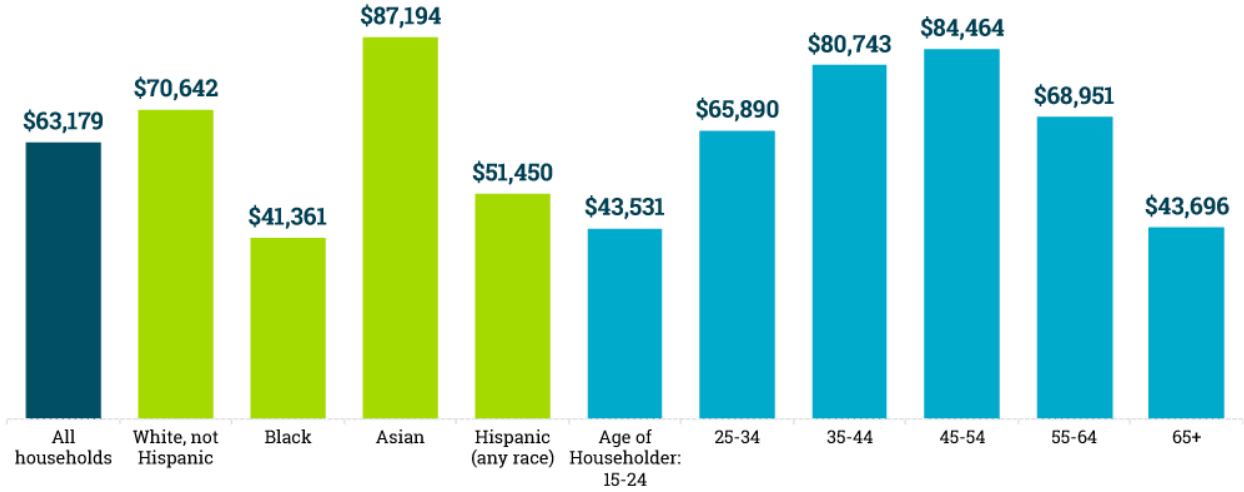
Household Income in the United States

STAT 471 Final Project



Victor Castillo and Ethan Reiser

December 19, 2021

Github Repository:
https://github.com/victorca2000/household-income-project

Executive Summary

Poverty is a pervasive problem in the United States, impacting over 37 million Americans in 2020 according to the United States Census Bureau. Given that households are determined to be impoverished or not based on their level of household income, we sought to discover which factors were the most predictive of total household income. For our final project, we constructed various models in order to best predict total household income based on a variety of socioeconomic characteristics of households in the United States including geographic region, health insurance coverage, access to a telephone, transfer payments from the government, as well as many others.

Our dataset comes from the 2018 Current Population Survey Annual Social and Economic Supplement (CPS ASEC) conducted by the United States Census Bureau. In order to collect this data, the Census Bureau surveyed over 92,000 households across 1300 counties in the United States. While we do believe that this data includes many factors that contribute to total household income, and hence poverty, we acknowledge that the data incorporates one-time income sources into the analysis such as severance pay or survivor income, which may slightly alter the prediction of the model. In order to maintain consistency, we removed all observations from the data where the household did not complete an interview, leading to all null values being removed from the dataset. We were left with approximately 67,000 observations, which were then used to construct our predictive models.

Before completing our analysis, we performed some data exploration in order to further examine the distribution of our response, as well as the relationship between the response and individual features. Upon completing this analysis, we decided to transform total household income onto the log-scale in order to prevent violations of normality assumptions. We subsequently split our data into a training dataset and a test dataset based on an 80-20 split. Using the training dataset, we constructed seven predictive models, four of which are regression methods and the remaining three being tree based methods. In order to determine which model performed the best, we used the test data to compute the root mean squared error (RMSE) for each of our predictive models.

After conducting our analyses, we found that the regression-based methods as well as the tree-based methods selected similar variables as having significant predictive power. These variables included existing sources of income (self-employment income, wages), property value, private healthcare coverage, and the composition of the household (married couple, nonfamily householder). Unsurprisingly, variables related to the possession of certain assets such as the ownership of stock, property value, and investments in money market funds and savings bonds, were predictive of greater total household income.

Introduction

Poverty is a crippling socioeconomic problem in the United States that continues to plague millions of people annually. According to the United States Census Bureau, the poverty rate in 2020 climbed slightly to 11.4%, indicating that over 37 million Americans live below the poverty line.[1] A different study conducted by the Social Security Administration found that nearly half of working Americans earn less than $35,000 in annual wages.[2] In most American metropolitan areas, these earnings would just barely allow Americans to make ends meet and provide for their families. The poverty trap that exists in the United States directly contributes to other pervasive socioeconomic issues such as food insecurity, malnutrition, homelessness, and lack of access to proper health care.[3] Hence, from a young age, a sizable percentage of people across the United States are entrenched in this poverty trap, spending their lives attempting to remove themselves from it.

Given that poverty thresholds are determined based on household income, it is important to analyze available data to determine which social and economic factors impact household income the most. Thus, we are specifically interested in how the various socioeconomic features presented in the data (e.g. geographic region, metropolitan area size, access to a telephone, value in food stamps received, etc.) impact household income. For each of our predictive models, we will compute the root mean squared error using the test data. We will deem the model with the lowest test root mean squared error as the most successful in accurately predicting total household income.

Determining the relative importance of factors contributing to low levels of household income will allow economists and government officials to craft better place-based as well as people-based policies in order to better address systemic poverty. Furthermore, we hope that our analysis will initiate a larger discussion regarding supplemental income and transfer payments, as welfare expenditures in the United States constitute the largest percentage of the federal budget, totalling approximately $2 trillion annually.[4]

---

[1] Income and Poverty in the United States: 2020

[2] Half of American Workers Made Less Than $35,000 in 2019, Report Shows

[3] Effects of poverty, hunger and homelessness on children and youth.

[4] http://www.usgovernmentspending.com/us_welfare_spending_40.html

Data

Data Sources

Our dataset comes from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) conducted by the United States Census Bureau. Each year, the Census Bureau collects this data in order to create income, poverty, and health insurance coverage estimates based on a detailed questionnaire that they pose to a sample of households throughout the United States.

For this data analysis, we are using the 2018 CPS ASEC data, which surveyed each individual across approximately 92,000 American households regarding various social and economic factors. The CSP ASEC estimates sample households within 1300 American counties out of a total of 3100 counties across the nation. Furthermore, all observations within the dataset were collected in the same month (March 2018), which ensures that responses for different households were not reflective of sudden economic booms or shocks.

Data Description

Our dataset has 92,138 observations, corresponding to each of the households surveyed by the Census Bureau. Our response variable is household income, measured in US dollars. We chose this variable as the response, as household income determines poverty status. However, household income in this dataset includes one-time payments such as severance payments or survivor income, which may slightly alter the predictive models. Upon cleaning the dataset, we included 88 features into our analysis. A short explanation of each variable's meaning and its classification as categorical or continuous can be found in the appendix section at the end of the paper.

Data Cleaning Process

Our central task for the data cleaning process was to remove the columns that were not useful for prediction. This included any allocation flag column for the respective feature, as well as any variable that did not offer any demographic or geographic information. This included identifiers and indexing information.

We first filtered the dataset to only include those rows of households that were interviewed, removing all cases where values were completely imputed. We then considered removing the households with a total income of 0, as the 0's could have been a replacement for nulls in the rawest form of the dataset. Since removing these households only accounted for 0.97% of the rows, we decided to remove them.

After, we began removing unwanted columns, starting with the allocation flags. This was done through regex matching, as all allocation flag columns started with “H1” or “I_”. Next, there were 20 known variables that added up to the total income. The linear combination of these features were correlated against the total income, showing a correlation matrix of 1s and confirming that those 20 variables were indeed just addends of the total income. Therefore, these variables were removed. Finally, columns that either had no variation in their values or were insignificant for prediction were removed.

Finally, we had to one-hot encode categorical variables to make them suitable for prediction. Not doing so would have implied that the categories, which were in terms of numbers representing a class, had quantitative meaning. Since this was not the case, all categorical variables were changed to dummy variables. A total of 35 variables were one-hot encoded, all typically having ranges from 2-4. These columns were also selected through regex filtering, as numerical variables have specific string patterns.

After applying these changes, the cleaned dataset had a total of 154 columns and 67,015 rows.

Data Transformation

Through our exploratory data analysis, we decided to review the distribution of the total household income. We noticed that this variable was extremely skewed to the right, which was an issue given that it is our response variable. Knowing that modeling usually yields better predictive performance when the response variable is normally distributed, we transformed total household income by taking its base-10 log. Figure 1 below shows the difference in normality for total household income before and after transforming it. Using a log transform yielded a more normal distribution. Since we had already filtered out all rows with total household income equal to zero, we did not create null values by transforming the variable.

Data Allocation

For the purposes of our analysis, we randomly sampled 80% of our observations to be included in the training data, with the remaining 20% allocated for test data. We set a seed before splitting the data to ensure that our results would be able to be reproduced if the analysis were to be performed again.

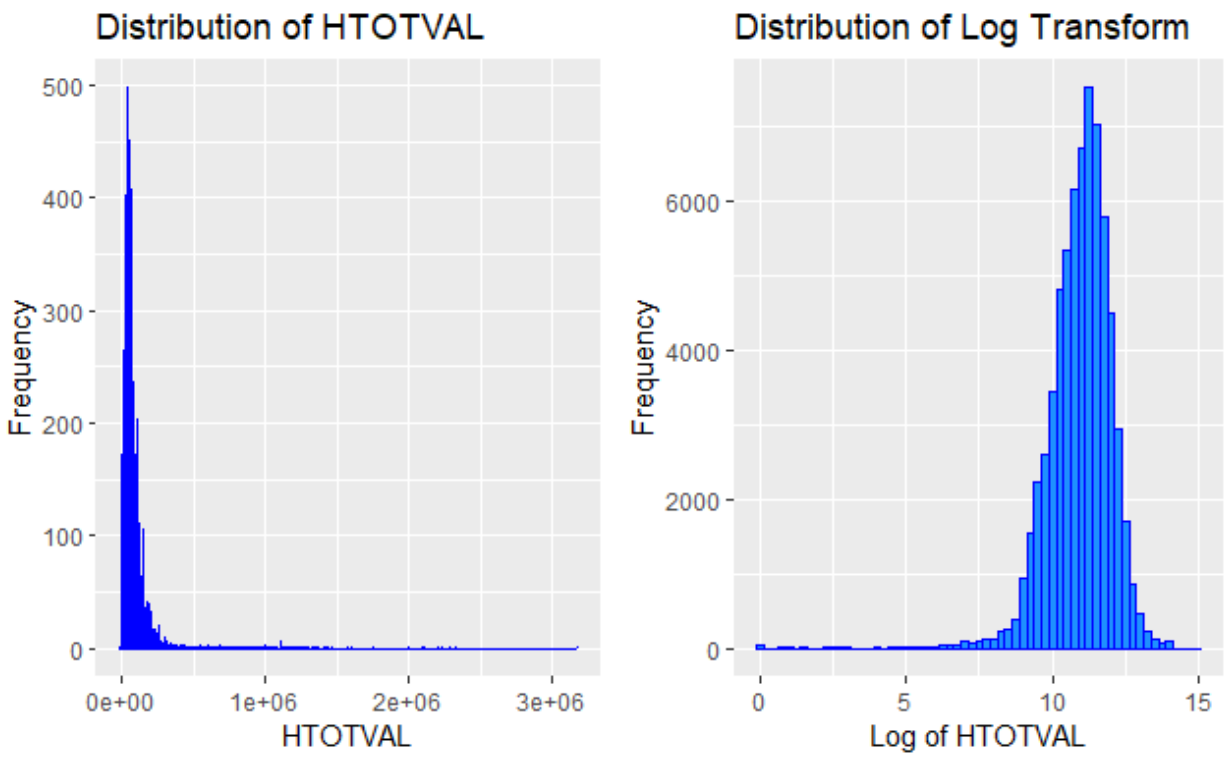


Figure 1. This figure shows the distribution of total household income before and after the log transformation.

Feature Selection

A highly dimensional dataset can create problems when performing predictive analysis, as it can lead to massively overfitting the model. It can also make every point in the model seem equidistant from each other. Unfortunately, our cleaned dataset still contained 154 columns, which hinted that our dataset had high dimensionality. Having too many columns meant we were asking our models to train for too many parameters. Therefore, we chose to conduct some feature selection and plotted a correlation matrix of all features to assess whether there were pairs of features that were highly correlated. Two features that are highly correlated to each other would most likely have the same effect on the response variable. Therefore, removing one would reduce the model complexity without jeopardizing accuracy. Figure 2 presents the plot for the correlation matrix. Although pairs of features seemed uncorrelated, indicated by the high area of white space, there were some prominent dark spots that meant some pairs of features were too correlated. Having a cutoff correlation of 0.7, one of the two features for each of these highly correlated pairs was therefore removed. This reduced our number of features for prediction from 153 to 88.

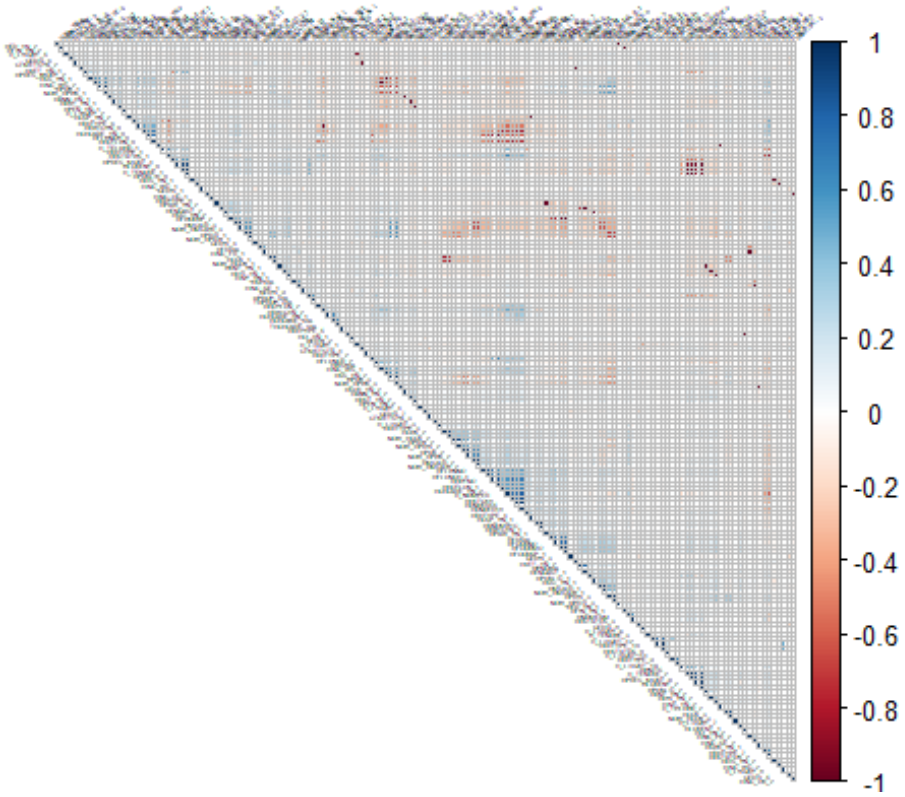


Figure 2. This is a plot of the correlation matrix of all features in the cleaned dataset to assess whether feature selection is needed.

Data Exploration

We decided to conduct some exploratory data analysis to better understand the features we were working with and how these relate to the total household income. Our most interesting findings are found below.

Household Income Versus Metropolitan Status:

We first wanted to understand the relationship between household income and whether the household was located in a metropolitan versus a non-metropolitan area. We expected that household income was going to be significantly higher in the former, as we expect higher salaries in high urbanized areas. However, Figure 3 shows that the difference in median household income for those in metropolitan areas only differs by 0.2, remember that these values are in log scale. The median of the log incomes for metropolitan versus non-metropolitan households are 11.1 and 10.9, respectively. The distribution for metropolitan households also extended slightly more than that for non-metropolitan households, as the former contained more

outliers. The range for the metropolitan distribution is 15.0, while that for the non-metropolitan distribution is 14.6.

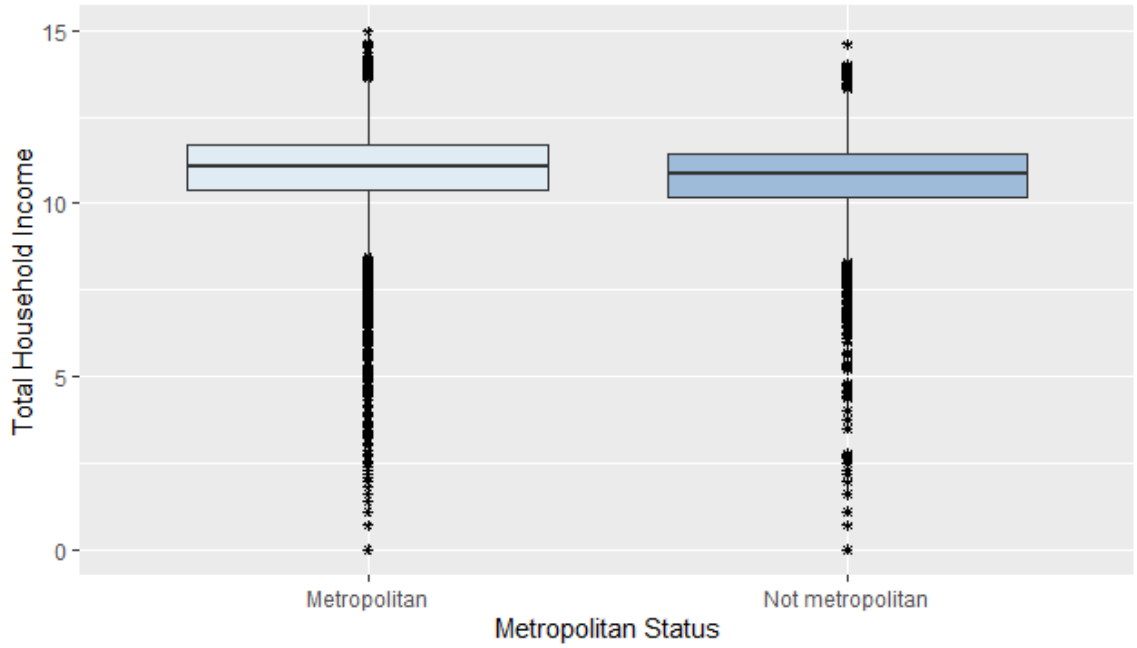


Figure 3. This figure shows side-by-side boxplots of the total household income distribution for metropolitan versus non-metropolitan households.

Household Income Versus Property Value For Each Region:

We assumed property value would be a good range for household income. Generally, households with higher income tend to spend more on property, buying those that are of higher value in the market. Figure 4 depicts a scatter plot between household income and property value for each of the nine regions reported in the dataset. All regions showed similar ratios between income and property value, because mean property value tends to be proportional in some way to mean household income for each region. The shape of the curve trends were fairly horizontal and linear for all plots, showing that in general property value is not too representative of total household income. This trend would probably be different if considering salary rather than income, as the latter accounts for additional monetary factors.

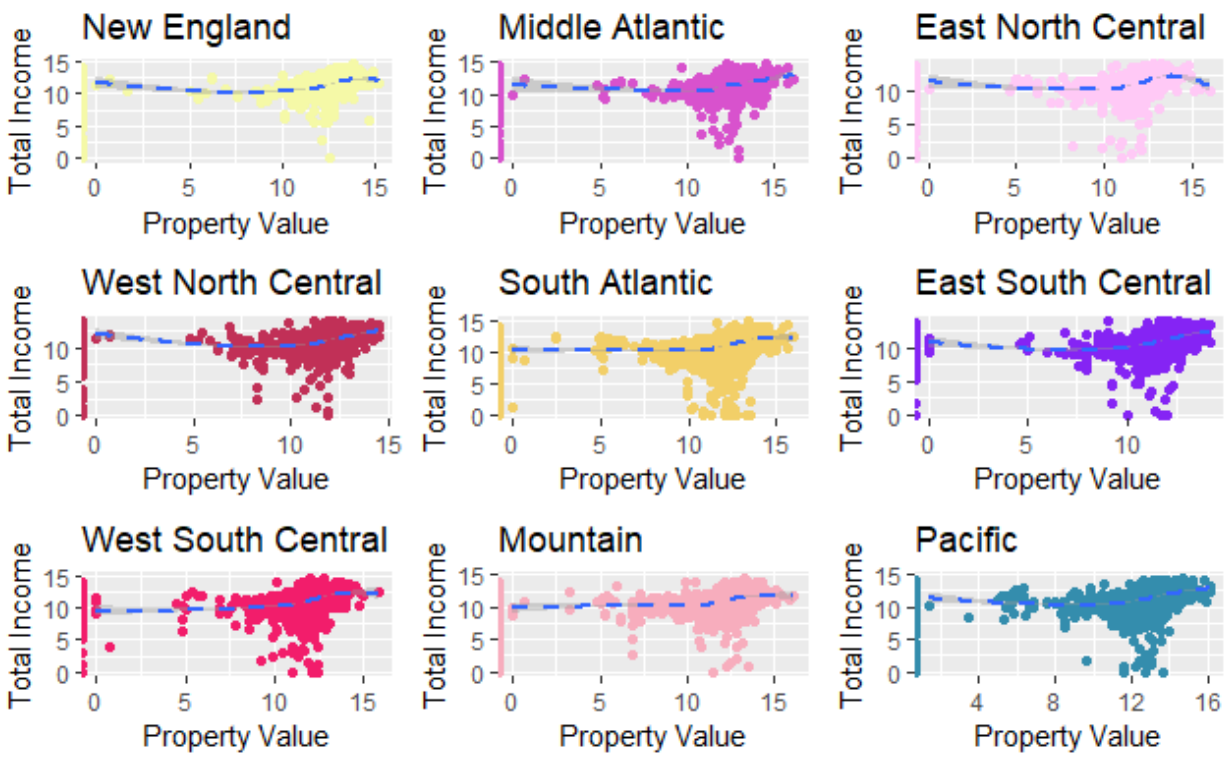


Figure 4. This figure shows the relationship between household income and property value in each of the nine regions of the United States.

Medicaid Eligibility And Government Assistance With Rent:

We also performed some exploratory analysis to study the relationship between features. Generally, it is expected that those households that need government assistance with rent are also enrolled in Medicaid, as this insurance is income-based to help households with financial struggles. Figure 5 compares the number of households not eligible for Medicaid for those being assisted versus not being assisted with rent. The plot effectively shows the difference between these two groups: there are over 50,000 households not eligible for Medicaid who do not receive government assistance for rent, while there are less than 1,000 households not eligible for Medicaid who do receive government assistance for rent. This validates the household selection process for government assistance with rent and Medicaid, as both typically target the same households: low-income ones.

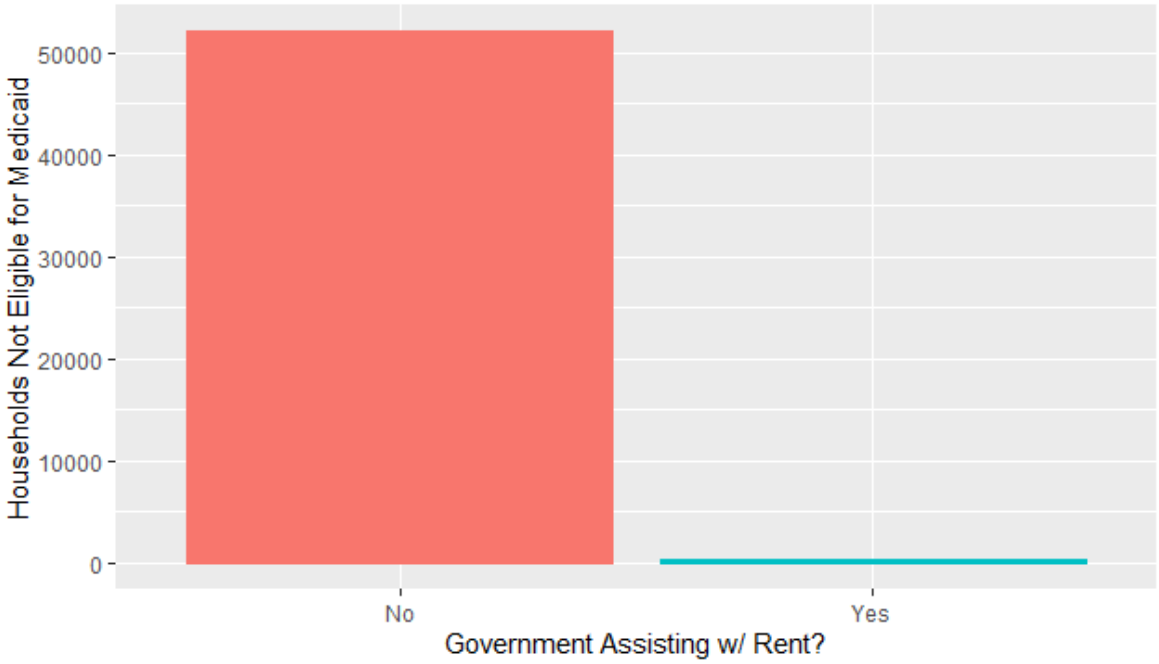


Figure 5. This plot shows the difference in number of households not eligible for Medicaid between groups depending on whether they receive government assistance for rent.

Modeling

Regression-Based Methods

Ordinary Least Squares

We began our regression-based predictive methods with an ordinary least squares (OLS) regression of total household income on all 88 features selected. The maximum residual value computed was 3.0458, and the lowest was -10.5365. The median residual was 0.0748. The model had 53,513 degrees of freedom, and presented a multiple R-squared of 0.6215. This R-squared is fairly good, as it explains that 62.15% of the variability in total household income is accounted for by the variability in the features. The p-value computed for the OLS regression model was below 2.2e-16, indicating that there is enough evidence to suggest that the slope of the regression line is not zero. There is a statistically significant relationship between total household income and the features.

Ridge Regression

We then proceeded with the penalized regressions, starting with ridge. Penalizing the regression would reduce the model complexity, which would resolve any issues generated from the high dimensionality of our data. A ridge regression shrinks the feature coefficients to a value close to zero, but does not operate on any feature selection. We first trained the model with a 10-fold cross-validation to select the lambda value that would reduce the CV error the most. This lambda value was 0.06895063. According to the one-standard error (1SE) rule, we chose 0.2783549 as our lambda value to make the model simpler. The left plot on Figure 6 represents the CV error at each lambda value: the left dashed line resembles the lambda value that minimizes the error, yet the right dashed line represents the lambda chosen by the 1SE rule.

We computed the coefficients for each of the 88 features using the 1SE rule. This can be visualized by the standardized coefficients falling on the black dashed line on the right plot on Figure 6. The highlighted curves represent the ten largest coefficients in magnitude with the chosen lambda: HDIV_YN_2, HINC_WS_2, HPEN_YN_2, HPROP_VAL, HSEVAL, HHSTATUS_2, HINT_YN_2, HPRIV_3, HRHTYPE_1, and HWSVAL.

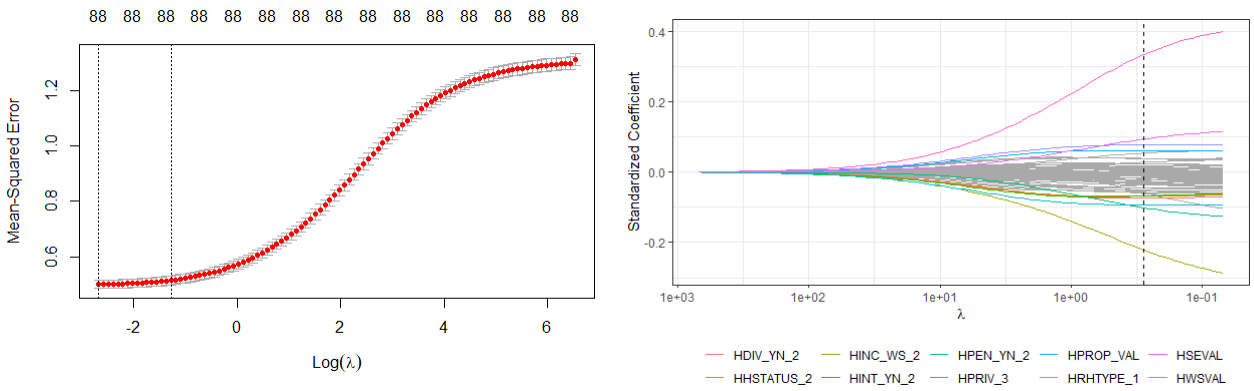


Figure 6. This figure shows the CV plot for the ridge regression, as well as a plot of the standardized coefficients for each feature at different lambda values.

Lasso Regression

The lasso regression showed similar results to the ridge regression in terms of coefficients. However, we must remind ourselves that lasso not only shrinks coefficients, but also performs feature selection by sparsing out the coefficient matrix. Similar to ridge, we first trained the model with a 10-fold cross-validation to select the lambda value that would reduce the CV error the most. This lambda value was 0.0006430359. According to the one-standard error (1SE) rule, we chose 0.02009947 as our lambda value to make the model simpler. Instead of using 88 features, the lasso fit was only performed on 35 selected features. The left plot on Figure 7 represents the CV error at each lambda value: the left dashed line resembles the lambda value that minimizes the error, yet the right dashed line represents the lambda chosen by the 1SE rule. The values above represent the number of nonzero coefficients based on feature selection.

We computed the coefficients for the features selected by lasso using the 1SE rule. This can be visualized by the standardized coefficients falling on the black dashed line on the right plot on Figure 7. The highlighted curves represent the ten largest coefficients in magnitude with the chosen lambda: HDIV_YN_2, HINC_WS_2, HPRIV_3, HRHTYPE_1, HWSVAL_ HHSTATUS_2, HINT_YN_2, HPROP_VAL, HSEVAL, NOW_HMCAID_1. The lasso and ridge regressions share 9 out of the 10 features with the largest coefficients in magnitude, meaning that the features highlighted on the right plots of Figures 6 and 7 are consistently indicative of total household income.

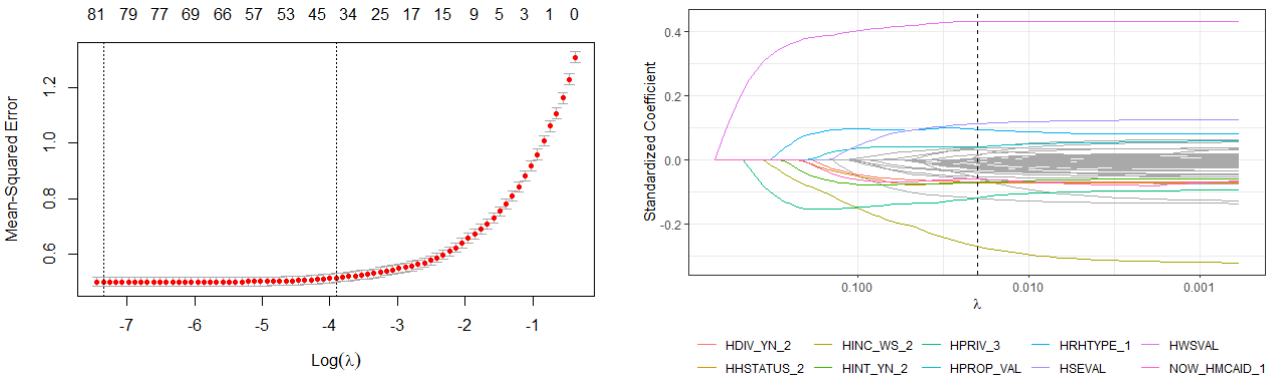


Figure 7. This figure shows the CV plot for the lasso regression, as well as a plot of the standardized coefficients for each feature at different lambda values.

Elastic Net Regression

The final regression trained for the regression-based methods was an elastic net regression. This is a combination of ridge and regression, as it does both coefficient shrinkage and feature selection but to a lower degree. The alpha value for an elastic net regression is not fixed, but rather varies in values between 0 and 1, not inclusive. Therefore, to train an elastic net regression on the train data, the first step was to compute a cross-validation on the alpha parameter to get the optimal value. The dashed line on Figure 8 corresponds to this alpha value: 0.216. Using this alpha, we then proceeded to train the model with a 10-fold cross-validation to select the lambda value that would reduce the CV error the most with the given alpha. This lambda value was 0.002977018. According to the one-standard error (1SE) rule, we chose 0.07725434 as our lambda value to make the model simpler. Instead of using 88 features, the lasso fit was only performed on 44 selected features. The left plot on Figure 8 represents the CV error at each lambda value: the left dashed line resembles the lambda value that minimizes the error, yet the right dashed line represents the lambda chosen by the 1SE rule. The values above represent the number of nonzero coefficients based on feature selection.

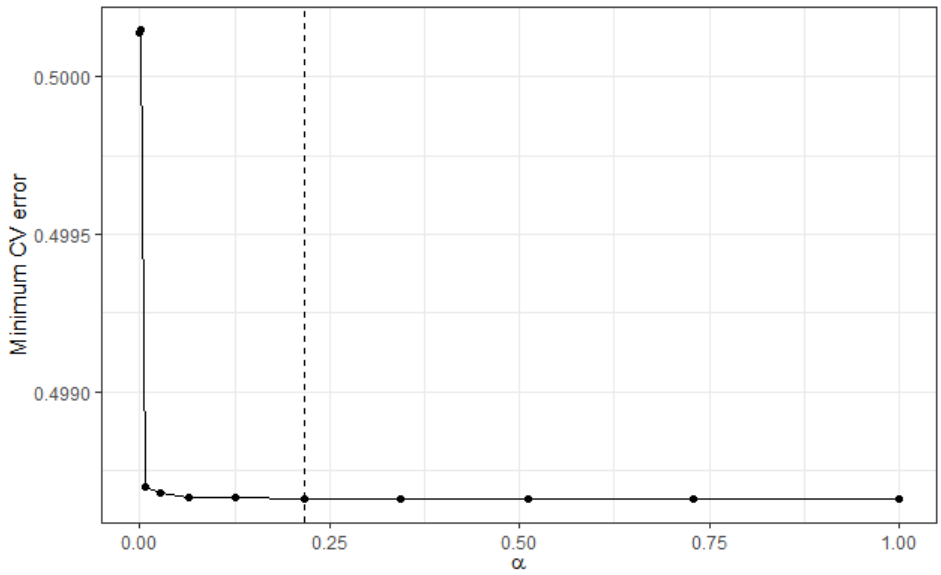


Figure 8. This plot highlights the CV error at each alpha value for the elastic net regression.

We computed the coefficients for the features selected by the elastic net using the 1SE rule. This can be visualized by the standardized coefficients falling on the black dashed line on the right plot on Figure 9. The highlighted curves represent the ten largest coefficients in magnitude with the chosen lambda: HDIV_YN_2, HINC_WS_2, HPRIV_3, HPUB_1, HWSVAL, HHSTATUS_2, HINT_YN_2, HPROP_VAL, HRHTYPE_1, and NOW_HMCAID_1. This model shares 9 out of the 10 features with the largest coefficients in magnitude with the lasso regression, and shares 8 out of 10 with the ridge regression.

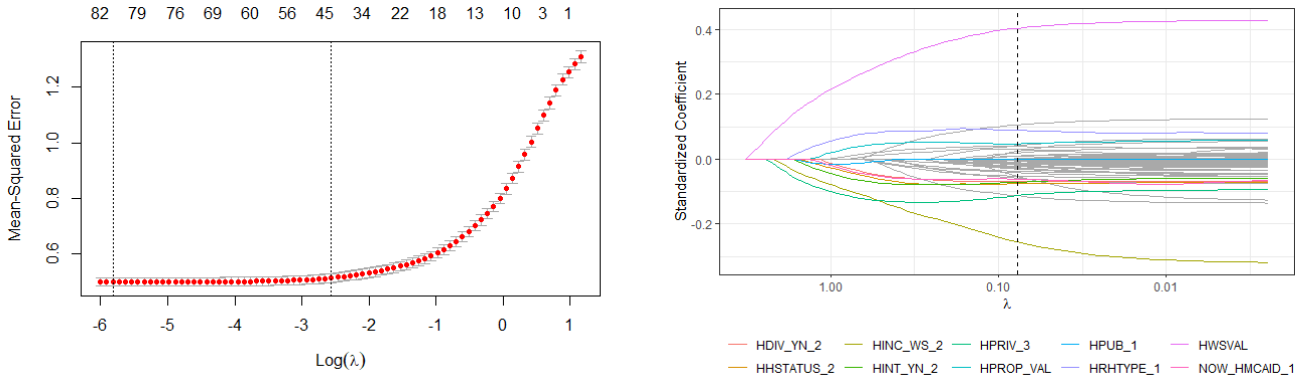


Figure 9. This figure shows the CV plot for the elastic net regression, as well as a plot of the standardized coefficients for each feature at different lambda values.

Tree-Based Methods

Regression Tree

In contrast to the regression-based methods described above, we fit various tree-based methods to the data to attempt to reduce the RMSE of the predictive model. When we fit a basic regression tree to our training data, the regression tree produced a tree with 12 splits and 13 terminal nodes. In order to tune the model, we produced a plot of the cross-validation error based on the number of terminal nodes in the tree, as seen in Figure 10. Using the one-standard error rule, we found that the optimal tree has 11 splits and 12 terminal nodes.

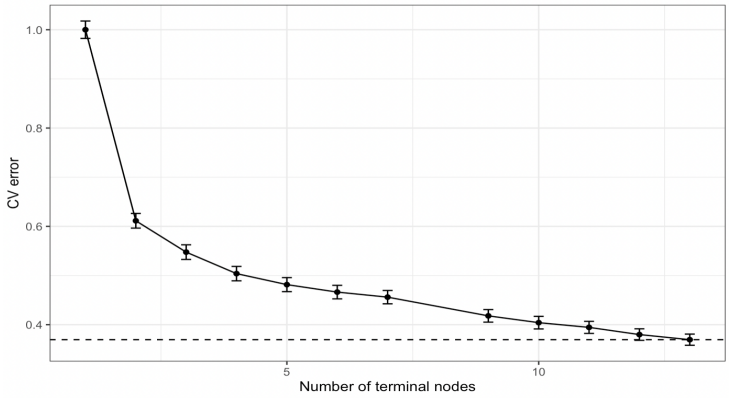


Figure 10: Plot of the CV error versus the number of terminal nodes.

Hence, as seen in Figure 11, our optimal tree is a slightly pruned version of the original tree. Upon further examination, we observed that the model determined self-employment income and whether the household received social security payments from the United States government to be especially predictive, as these were the variables that were most frequently used in the model to split the training data.

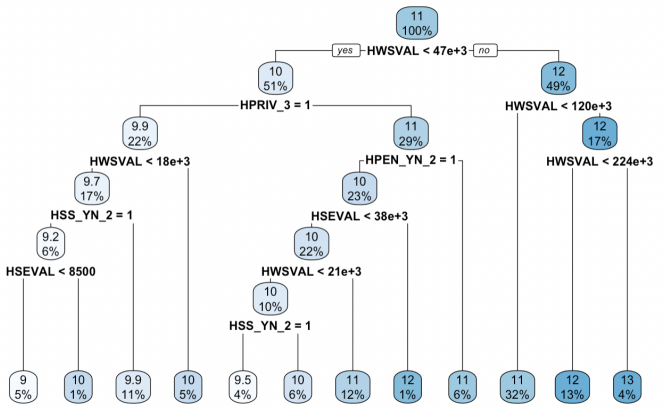


Figure 11: Optimal regression tree.

Random Forest

In order to reduce the variance of the decision tree, we then fit a random forest to our training data. A random forest involves taking bootstrapped samples and splitting based on a subset of features from the training data. This process decorrelates the trees, which improves the predictive power of the model. Due to the limitations of computer processing with such a large dataset, we subsampled 20% of our training data to use in the random forest model, which corresponded to approximately 10,000 observations.

After fitting the model, we produced Figure 12, which depicts the training error of the random forest versus the number of bootstrapped samples (labeled as trees on the graph). Given that the error appears to stabilize at approximately 200 trees, we chose $B = 200$ as the tuned value of bootstrapped samples to use in the training of our final random forest.

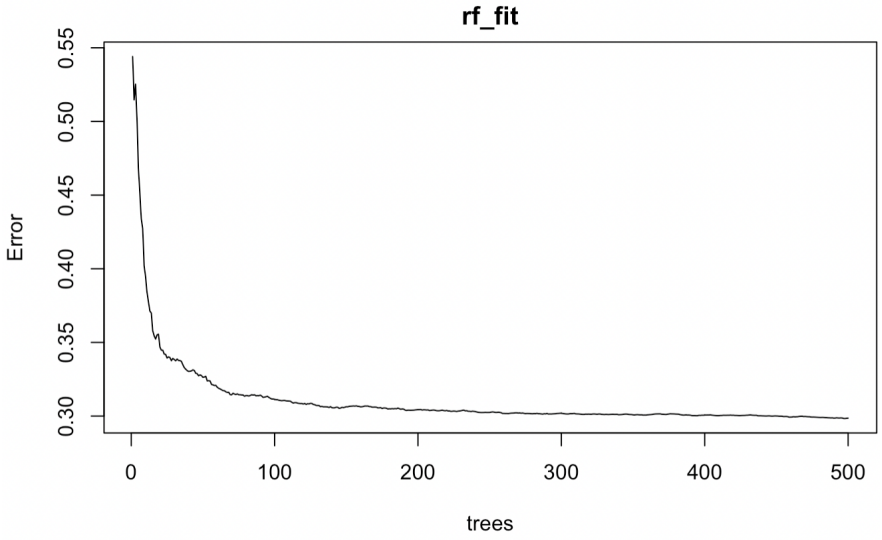


Figure 12: Training error of the random forest versus the number of trees.

We then attempted to tune the *m* parameter, which corresponds to the number of features selected to be sampled at each split in the tree. To this end, we trained random forests utilizing different values of *m* from 1 to 90. For each successive training model, we increased *m* by 11, which resulted in 9 trained random forests. The results of this model tuning are apparent in Figure 13, with $m = 56$ clearly minimizing the out-of-bag error. When comparing the random forest with $m = 56$ to the original random forest, we find that the tuned random forest has a reduced out-of-bag error, as evidenced in Table 1 below.

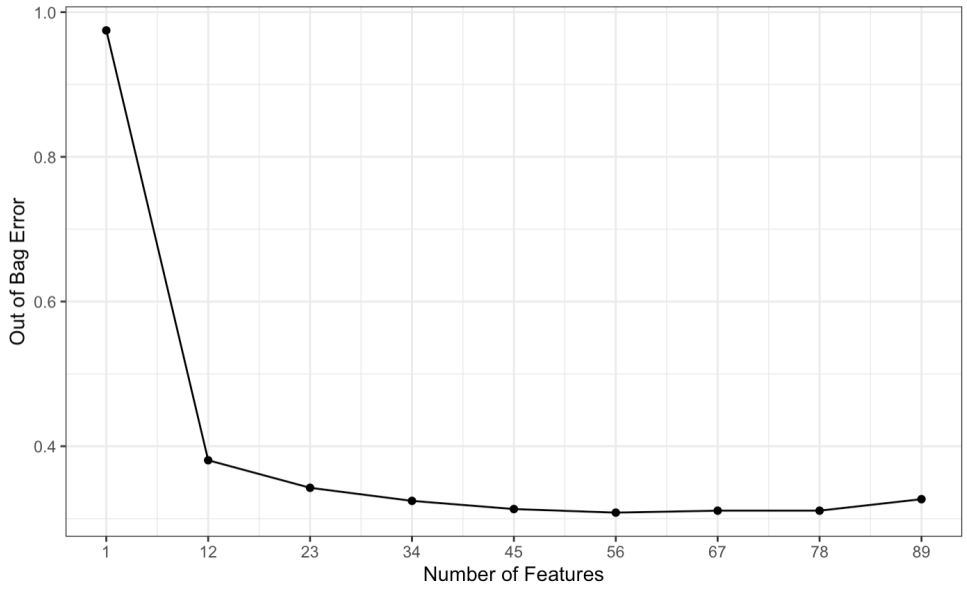


Figure 13: Out of bag error versus number of features when tuning the random forest.

Table 1: Error for default random forest and tuned random forest.

Default MSE ($m = 29$)	Out Of Bag Error ($m = 56$)
0.3016518	0.2920743

After training our final random forest using the *B* and *m* parameters chosen through the tuning process, we produced a variable importance plot, which depicts the variables contributing to the largest increases in node purity and the largest decreases in out-of-bag error. According to our variable importance plot, as seen in Figure 14, features concerning the receipt of social security payments, pension income, and self-employment income contribute the most to the percentage increase in MSE. On the other hand, income from wages and salaries, the receipt of private insurance coverage, and property value contribute to the greatest increases in node purity.

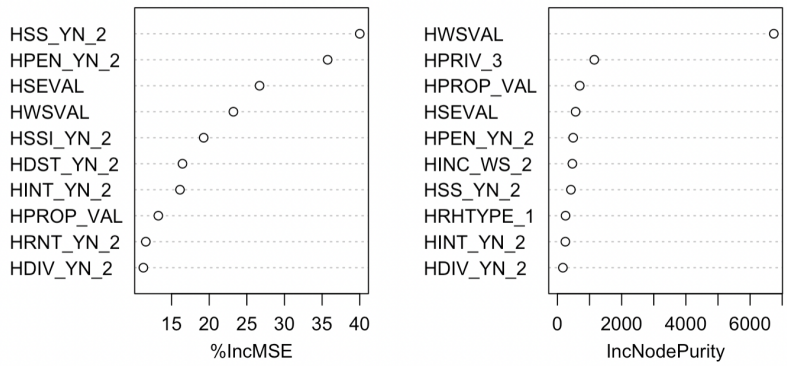


Figure 14: Variable importance plot for the tuned random forest.

These features are clearly predictive of household income, as these variables either are components of total household income (pension, wages, self-employment income) or are indicative of the relative income bracket the household is in (receipt of social security, property value, insurance coverage).

Boosting

Fitting based on a differing number of interaction depths, which refers to the maximum number of splits needed to get to a terminal node. To this end, we fit three boosted tree models with interactions depths of 1, 2, and 3. We kept the rest of the parameters constant across all three fitted models: shrinkage factor = 0.1, number of folds used for cross-validation = 5, and the number of trees = 1000. As observed in Figure 15 below, for each interaction depth, the cross-validation error monotonically decreases as the number of trees increases. Furthermore, we confirmed that the optimal number of trees is 2979, as that produced the minimum error value at an interaction depth equal to 3.

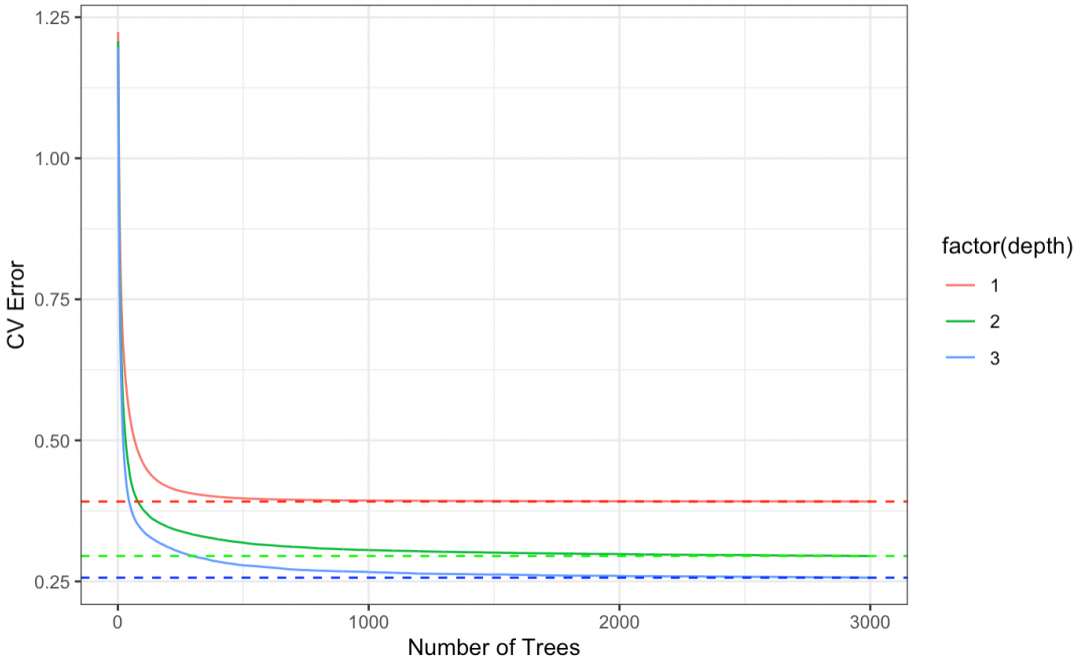


Figure 15: Plot of the CV error versus the number of trees at various interaction depths for the boosted model.

In order to assess variable importance in our model, we produced a relative influence table, and partial dependence plots. As per the relative influence table (Table 2), the top three features from the optimal boosted model were income from wages and salaries, self-employment income, and whether or not the household received social security payments from the government. Using these features, we created partial dependence plots, depicted in Figure 16 , to further examine the relationship between these features and the response within the boosted model.

Table 2: Relative influence table for the optimal boosted model.

Variable	Relative Influence
HWSVAL	63.144013
HSS_YN_2	6.053412
HSEVAL	5.641990
HPEN_YN_2	4.725551
HPROP_VAL	2.852001
HPRIV_3	2.128405
HRHTYPE_1	1.521752
HDIV_YN_2	1.160603

HDST_YN_2	1.157106
HINT_YN_2	1.145854

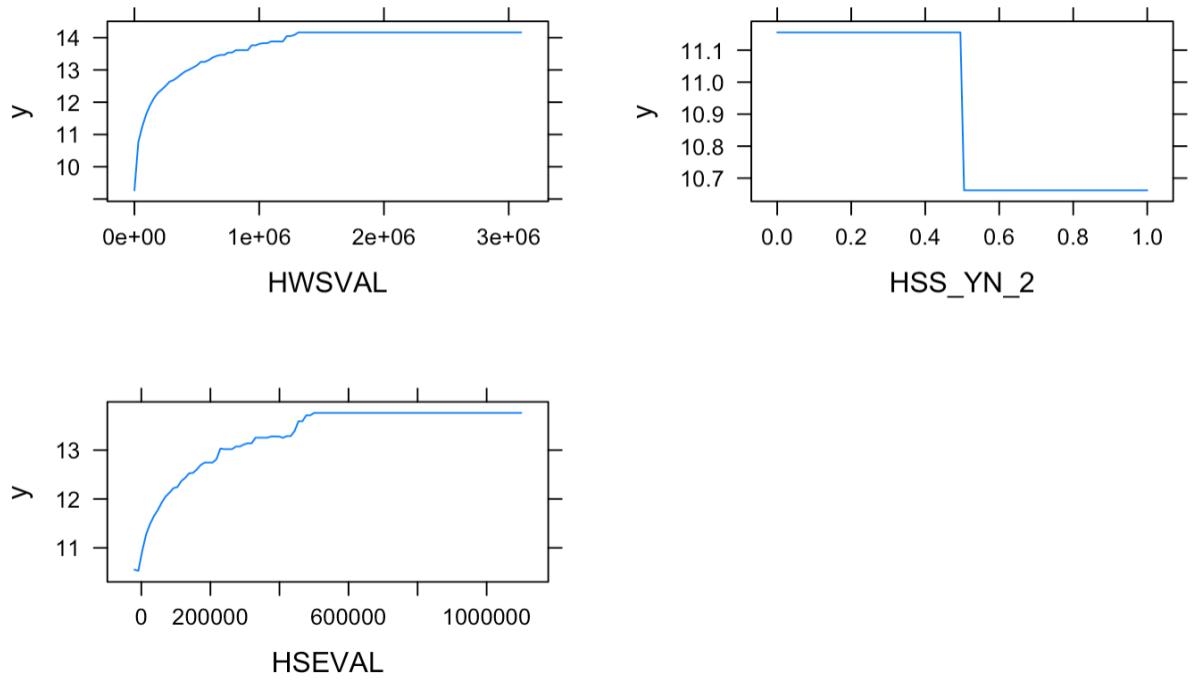


Figure 16: Partial dependence plots for selected features based on their relative importance.

We found that as wages and self-employment income increases, the model predicts total household income to increase as well, which is the relationship we would have predicted. This relationship appears to stabilize at a salary of $1.25 million and at $500,000 for self-employment income. Furthermore, as shown in Figure 16 above, a HSS_YN_2 value closer to 1, which infers that the household did not receive social security assistance, resulted in a lower value of total household income. Hence, we found that receiving social security payments from the government leads to the model predicting greater total household income.

Model Comparison

To compare all four regression-based models trained, as well as all three tree-based models, we tried predicting total household income for the test data using all models. Then, the root mean squared error (RMSE) was calculated for each set of predictions, yielding the following results:

Table 3: Regression-based RMSEs.

OLS RMSE	Ridge RMSE	Lasso RMSE	Elastic Net RMSE

0.6990139	0.713461	0.7108392	0.713461

Table 4: Tree-based RMSEs.

Regression Tree RMSE	Random Forest RMSE	Boosted Tree RMSE
0.6866733	0.5117325	0.4885832

Conclusion

Method Comparison

As seen in the tables above, all three tree-based methods were better at predicting household income compared to the regression-based methods. This could be due to the fact that a significant percentage of our features were categorical, which are better accounted for when fitting tree-based models. Out of the three tree-based methods, the boosted tree model performs the best, followed by the random forest and the regression tree. This is not surprising, as random forests and boosted tree models are algorithms designed to improve prediction performance by reducing the variance and the bias, respectively. Clearly, focusing on reducing the bias of the original regression tree through boosting was slightly better than reducing its variance through a random forest.

In terms of the regression-based methods, which did not perform as well as the tree-based methods, the ordinary least squares regression performed the best at predicting the total household income for the test data, followed by the lasso, and then the elastic net and ridge regressions. These findings are interesting, given that penalized regressions were not able to outperform the ordinary least squares model. A possible explanation for this is that our penalized methods over-regularized too much to the extent of being outperformed by the OLS model.

However, despite the differences in test RMSE for the regression-based methods and the tree-based methods, similar features were determined by each model to hold significant predictive power. For example, the Ridge regression, Lasso regression, random forest, and boosted tree model all determined wage and salary income (HWSVAL), self-employment income (HSEVAL), property value (HPROPVAL), and private insurance coverage (HPRIV_3) to be some of the most significant features when attempting to predict total household income.

Takeaways

As selected by our models, one of the most important variables was whether or not the household receives social security payments. Before conducting our analyses, we hypothesized that the presence of social security payments would lead to a lower total household income. Households that receive social security payments would typically have older residents, who are either retired or are working less than younger people. Hence, while they would receive social security, their income generated from other sources would undoubtedly be lower, leading to the model predicting lower total household income overall. However, our hypothesis was contradicted by our regression tree model, as well as the boosted tree model. Given this result, we concluded that the presence of social security payments must indicate that at least one member of the household previously had a decently well paying job if he/she was able to pay into the social security program for at least 10 years.

Given the relatively few location-based variables in our dataset, it is interesting that property value was deemed significant in almost all of the seven models. While economists have traditionally been focused on people-based policies such as welfare payments to relieve people from poverty, it is clear that more effort must be devoted to transforming impoverished areas in order to lift residents from the cycle of poverty. Many economists raise concerns that implementing place-based policies will lead to gentrification and ultimately displace poorer people from these neighborhoods. However, efforts to improve local schools and reduce crime in certain areas can undoubtedly increase property values within the area, leading to increased total household income and lower levels of poverty.

The importance of insurance is one final noteworthy takeaway. The majority of our models, especially the regression models, deemed private insurance coverage to be an important indicator of total household income. This is significant due to the nature of the health care system in the United States where private insurance coverage results in significantly better medical benefits and access to more advanced medical care. People who are in poverty are more likely to be in poorer health either due to lack of access to good medical care or general malnourishment. While it is not surprising that private insurance coverage is reserved for higher income individuals, it is concerning that it is one of the most predictive features, as access to quality health care should not be entirely reliant on a person's salary.

Limitations

One important limitation that we faced was that our data was mostly composed of categorical variables. This made our data cleaning process significantly more challenging, as we had to individually one-hot encode many of our variables. We had to remove many of the continuous features early on in our data cleaning process, as many of the continuous features were linear combinations of total household income. We believe that this hampered prediction capabilities, as it is difficult for categorical variables to predict the specific impact on total household income. For example, as mentioned above, the presence of social security payments was a significant variable in many of our models. However, this does not tell us how much social security income each household received.

One other important limitation was the lack of features that pertained to individual members of the household. In the context of household income, it is important to remember that household income is the sum of all individuals within the household. Members of the household may have drastically different incomes based on their education, their expertise, their tenure in the labor force, etc. We believe the lack of data on an individualized-level led to the exclusion of important variables with respect to the predictive capabilities of our model.

Follow-Ups

Our first suggestion for next steps would be to try out with a different model, such as a neural network. Although we were able to achieve a low RMSE with our tree-based methods, we believe a neural network might be even more accurate in predicting total household income based on weights on different features. Second, as explained above, the inclusion of data pertaining to individuals within the household could help improve the prediction power of the model. We believe the inclusion of features such as race (majority within the household) or education level (median) would be influential if we were to incorporate them into our existing models. Finally, as we began the paper focusing on poverty, we would like to see further analysis regarding the poverty thresholds based on predicted household income. Given that the poverty thresholds vary based on the number of people living in the household, it would be interesting to explore the predicted level of poverty in the United States based on predicted total household income for each number of household occupants.

Appendix

Explanatory Variables

Provided below are the explanatory variables that were included in our model, as well as a description of each. All of the categorical variables were encoded into our model as factors with dummy variables.

Continuous Variables

- Child Care (HCHCARE_VAL): Annual amount paid for child care by household members.
- Energy Assistance (HENGVAL): Amount of energy assistance received in 2018.
- Food Stamps (HFDVAL): Value of all food stamps received in 2017.
- Farm Income (HFRVAL): Value of income received through the operation of a farm.
- Hot Lunch (HHOTNO): Number of children in the household who ate hot lunch
- Families (HNUMFAM): Number of families within a household.
- Property Value (HPROP_VAL): Estimate of current property value (equal to 0 for renters).
- Nutrition (HRNUMWIC): Number of people in the household receiving WIC (Special Supplemental Nutrition Program for Women, Infants, and Children).
- Self-Employment (HSEVAL): Value of self-employment income
- Units (HUNITS): Number of units in the structure
- Wages (HWSVAL): Value of all wages in the household

Categorical Variables

- Topcode Flag (THCHCHARE_VAL): Whether or not the value of child care paid by a household was topcoded.
    - 0 = Non topcoded
    - 1 = Topcoded
- Data Swapping Flag (THPROP_VAL): Whether or not data swapping occurred for the property value feature.
    - 0 = No swapping
    - 1 = Variable value was swapped with another record
- Area of Residence (GEDIV): Nine-level categorical variable for census division of current residence.
    - 1 = New England
    - 2 = Middle Atlantic
    - 3 = East North Central

    - 4 = West North Central
    - 5 = South Atlantic
    - 6 = East South Central
    - 7 = West South Central
    - 8 = Mountain
    - 9 = Pacific
- Household Type (HRHTYPE): Ten-level categorical variable for classification of persons within the household.
    - 1 = Married couple primary family (neither spouse in Armed Forces)
    - 2 = Married couple primary family (one spouse in Armed Forces)
    - 3 = Unmarried civilian male primary family householder
    - 4 = Unmarried civilian female primary family householder
    - 5 = Primary family household - reference person in Armed Forces and unmarried
    - 6 = Civilian male nonfamily householder
    - 7 = Civilian female nonfamily householder
    - 8 = Nonfamily householder household - reference person in Armed Forces
    - 9 = Group quarters with actual families
    - 10 = Group quarters with secondary individuals only
- No Cash Rent (H_TENURE_3): Whether or not a household pays no cash rent.
    - 0 = Household pays a cash rent or owns the household
    - 1 = Household pays no cash rent
- Living Quarters (H_LIVQRT): Twelve-level categorical for type of living quarters.
    - 1 = House, apt., flat
    - 2 = HU in nontransient hotel, etc.
    - 3 = HU, perm, in trans. hotel, motel, etc.
    - 4 = HU in rooming house
    - 5 = Mobile home or trailer with no permanent room added
    - 6 = Mobile home or trailer with 1 or more perm rooms added
    - 7 = HU not specified above Other Unit
    - 8 = Qtrs not hu in rooming or boarding house
    - 9 = Unit not perm in trans. hotel, motel, etc.
    - 10 = Tent or trailer site
    - 11 = Student quarter
    - 12 = Other not HU
- Telephone in Household (H_TELHHD_2): Whether or not a household owns a telephone.
    - 0 = Household owns a telephone
    - 1 = Household does not own a telephone
- Annuity Income (HANN_YN_2): Whether or not any member of the household received income from an annuity.
    - 0 = Annuity income received

    - 1 = Annuity income not received
- Child Care Payments (HCHCARE_YN): Whether or not any member of the household paid for child care for the child of anyone in the household.
- Health Insurance Coverage (HCOV): Three-level categorical for status of current health insurance coverage in the household.
    - 1 = All members of the household
    - 2 = Some members of the household
    - 3 = No members of the household
- Child Support Payments (HCSP_YN_2) : Whether or not any member of the household received any child support payments.
    - 0 = Child support payment received
    - 1 = No child support payment received
- Disability (HDIS_YN_2): Whether or not any member of the household had a disability or health problem that impeded their ability to work.
    - 0 = Disability
    - 1 = No disability
- Shares of Stock (HDIV_YN_2): Whether or not any member of the household owned any shares of stock in corporations or in mutual funds.
    - 0 = Owned shares of stock
    - 1 = Did not own shares of stock
- Retirement Income (HDST_YN_2): Whether or not any member of the household received retirement distribution income for people aged 58 and over.
    - 0 = Received retirement distribution income
    - 1 = Did not receive retirement distribution income
- Educational Assistance (HED_YN_2): Whether or not any member of the household received any educational assistance for tuition, fees, books, or living expenses in 2018.
    - 0 = Received educational assistance
    - 1 = Did not receive educational assistance
- Regular Financial Assistance (HFIN_YN_2): Whether or not any member of the household received any financial assistance from friends or relatives not living in the household.
    - 0 = Received financial assistance
    - 1 = Did not receive financial assistance
- Free or Reduced Price Lunch (HFLUNCH_2): How many children in the household received free or reduced price lunch.
    - 0 = All or some of the children in the household
    - 1 = None of the children in the household
- Hot Lunch (HHOTLUN_2): How many children in the household usually ate a complete hot lunch offered at school.
    - 0 = All or some of the children in the household

    - 1 = None of the children in the household
- Household Status (HHSTATUS): Three-level categorical for the household status.
    - 1 = Primary Family
    - 2 = Nonfamily householder living alone
    - 3 = Nonfamily householder living with nonrelatives
- Farm Self-Employment (HINC_FR_2): Whether or not any member of the household is self-employed in a farming-related job.
    - 0 = Self-employed in farming
    - 1 = Not self-employed in farming
- Business Self-Employment (HINC_SE_2): Whether or not any member of the household is self-employed in their own business.
    - 0 = Self-employed in own business
    - 1 = Not self-employed in own business
- Unemployment Compensation (HINC_UC_2): Whether or not any member of the household received unemployment compensation
    - 0 = Received unemployment compensation
    - 1 = Did not receive unemployment compensation
- Workers Compensation (HINC_WC_2): Whether or not any member of the household received workers compensation.
    - 0 = Received workers compensation
    - 1 = Did not receive workers compensation
- Wage and Salary (HINC_WS_2): Whether or not any member of the household received a wage or salary.
    - 0 = Received a wage or salary
    - 1 = Did not receive a wage or salary
- Saved Money - Accounts/Funds (HINT_YN_2): Whether or not any member of the household had money saved in a savings account, checking account, money market fund, certificate of deposit, savings bond, any other (non-retirement) investment which pays interest, or retirement account.
    - 0 = Saved money in accounts or funds
    - 1 = No saved money in accounts or funds
- Subsidized Rent (HLORENT_2): Whether or not the household is paying lower rent because the government is paying part of the cost.
    - 0 = Paying lower rent
    - 1 = Not paying lower rent
- Medicaid (HMCAID_2): Whether or not the household has Medicaid, PCHIP, or any other means-tested coverage last year.
    - 0 = All or no members of the household
    - 1 = Some members of the household

- Income from Other Sources (HOI_YN_2): Whether or not the household received any income from foster child care, alimony, jury duty, armed forces reserved, severance pay, hobbies, or any other source.
    - 0 = Income from above sources
    - 1 = No income from above sources
- Public Assistance/Welfare Payments (HPAW_YN_2): Whether or not any member of the household received public assistance or welfare payments from the state or local welfare office.
    - 0 = Received public assistance/welfare payments
    - 1 = Did not receive public assistance/welfare payments
- Pension Income (HPEN_YN_2): Whether or not any member of the household received pension income from a previous employer or union.
    - 0 = Received pension income
    - 1 = Did not receive pension income
- Home Mortgage (HPRES_MORT): Presence of a home mortgage
- Private Coverage (HPRIV_3): Whether or not any member of the household had private insurance coverage.
    - 0 = All or some members of the household had private coverage
    - 1 = No members of the household had private coverage
- Government coverage (HPUB): Three-level categorical for if any member of the household had government insurance coverage.
    - 1 = All members of the household
    - 2 = Some members of the household
    - 3 = No members of the household
- Public Housing Project (HPUBLIC_1): Whether or not the household is part of a public housing project that is owned by a local housing authority or other public agency.
    - 0 = Not part of a public housing project
    - 1 = Part of a public housing project
- Rental or Estate Income (HRNT_YN_2): Whether or not any member of the household owned any land or property that was rented to others and/or received income from royalties, estates, or trusts.
    - 0 = Received rental or estate income
    - 1 = Did not receive rental or estate income
- WIC Program (HRWICYN_2): Whether or not any member of the household was enrolled in the Woman, Infants, and Children Nutrition Program.
    - 0 = Enrolled in WIC
    - 1 = Not enrolled in WIC
- Supplemental Security Income (HSSI_YN_2): Whether or not any member of the household received supplemental security income payments.
    - 0 = Received SSI payments

    - 1 = Did not receive SSI payments
- Social Security Payments (HSS_YN_2): Whether or not any member of the household received social security payments from the U.S. government.
    - 0 = Received social security payments
    - 1 = Did not receive social security payments
- Survivor Income (HSUR_YN_2): Whether or not any member of the household received survivor income such as widow's pensions, estates, trusts, or annuities.
    - 0 = Received survivor income
    - 1 = Did not receive survivor income
- Veterans Payments (HVET_YN_2): Whether or not any member of the household received payments from the veterans administration.
    - 0 = Received veterans payments
    - 1 = Did not receive veterans payments
- Current Medicaid (NOW_HMCAID): Three-level categorical for if any member of the household has Medicaid, PCHIP, or any other means-tested coverage.
    - 1 = All members of the household
    - 2 = Some members of the household
    - 3 = No members of the household
- Current Private Coverage (NOW_HPRIV_2): Whether or not any member of the household has private insurance coverage.
    - 0 = All or no members of the household
    - 1 = Some members of the household
- Metropolitan Area (GTMETSTA): Three-level categorical for whether the household is in a metropolitan area.
    - 1 = Metropolitan area
    - 2 = Non-metropolitan area
    - 3 = Not identified