Predicting Median Household Income

Arnold Jia and Kevin Chen

December 19, 2021

https://github.com/kevc528/county-analysis

1 Executive Summary

Problem: Median household income is a canonical measure of well-being in the United States. Not only is it indicative of one's economic status, it is closely tied to other demographic and sociopolitical factors, from educational attainment to health. As shown in a NCBI study, median household income was associated with longer life expectancy throughout the income distribution. The gap in longevity between the wealthiest 1% and the poorest 1% of individuals was 14.6 years for men and 10.1 years for women. Furthermore, this study also found that inequality in life expectancy has only increased throughout the years. Between 2001 and 2014, life expectancy for the top 5% increased by 2.34 years for men and 2.91 years for women; but in the bottom 5%, life expectancy only increased by 0.32 years for men and 0.04 years for women. This means that median household income is closely, and statistically significantly, tied to health, and is growing evermore important. Beyond health, median household income is also closely tied to measures of happiness and other standards of living, making it a valuable measure to explore in-depth.

Data: The data we use includes information about median household income, housing, education, ethnicity, etc. for 3,142 counties in the United States. See **Data Description** below for an in-depth description of all of the features. This data was collected from the American Community Survey in 2019, which was conducted by the U.S. Census Bureau to better understand communities across the country.

Analysis: After cleaning the data and splitting it into training and testing sets, our first step was to explore various relationships in the data to obtain a preliminary sense of which features are important and how they impact household income. Then, we fit four regression methods (ordinary-least-squares regression, ridge regression, lasso regression, and elastic net regresion) and three tree-based methods (regression tree, random forest, and boosted model). With each method, we tune parameters when necessary to best minimize expected test error. In the end, we test all models on the test dataset.

Conclusions: The boosted model had the best performance, with an RMSE of $4,922. Ultimately, our models were effective in explaining the variance in median household income, and there was significant overlap in which features were selected as important. The main sets of features that were important in predicting median household income across all models were (1) technology ownership (computers, smartphones, broadband internet), (2) economic factors (unemployment, lack of insurance), and (3) education (having a bachelor's degree or being a high-school graduate). For relevant stakeholders, we believe that analysis of these aforementioned features will be insightful into improving well-being across the country, especially in struggling communities. Ultimately, investing in access to technology, infrastructure, and education is tremendously important in improving median household income.

Contents

2 Introduction

2.1 Background Information

Household income has long been a measure of the well-being of members in a community. By definition, household income is the pre-tax earnings of every member in a house over age 15. Among the measures for household income, one of the most popular is the median household income of a region, which removes the effect of outliers compared to the mean household income. We are specifically interested in the median household income of different counties in the United States.

Median household income has been used heavily to determine standards-of-living, which is a very important factor for people when deciding where to live. Additionally based on a research study in January 2021, higher income in the United States has the potential to increase people's day-to-day well-being. With our data analysis, we are looking to explore what other factors relate to median household income. From this, we will glean insights into the characteristics that make a community attractive to residents and features that may have an influence on household income, standards of living, and more importantly, measures of happiness.

2.2 Analysis Goals

While there are intuitive factors that relate to an individual's income, such as educational attainment and employment status, we seek to understand which factors are most important in predicting median household income of an entire county. We will be using features related to housing, education, ethnicity, age distributions, etc. provided by the American Community Survey to predict median household income of a county. It will be interesting to see which features our models use to make predictions.

For prediction, we will be running various regression models and tree-based models and success will be evaluated by the root mean square error (RMSE) on our test data set. Perhaps more important than prediction performance, we seek to understand to what extent the features chosen by the models are interpretable.

2.3 Significance

We hope that our analysis will help people understand the factors that differentiate counties with a high median household income from counties with a low median household income. This may help officials better understand the factors that lead to differences in median household income and the disparities that arise from this difference. Additionally, our analysis should help people better understand the communities they live in and ways to improve them. Our analysis is on a wide range of features; however we hope that this report can lead people to perform their own research on more specific features and expand our understanding of this subject.

3 Data

3.1 Data Sources

The data we collected was from OpenIntro and can be found here. This data was originally collected from the American Community Survey in 2019, which is conducted by the U.S. Census Bureau. This is a demographics survey to help local officials and leaders understand their community. Included is information about housing, education, ethnicity, etc. for 3,142 counties in the United States. Data is released every year and this data set is from the 2019 survey. We believe that the 2019 survey would be best since this was conducted pre-COVID and would be more representative of both the past and future.

3.2 Data Cleaning

The first step in the data cleaning process was removing the margin of error columns in our data set. The original data set had 46 features that were the margin of error for continuous features and these would not be helpful for our models. Additionally, these margin of error columns contained many null values.

After this, we found which of the remaining columns had the most null values. We discovered that only 4 columns had null values: `poverty`, `poverty_65_and_over`, `poverty_under_18`, and `mean_work_travel`. However, if we were to remove all observations with null values, we would lose over half of our dataset. On the other hand, we did not want to drop these features, as we believed that they would be important predictors of median household income. To combat this problem, we decided to impute these missing values. For each observation containing null values, we replaced them with the average values for their respective state. We decided to use state instead of country-wide averages because the entire United States is very heterogeneous, so we thought state-wide values would result in more precise estimates for each county. By doing this, we were able to retain all observations in the data.

3.3 Data Description

Our cleaned data has 3,142 observations and 49 features. Each observation represents a single county. The response variable we chose was `median_household_income` which is the median household income of a county. This response variable is continuous. Below is a description of the 49 features based on the source of the data.

Categorical Variables

- `state`: State that the county is in
- `name`: Name of the county
- `fips`: FIPS code for the county

Continuous Variables

- `age_over_18`: Percent of population 18 and over (2015-2019)
- `age_over_65`: Percent of population 65 and over (2015-2019)
- `age_over_85`: Percent of population 85 and over (2015-2019)

- `age_under_5`: Percent of population under 5 (2015-2019)
- `asian`: Percent of population that is Asian alone (2015-2019)
- `avg_family_size`: Average family size (2015-2019)
- `bachelors`: Percent of population 25 and older that earned a Bachelor's degree or higher (2015-2019)
- `black`: Percent of population that is black alone (2015-2019)
- `hispanic`: Percent of population that identifies as Hispanic or Latino (2015-2019)
- `household_has_broadband`: Percent of households that have broadband internet subscription (2015-2019)
- `household_has_computer`: Percent of households that have desktop or laptop computer (2015-2019)
- `household_has_smartphone`: Percent of households that have smartphone (2015-2019)
- `households`: Total households (2015-2019)
- `households_speak_asian_or_pac_isl`: Percent of households speaking Asian and Pacific Island language (2015-2019)
- `households_speak_limited_english`: Percent of limited English-speaking households (2015-2019)
- `households_speak_other`: Percent of households speaking non European or Asian/Pacific Island language (2015-2019)
- `households_speak_other_indo_euro_lang`: Percent of households speaking other Indo-European language (2015-2019)
- `households_speak_spanish`: Percent of households speaking Spanish (2015-2019)
- `housing_mobile_homes`: Percent of housing units in mobile homes and other types of units (2015-2019)
- `housing_one_unit_structures`: Percent of housing units in 1-unit structures (2015-2019)
- `housing_two_unit_structures`: Percent of housing units in multi-unit structures (2015-2019)
- `hs_grad`: Percent of population 25 and older that is a high school graduate (2015-2019)
- `mean_household_income`: Mean household income (2019 dollars, 2015-2019)
- `mean_work_travel`: Mean travel time to work (2015-2019)
- `median_age`: Median age (2015-2019)
- `median_household_income`: Median household income (2015-2019)
- `median_individual_income`: Median individual income (2019)
- `median_individual_income_age_25plus`: Median individual income (2019 dollars, 2015-2019)
- `native`: Percent of population that is Native American alone (2015-2019)
- `other_single_race`: Percent of population that is some other race alone (2015-2019)
- `pac_isl`: Percent of population that is Native Hawaiian or other Pacific Islander alone (2015-2019)
- `per_capita_income`: Per capita money income in past 12 months (2019 dollars, 2015-2019)
- `persons_per_household`: Persons per household (2015-2019)

- `pop`: 2019 population
- `poverty`: Percent of population below the poverty level (2015-2019)
- `poverty_65_and_over`: Percent of population 65 and over below the poverty level (2015-2019)
- `poverty_under_18`: Percent of population under 18 below the poverty level (2015-2019)
- `two_plus_races`: Percent of population that is two or more races (2015-2019)
- `unemployment_rate`: Unemployment rate among those ages 20-64 (2015-2019)
- `uninsured`: Percent of civilian noninstitutionalized population that is uninsured (2015-2019)
- `uninsured_65_and_older`: Percent of population 65 and older that is uninsured (2015-2019)
- `uninsured_under_19`: Percent of population under 19 that is uninsured (2015-2019)
- `uninsured_under_6`: Percent of population under 6 years that is uninsured (2015-2019)
- `veterans`: Percent among civilian population 18 and over that are veterans (2015-2019)
- `white`: Percent of population that is white alone (2015-2019)
- `white_not_hispanic`: Percent of population that is white alone, not Hispanic or Latino (2015-2019)

3.4 Data Allocation

For data allocation, we performed a 80-20 split, where 80% of the observations were used for training and exploration and 20% were used for testing. We made sure to use a seed so our training and testing data stayed consistent throughout different runs. Additionally, we removed highly correlated income features and only kept `median_household_income`. More specifically, we removed `mean_household_income`, `median_individual_income`, `median_individual_income_age_25plus`, and `per_capita_income`. This is because we feared keeping these features would cause our models to only look at them and not other, more interesting features. Note that the `state, name`, and `fips` features are kept for exploration, but removed for training, because we want to model median household income based on the county's meaningful features and not its specific location or name and fips code.

3.5 Data Exploration

Below, we will perform data exploration on the training data.

3.5.1 Response Variable

As mentioned previously, the response variable is the median household income per county in the United States. First, we look to explore the distribution of the response variable among the counties in the United States for the training data.

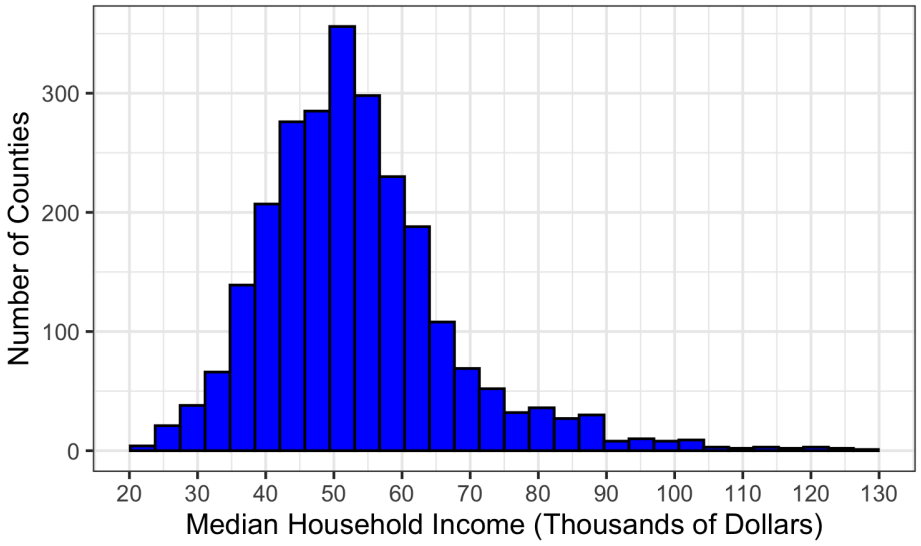

Figure 1: Histogram of Median Household Income

As we can see in Figure 1, which shows a histogram of median household income among U.S. counties, median household income seems to follow a normal distribution, where most of the counties lie at a median household income of around $40,000 to $65,000. However, there are a few counties that have extremely high median household incomes, as signified by the long right tail.

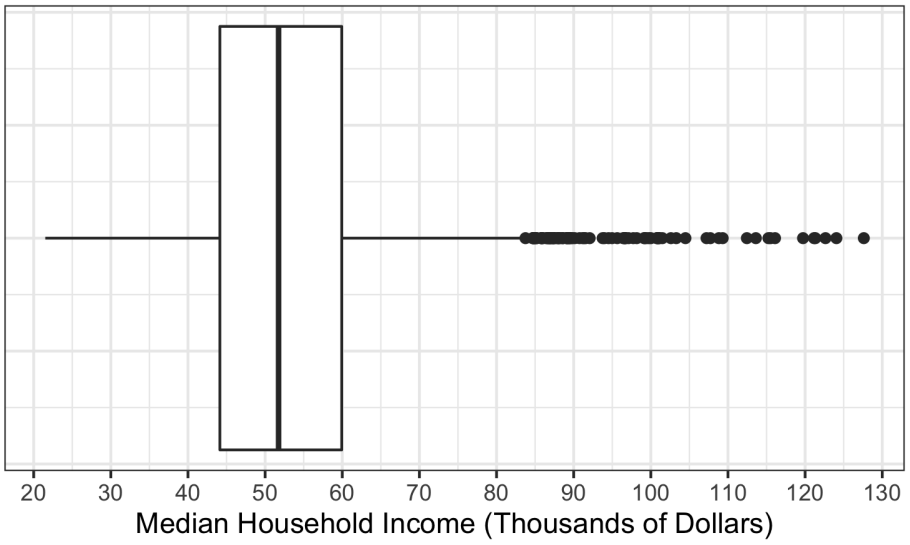

Figure 2: Boxplot of Median Household Income

This is confirmed when we look at the boxplot in Figure 2. We notice that there are many outliers on the right-hand side, meaning many counties have extremely high median household incomes. To attain more specific numbers, we can take a closer look at the quartiles in the distribution of median household income.

Percentile	Median Household Income
0	21504
25	44135
50	51745
75	59931
100	127610

Table 1: Quartiles for Median Household Income

In Table 1, each of the quartiles are shown. In the training dataset, the median value for median household income in all counties in the United States is $51,745, and ranges from a minimum of $21,504 to a maximum of $127,610.

Now, we display the median household income on a map of the United States. Doing this should allow us to get a better understanding of how median household income is distributed throughout the country.

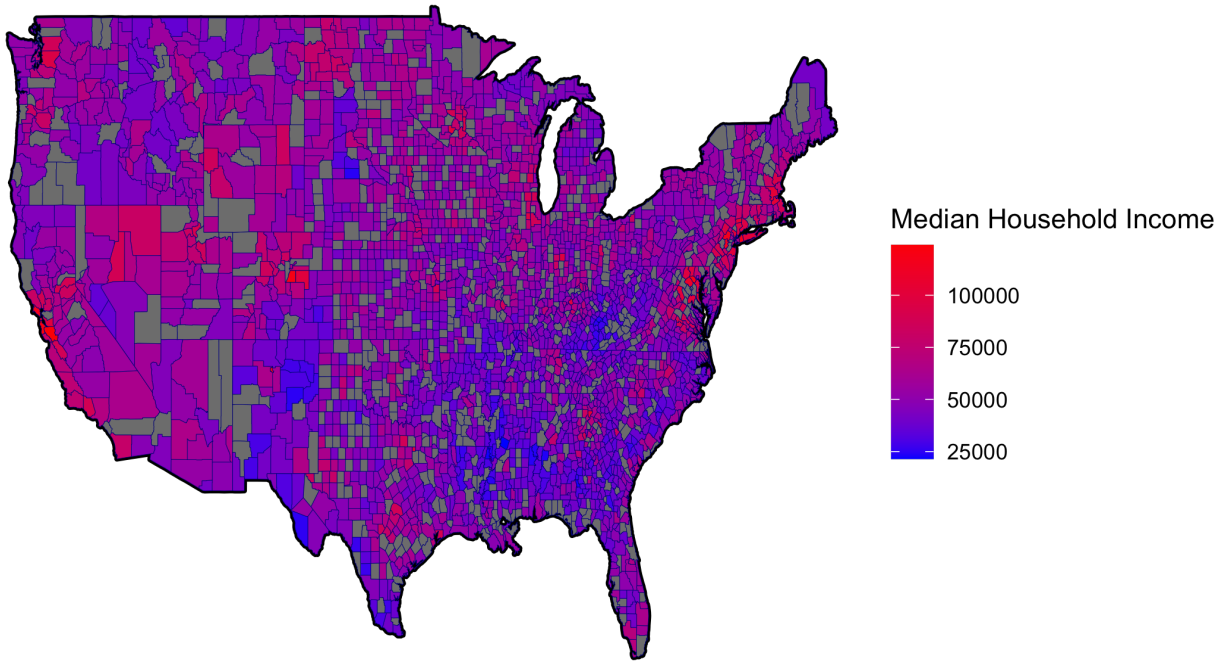

Figure 3: Heatmap of United States for Median Household Income

In Figure 3, we see a heatmap of the United States, where blue indicates counties with a lower median household income and red indicates counties with a higher median household income. Note the grey values are counties without data and are likely the ones being used in the test dataset. In this map, we notice that there is a large cluster of counties with low median household incomes in the south, especially around Mississippi, Alabama, and Louisiana. On the other hand, counties with higher median household incomes can be found near the Northeast Coast and the Californian Coast. This is because higher paying jobs tend to be here (i.e. New York and California). Likewise, these regions have higher costs of living, which then

attract wealthier households.

Next, we take a snapshot of which counties have the highest and lowest median household incomes.

Name	State	Median Household Income
Falls Church city	Virginia	127610
Santa Clara County	California	124055
San Mateo County	California	122641
Los Alamos County	New Mexico	121324
Howard County	Maryland	121160
Douglas County	Colorado	119730
Nassau County	New York	116100
Morris County	New Jersey	115527
Marin County	California	115246
Somerset County	New Jersey	113611

Table 2: Top 10 Counties with Highest Median Household Income

Name	State	Median Household Income
Holmes County	Mississippi	21504
Clay County	Georgia	22325
East Carroll Parish	Louisiana	22346
Perry County	Alabama	23447
Greene County	Alabama	24145
Issaquena County	Mississippi	24208
Sumter County	Alabama	24320
Todd County	South Dakota	24331
Wolfe County	Kentucky	24623
Guadalupe County	New Mexico	24798

Table 3: Top 10 Counties with Lowest Median Household Income

In Table 2 we see the top 10 counties in our training data set with the highest median household incomes. Here, we see that many of the counties are in the Bay Area (Santa Clara, San Mateo, Marin) or in the New York Metro Area (Nassau, Morris, Somerset). The remaining counties have relatively small populations and size, with relatively low poverty rates. Because of this, these counties likely have more wealthy individuals and fewer poor individuals compared to larger counties with higher poverty rates.

In Table 3 we see the top 10 counties in our training data set with the lowest median household incomes. Based on this, we see that most of the counties in this table are located in the deep south: in states like Mississippi, Georgia, Louisiana, and Alabama. The other counties that aren't in the south are in exceptionally rural areas of the United States. This makes sense because in the south and other rural areas, there are often fewer high school and college graduates, and fewer skilled jobs. Most people in high earning positions choose to leave the south/rural areas to work in bigger cities in regions named in the previous paragraph.

3.5.2 Explanatory Variables

Next, we will look into some of our explanatory variables. First, we obtain a broad understanding of how explanatory variables correlate with one another and the response variable.

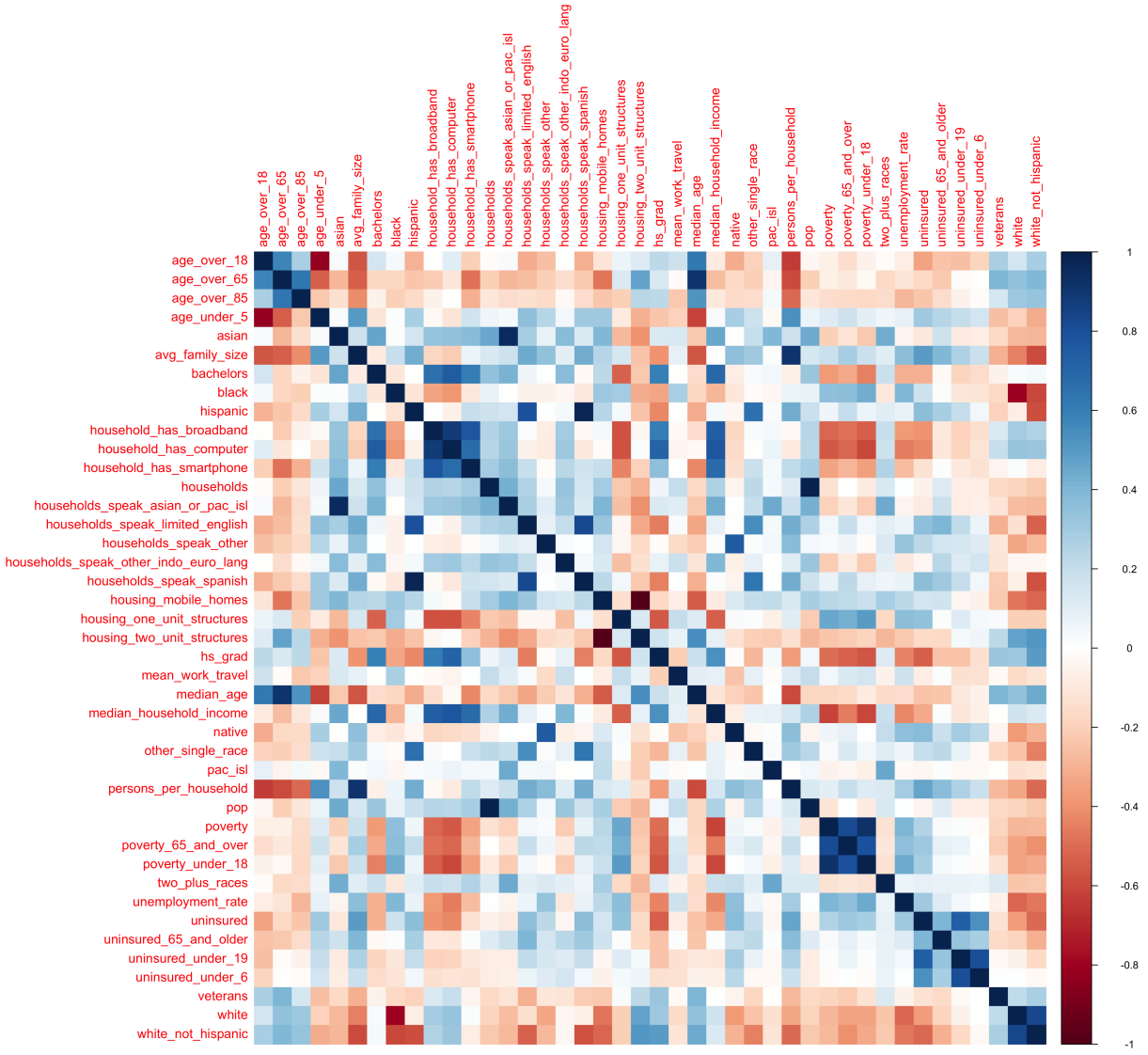

Figure 4: Correlation Plot of All Features

Above, in Figure 4, we see a correlation plot of all the features and the response variable. Among the features, we see there are some features that are extremely correlated. One example of this is the features `household_has_computer`, `household_has_broadband`, and `household_has_smartphone`, which have a strong positive correlation. This makes sense because households that can afford a computer, can likely also afford broadband and smartphones. Another example are the features `hs_grad` and `poverty`, which have a

strong negative correlation. This also makes sense because it is likely that people who have graduated high school can find a job more easily and earn a higher income.

Now, let's see the features that have the strongest positive correlation with our response variable, median household income.

Feature	Correlation
household_has_computer	0.78
household_has_broadband	0.74
bachelors	0.70
household_has_smartphone	0.68
hs_grad	0.54
asian	0.46
households_speak_asian_or_pac_isl	0.43
households_speak_other_indo_euro_lang	0.40
households	0.27
pop	0.26

Table 4: Top 10 Features having Positive Correlation with Median Household Income

In Table 4, we see that the top three features having the strongest positive correlation with median household income are `household_has_computer`, `household_has_broadband`, and `bachelors`. This makes sense because if a county has many households with computers and broadband, it indicates that families have enough money to buy these "non-essential" goods and are not struggling to make ends meet. And `bachelors` makes sense because many high paying jobs require bachelor's degrees. Furthermore, these features have a causal relationship with income, as having a computer and a degree allows one to advance in their career.

Because `household_has_computer` has the strongest positive correlation, we explore its distribution:

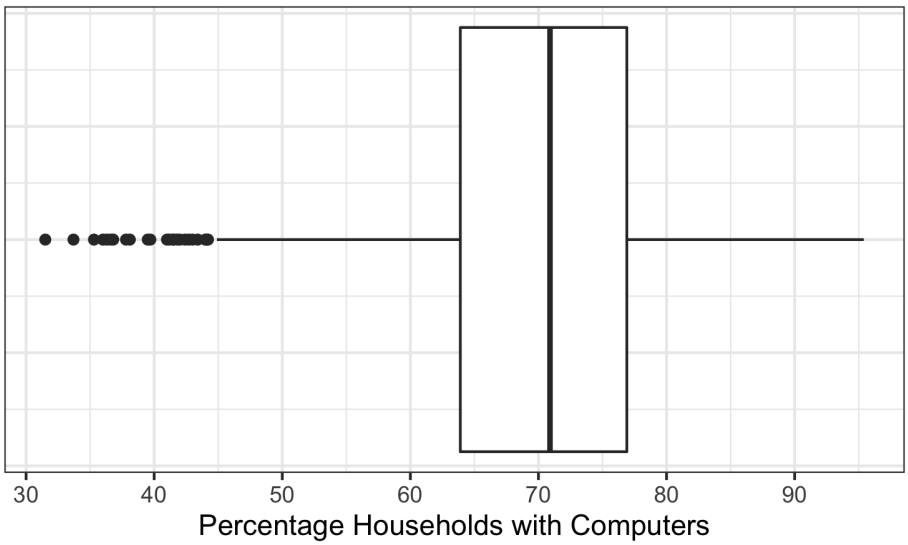

Figure 5: Boxplot of Percent of Households with Computers

In the boxplot in Figure 5, we can visualize the distribution of the feature. We notice a slight left skew as the left side of the median line is longer than the right side. Additionally, we notice a lot of outliers on the left side for low percentage of households having a computer. This means there are many counties with an extremely low percentage of households having a computer, and this might be a good indicator for those counties having low median household incomes. As for summary statistics of this feature, it appears that the 25th percentile is around 63%, the median is around 71%, and the 75th percentile is around 77%.

Next, let's explore how some of these above features work together in relation to median household income.

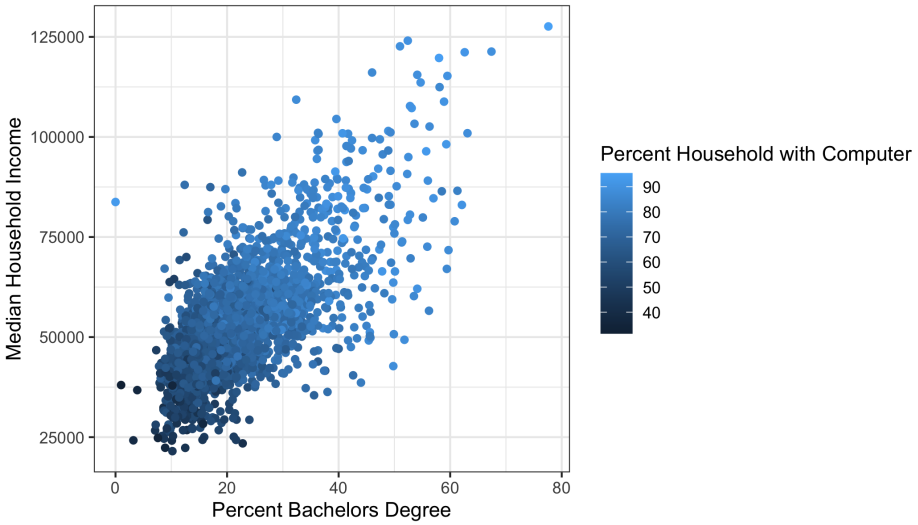

Figure 6: Median Household Income vs. Percent Bachelor's Degree and Households with Computer

In Figure 6, there are a few important observations. First, median household income is positively correlated with the bachelor's degrees, given the positive slope of the scatterplot. Second, the points are colored by their ownership of computers, with darker points representing low computer ownership and lighter points representing high computer ownership. We see that household income is similarly correlated with computer ownership, as the darker points are on the bottom of the graph and the lighter points on the top. Finally, we notice a positive relationship between bachelor's degrees and computer ownership. The dark points are on the left (counties with few bachelor's degrees have low computer ownership), and the light points are on the right (counties with many bachelor's degrees have high computer ownership). There appears to be a slight "fanning out" effect in the upper right quadrant of the graph, indicating there is more variation in household income as the proportion of bachelor's degrees increase. However, the relationships between the features and response is still clear.

Now let's take a look at features having strong negative correlations with median household income.

Feature	Correlation
poverty	-0.61
poverty_under_18	-0.61
housing_one_unit_structures	-0.54
poverty_65_and_over	-0.46
unemployment_rate	-0.43
uninsured	-0.32
black	-0.28
age_over_65	-0.26
median_age	-0.11
uninsured_under_19	-0.11

Table 5: Top 10 Features having Negative Correlation with Median Household Income

In Table 5, we see that the top features having the strongest negative correlation with median household income are measures of poverty and percent of housing units in 1-unit structures. Poverty measures make sense: more people in poverty will drive down the median household income. However, we decided to keep these measures of poverty as features because higher poverty doesn't necessarily directly translate to lower median household income; there are more factors at play. At first glance, `housing_one_unit_structures`, or percent of housing units in 1-unit structures, having a negative correlation with median household income seems confusing. But, this likely points to the fact that if more people in a county live in 1-unit structures, the property in the county might be cheap. Cheaper property implies lower cost of living, which may lead to lower paying jobs and lower income.

Now, lets take a closer look at some of these features. Let's explore the distribution of `poverty`, or the percent of population below the poverty level, as it has the strongest negative correlation.

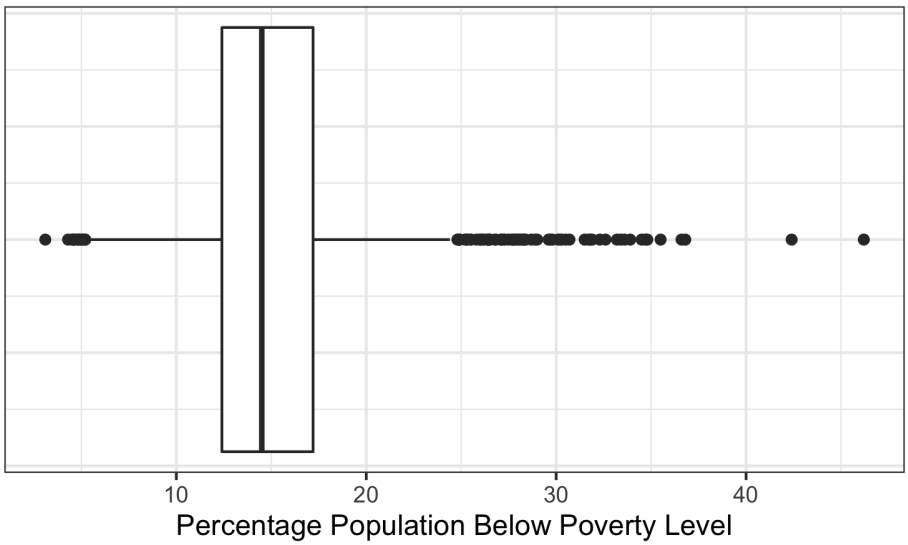

Figure 7: Boxplot of Percent Below Poverty Level

In the boxplot in Figure 7, we can visualize the distribution of `poverty`. We notice a slight right skew as

the right side of the median line in the boxplot is longer than the left side. Additionally, we notice a lot of outliers on the right side for high percentage of the population living below the poverty level. These are the counties with an extremely high rate of poverty. Additionally, on the left side, there are also a few outliers, though fewer than the number of outliers on the right side. These outliers are for counties with extremely low rates of poverty. These outliers might be a good indicator for those counties with very low and very high median household income. As for summary statistics of this feature, it appears that the 25th percentile is around 12%, the median is around 14%, and the 75th percentile is around 17%.

Next, let's explore how some of these negatively correlated features work together in relation to median household income.

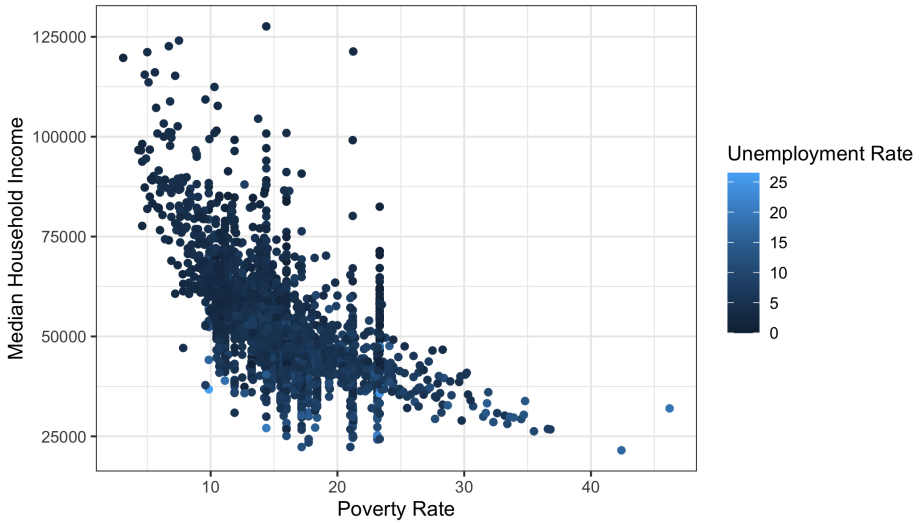

Figure 8: Median Household Income vs. Percent Poverty and Unemployed

In Figure 8 we can visualize the negative correlation that both poverty rate and unemployment rate have on median household income. Poverty has a negative correlation, based on the negative slope of the scatterplot. Likewise, unemployment rate has a negative correlation, as the lighter points (higher unemployment) lie towards the bottom of the graph while the darker points (low unemployment) lie towards the top of the graph. The gradient of colors is not as clear, however, in the middle of the graph for counties with mid-level household income ($40,000 - $60,000), but we still suspect that unemployment has a negative correlation with household income. When exploring the relationship between poverty and unemployment, the counties with the highest unemployment rates are not necessarily the same counties with the highest poverty rates, as there are light points that lie towards the left-hand side of the graph. Regardless, when we look at the counties with the highest household income, they all have very low unemployment rates and poverty rates.

4 Modeling

4.1 Regression Methods

4.1.1 Ordinary Least Squares Regression

To start the modeling task, we build an ordinary-least-squares regression to predict median household income using all 41 features. We believed this would be a good starting point to examine if the features were able to explain at least a sizable portion of the variance in the response. However, because ordinary-least-squares uses all 41 features, it is prone to common pitfalls such as overfitting and multicollinearity between features.

```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -4.420e+06  5.153e+06  -0.858 0.391084
age_over_18                         -2.131e+02  6.650e+01  -3.205 0.001370 **
age_over_65                         -1.454e+03  7.437e+01 -19.552  < 2e-16 ***
age_over_85                          1.115e+03  1.803e+02   6.186 7.18e-10 ***
age_under_5                          6.644e+02  1.609e+02   4.130 3.75e-05 ***
asian                                1.574e+03  1.442e+03   1.092 0.274926
avg_family_size                     -1.082e+04  1.058e+03 -10.229  < 2e-16 ***
bachelors                            4.770e+02  2.245e+01  21.245  < 2e-16 ***
black                                1.419e+03  1.436e+03   0.988 0.323141
hispanic                            -1.792e+02  1.542e+02  -1.162 0.245330
household_has_broadband              1.847e+01  3.157e+01   0.585 0.558663
household_has_computer               2.112e+02  3.165e+01   6.673 3.08e-11 ***
household_has_smartphone             2.387e+02  2.806e+01   8.507  < 2e-16 ***
households                          -3.941e-03  1.099e-02  -0.359 0.719862
households_speak_asian_or_pac_isl    7.092e+02  1.876e+02   3.780 0.000161 ***
households_speak_limited_english     4.519e+01  7.201e+01   0.628 0.530285
households_speak_other              -2.241e+01  4.843e+01  -0.463 0.643651
households_speak_other_indo_euro_lang  3.297e+02  5.514e+01   5.979 2.57e-09 ***
households_speak_spanish            -1.794e+02  5.214e+01  -3.440 0.000591 ***
housing_mobile_homes                 4.249e+04  5.147e+04   0.826 0.409108
housing_one_unit_structures         -7.554e+01  1.651e+01  -4.575 4.99e-06 ***
housing_two_unit_structures          4.275e+04  5.147e+04   0.831 0.406268
hs_grad                              5.364e+01  3.998e+01   1.342 0.179787
mean_work_travel                     1.789e+02  2.928e+01   6.109 1.16e-09 ***
median_age                           1.215e+03  6.501e+01  18.685  < 2e-16 ***
native                               1.472e+03  1.436e+03   1.026 0.305208
other_single_race                    1.769e+03  1.444e+03   1.225 0.220743
pac_isl                              1.011e+03  1.453e+03   0.696 0.486463
persons_per_household                1.903e+04  1.414e+03  13.455  < 2e-16 ***
pop                                 -7.347e-04  3.834e-03  -0.192 0.848076
poverty                             -6.337e+02  8.013e+01  -7.908 3.90e-15 ***
poverty_65_and_over                  3.656e+02  7.184e+01   5.089 3.88e-07 ***
poverty_under_18                     1.498e+01  4.908e+01   0.305 0.760302
two_plus_races                       1.436e+03  1.438e+03   0.999 0.318056
unemployment_rate                   -5.218e+02  5.779e+01  -9.030  < 2e-16 ***
uninsured                           -5.358e+02  4.559e+01 -11.751  < 2e-16 ***
uninsured_65_and_older              -1.296e+02  1.329e+02  -0.975 0.329740
uninsured_under_19                   2.919e+02  4.660e+01   6.264 4.40e-10 ***
uninsured_under_6                    2.210e+01  3.241e+01   0.682 0.495324
veterans                            -2.319e+01  5.345e+01  -0.434 0.664442
white                                1.754e+03  1.445e+03   1.213 0.225072
white_not_hispanic                  -4.003e+02  1.648e+02  -2.429 0.015212 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5131 on 2471 degrees of freedom
Multiple R-squared:  0.867,     Adjusted R-squared:  0.8648
F-statistic: 392.9 on 41 and 2471 DF,  p-value: < 2.2e-16
```

Figure 9: Summary of Ordinary-Least-Squares Regression

As we can see in Figure 9, the ordinary-least-squares regression has an R^2 of 0.867, which means that the 41 features explain 86.7% of the variance in median household income. This shows that there is a relatively strong relationship between the features and the response, which is promising for more robust models that can address the pitfalls of overfitting and multicollinearity through penalization.

Furthermore, an examination of which features were significant corroborates exploratory data analysis. 6 of the 10 features most positively correlated with median household income are significant, mainly related to technology (household having broadband, computer, and smartphone) and other demographics (having a bachelor's degree or speaking Asian/Indo-European language). Likewise, 8 of the 10 features most negatively correlated with median household income are significant, mainly related to poverty, unemployment, lack of insurance, and old age. The signs of these coefficients mostly match as well – positively correlated features have positive coefficients, and negatively correlated features have negative coefficients, as expected. One would expect having access to technology to be predictive of higher median household income, and one would expect unemployment and lack of insurance to be predictive of lower median household income.

4.1.2 Ridge Regression

We suspect that there is potential overfitting in the ordinary-least-squares regression. Furthermore, the correlation between variables could cause the regression to become unstable and to unnecessarily add features to the model when other features are sufficient. This would lead to a large increase in variance without a large decrease in bias, since the addition of a correlated variable may add little explanatory power but high variance. As such, we explore three robust regression methods that use different methods of penalization to address the aforementioned problems and output more parsimonious models.

We start with ridge regression, which disincentivizes large values to reduce variance. Specifically, ridge regression is beneficial when dealing with multicollinearity as it can "split the credit" among correlated features. To train a ridge regression fit, we used 10-fold cross-validation and the 1-standard-error rule to optimize the penalization parameter, λ, which controls the flexibility of the resulting model.

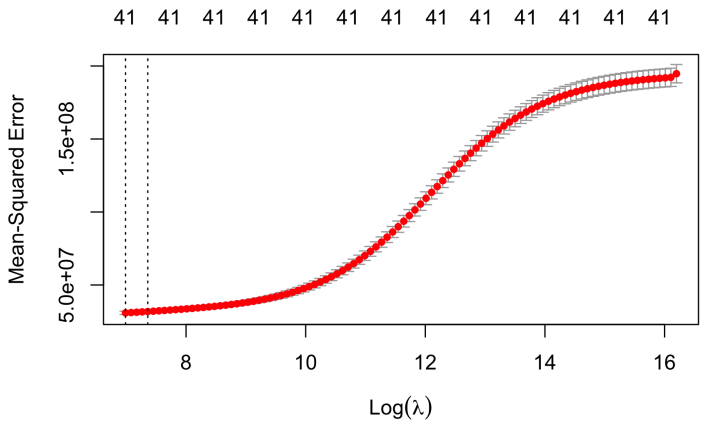

Figure 10: CV Plot of Ridge Regression

As we can see in Figure 10, which depicts the CV curve, the optimal value of λ is not very high, as indicated by the dashed vertical lines lying closer to the left-side of the plot. Furthermore, the CV curve is not U-shaped, as one may expect. Normally, the U-shape indicates that the penalization is effective in reducing variance (and thus expected test error), until over-penalization leads to an increase in bias. However, the CV curve only increases from the left-hand side, which indicates that penalization does not help reduce expected test error to a significant extent. This would lead us to believe that bias, not variance, is the primary driver of expected test error. This means that, from a preliminary examination of the CV plot, we may expect that the ridge regression will not represent a significant improvement from the ordinary-least-squares regression.

Even if this is the case, it may be insightful to examine the trace plot to see which features enter the model with the largest absolute standardized coefficients.

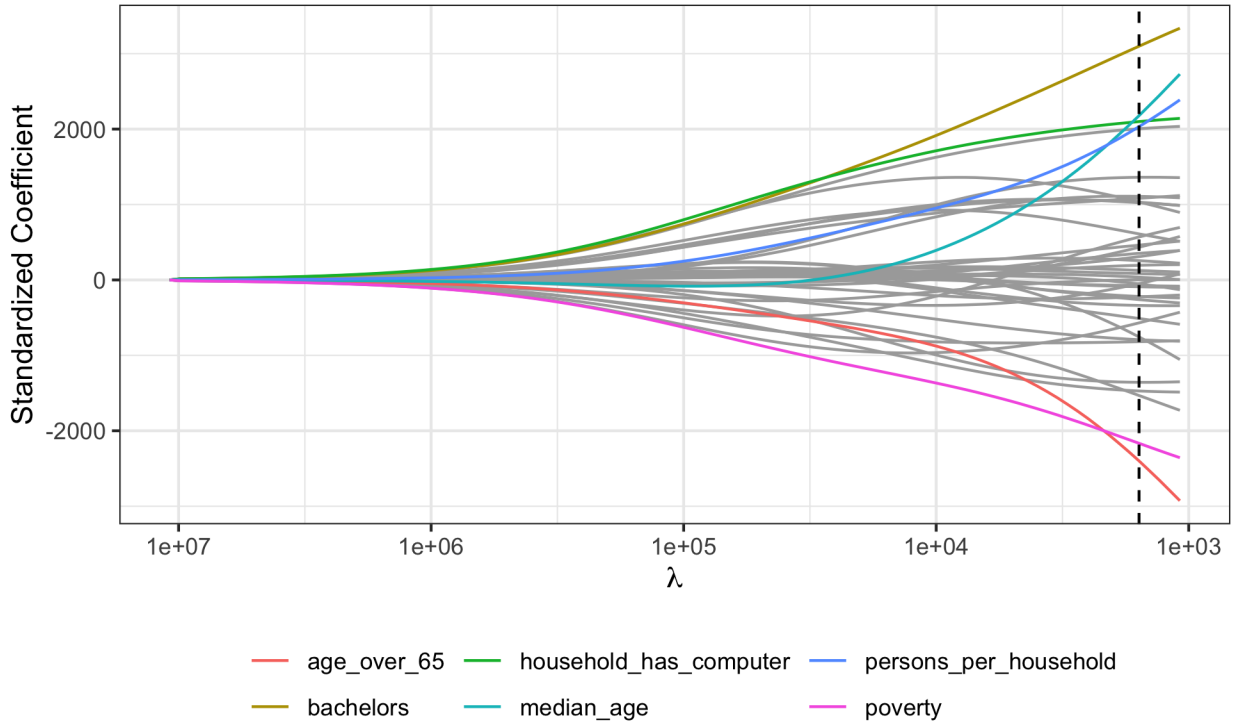

Figure 11: Trace Plot of Ridge Regression

Figure 11 highlights the 6 largest absolute standardized coefficients. As we can see, the features with the largest positive coefficients in the ridge regression are `bachelors, household_has_computer, median_age, persons_per_household`. These features make sense: people that have bachelor's degrees are more well-educated and thus likely work in higher-paying jobs and have higher household incomes. Similarly, having a computer also indicates access to discretionary income to purchase a laptop/desktop. Next, as median age of a county increases, we expect households to be more wealthy, since people generally earn higher incomes as they progress in their careers and get promoted/move to higher-paying jobs. Finally, having more people per household generally implies that the household as a whole is making a larger income in order to support everyone within the household. On the other hand, the features with the largest negative coefficients in the

ridge regression are `age_over_65, poverty`. While median age is a positive predictor of household income, we only expect this to hold true up to a certain point. Once a large portion of the county is over 65 years old, which is the canonical retirement age, then a larger percentage of the county consists of non-working members. This then results in lower household income. Likewise, we intuitively expect poverty to be a negative predictor of household income. When a county has a large percentage of its population living under the poverty level, it indicates that the county is most likely suffering economically, i.e. does not have access to high-paying jobs or high-quality education, further resulting in lower median household income.

4.1.3 Lasso Regression

We move to another form of penalization, lasso regression. The strength of lasso regression is in pushing coefficients to 0, which both reduces variance and leads to more interpretable, parsimonious models by selecting only the important features. To train a lasso regression fit, we used 10-fold cross-validation and the 1-standard-error rule to optimize the penalization parameter, λ, which controls the flexibility of the model. This results in the following cross-validation curve:

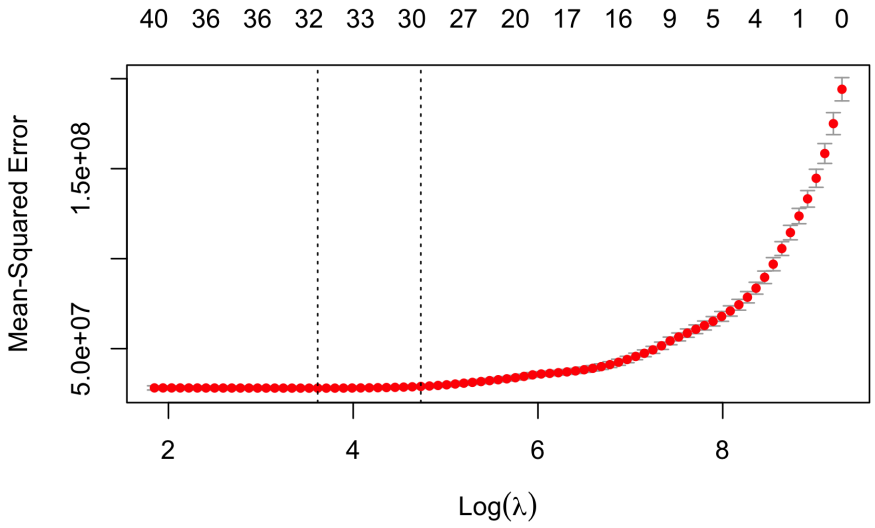

Figure 12: CV Plot of Lasso Regression

Once again, Figure 12 demonstrates some of the similar conclusions as Figure 10, the CV plot of the ridge regression. The lack of a large dip in CV error in the left-side of the CV plot indicates that variance is not the main driver in expected test error – bias is instead. Furthermore, given the numbers on the top of the CV plot, the resulting model chooses between 27 and 30 variables, which suggests that the model needs a relatively large number of variables (75% of all variables) to have enough power to explain household income.

Once again, it may be insightful to examine the trace plot of the lasso regression to see which variables first enter the model. We can also examine the top 10 most important coefficients (by magnitude) that the lasso regression selects.

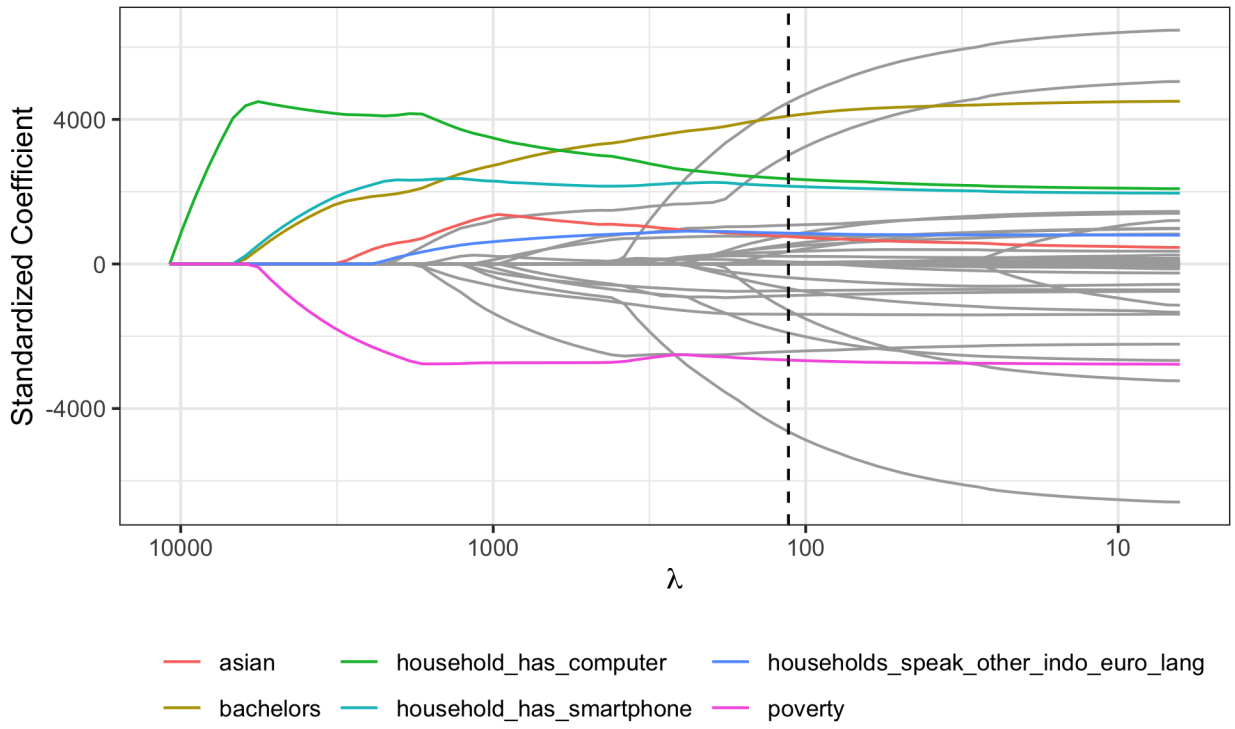

Figure 13: Trace Plot of Lasso Regression

Figure 13 shows the trace plot for the lasso regression. Because the trace plot shows the first variables to enter the model, and not necessarily the most important variables at the chosen λ, we output the top 10 important features below.

Feature	Coefficient
age_over_65	-4640.49
median_age	4469.63
bachelors	4093.39
persons_per_household	3010.23
poverty	-2659.27
housing_mobile_homes	-2422.82
household_has_computer	2355.95
household_has_smartphone	2152.38
uninsured	-1906.47
unemployment_rate	-1391.39

Table 6: Lasso Regression Variable Importances

Most of these features in Table 6 also appeared in the ridge trace plot as the most important features, and with the same signs. This corroborates our previous analysis of many of these predictors, which span across demographic, economic, technological, and educational circumstances. See **Ridge Regression** for an in-depth analysis of many of these features.

4.1.4 Elastic Net Regression

Because both ridge and lasso regression have their merits, it is often advisable to run an elastic net regression, which combines both penalties with a weight that comes from the α parameter. However, in this case, after multiple iterations of cross-validation, the optimal α parameter was chosen to be $\alpha = 1$, which reduces to lasso regression. Below is the CV plot that determined that the optimal value of α was 1.

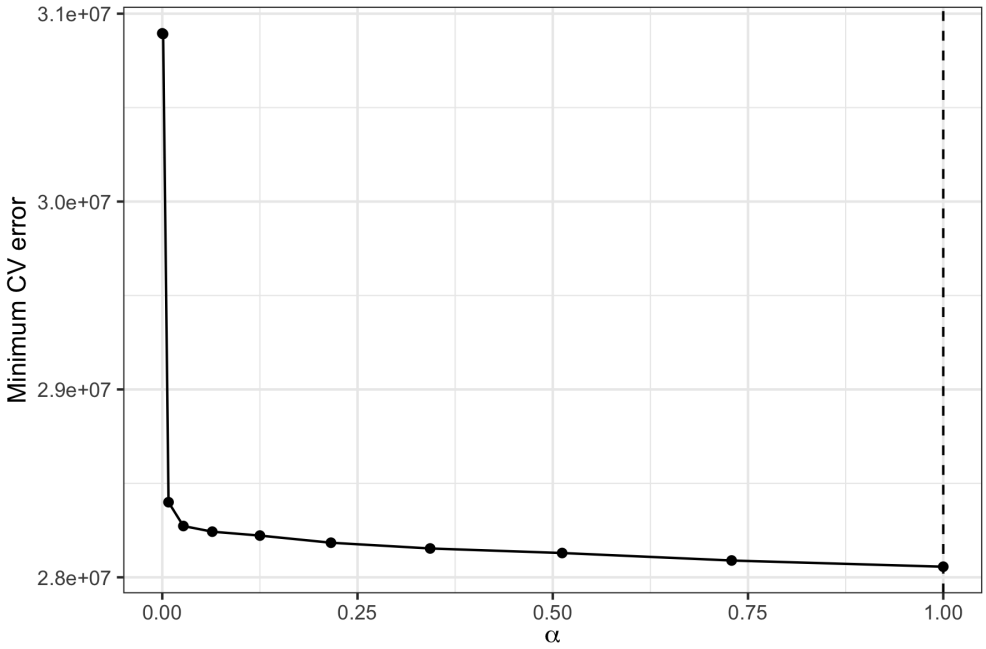

Figure 14: CV Plot for α for Elastic Net Regression

This suggests that the lasso penalty is more valuable than the ridge penalty, and that the lasso regression will likely have better performance than the ridge regression. Ultimately, our elastic-net regression is identical to our lasso regression, but we include this section for thoroughness.

Overall, given the strong starting point of our ordinary-least-squares regression, we expect that our regression-based models will perform fairly well on the test dataset. There seem to be many variables that have strong explanatory power for household income, and the intepretations of said variables is consistent with our preliminary hypotheses and exploratory data analysis.

4.2 Tree-Based Methods

While our regression methods may perform well, they can only capture linear relationships between features and the response variable. As a result, we implement tree-based methods in order to capture non-linear relationships. Furthermore, tree-based methods can be easily interpretable and inform us as to which variables are important or explain more variance in the response. For tree-based methods, we start with regression trees, then advance to more state-of-the-art models: random forests and boosted models.

4.2.1 Regression Tree

To start, we fit a conventional regression tree, which is highly interpretable. We begin with the default parameters, which results in a tree with 9 terminal nodes, as shown below:

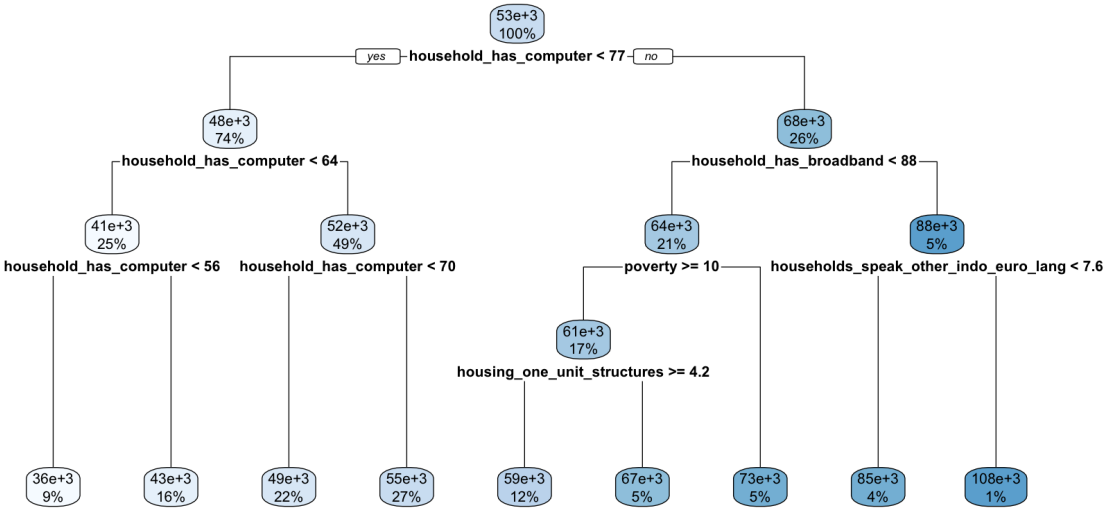

Figure 15: Default Regression Tree

From Figure 15, `household_has_computer` is very important, as it is present in 4 splits (the entire left-half of the tree) – this suggests that this feature alone may be powerful to predict whether a county has a low median household income, and if so, to what extent. Other features make sense, such as more broadband or less poverty leading to higher household incomes. However, the default decision tree may not be optimal, so we used cross-validation and cost-complexity pruning to find the optimal number of terminal nodes.

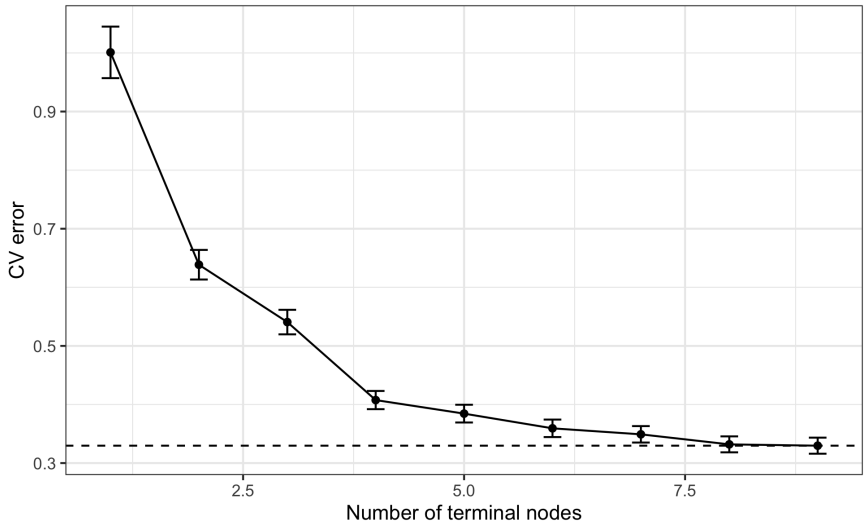

Figure 16: Regression Tree CV Plot

From the 1-standard-error rule in Figure 16, the optimal number of terminal nodes to minimize CV error is 8, one fewer than the default tree above. Then, we can use the pruning algorithm to extract the optimal tree, as shown below:

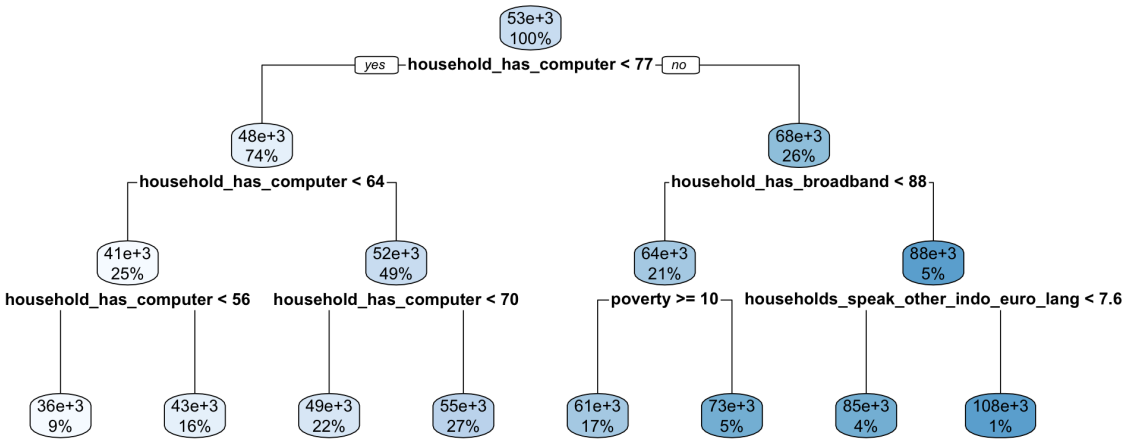

Figure 17: Optimal Pruned Regression Tree

Figure 17 shows the new pruned tree, which removed the split on `housing_one_unit_structures`. Once again, `household_has_computer` is evidently important. Following the splits and paths from the root to different nodes leads to interpretable results ($\geq$ 77% having computers, $\geq$ 88% having broadband, $\geq$ 7.6% speaking Indo-European languages leads to the highest predicted median household income). To further examine which features were selected as important by the optimal regression tree, we can examine the variable importances below.

Feature	Importance
household_has_computer	290379530522
household_has_broadband	224214185122
bachelors	120629493642
household_has_smartphone	90612310469
poverty	74048052051
asian	47312676681
hs_grad	28651357890
housing_one_unit_structures	19530196909
poverty_65_and_over	18471287019
poverty_under_18	16386865314

Table 7: Optimal Pruned Regression Tree

Table 7 is consistent with our analysis. `household_has_computer` is the most important feature, followed by `household_has_broadband`, with `household_has_smartphone` in fourth. Our regression tree thus indicates that owning technology, which is indicative of having higher discretionary income, is tremendously important in predicting median household income. Other features that we have seen and explored prior (see **Regression**

Methods also appear, such as `bachelors, poverty`, and `asian`. The features `hs_grad` and `bachelors` together suggest that education is an important predictor of household income, as well. These education features do not appear in our pruned tree (despite `bachelors` being the third most important feature); this is likely because the splits on `household_has_computer` is sufficient in capturing the explanatory power of `bachelors` as well. This is evident in Figure 6 in our exploratory data analysis, which demonstrates that owning computers and having a bachelor's degree are closely related, both with each other and with median household income.

4.2.2 Random Forest

While individual regression trees may be interpretable, they may not have the performance we desire, because they are subject to high variance. One solution to this would be bagging, which involves using bootstrap samples to fit different independent regression trees, which then aggregate their predictions. While this is an improvement from a single tree, considering all features at every split point nevertheless leads to high variance. As a result, we implement a random forest, which only uses a subset of the features at each split point and thus decorrelates the trees from one another, further reducing variance and improving performance.

To start, we fit a default random forest to obtain a preliminary sense of how many trees to use in the random forest and for tuning. This is displayed in a plot of OOB error vs. number of trees below:

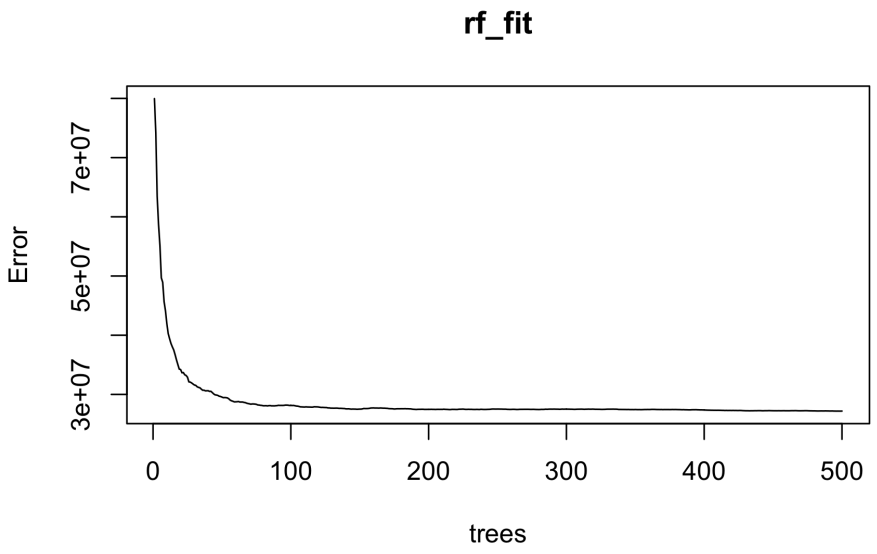

Figure 18: OOB Error vs. Number of Trees for Default Forest

As demonstrated in Figure 18, OOB error stabilizes after around 100 trees. This allows us to use 100 trees for tuning the m parameter (how many features to sample at each split), in order to save on computational cost without sacrificing accuracy. As such, the next step and main parameter to tune is m: if m is too small, the forest will lack predictive power and suffer from high bias; but if m is too large, the forest will approach bagging and thus suffer from high variance as the trees will be correlated with one another. To tune m, we

trained new random forests with 100 trees using m values ranging from 1 to 41 (there are 41 features), then plotted the OOB error for each value of m, shown below:

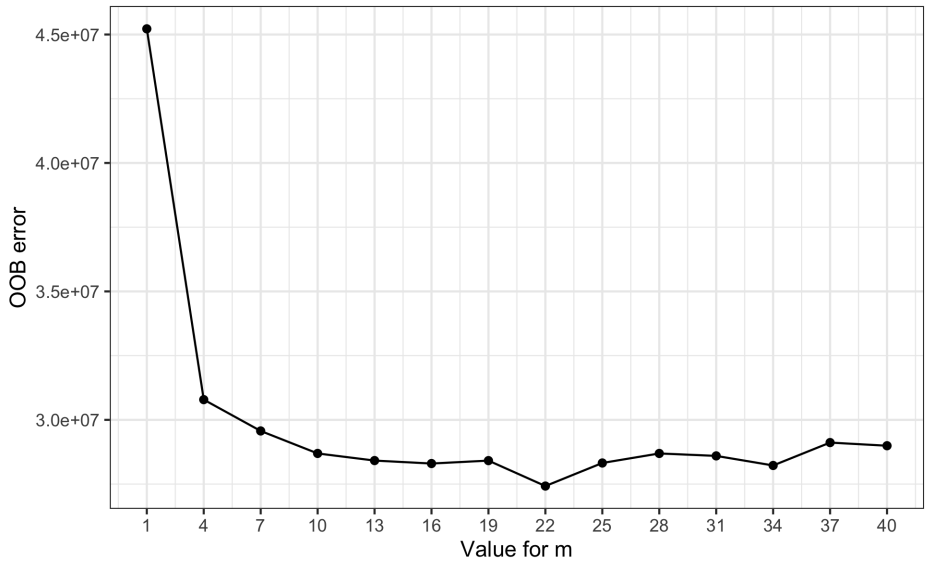

Figure 19: OOB Error vs. m

Figure 19 shows the U-shaped error curve. As we can see from the graph, the optimal value of m is 22, which is higher than the default value of $\lfloor p/3 \rfloor = 13$. This suggests that the optimal random forest uses more features to reduce bias more than the subsequent increase in variance, as can also be seen from the curve in the plot above. Then, we can extract the tuned random forest and train it using 500 trees (which we do not need to tune). For quick verification, we confirm that 500 trees is enough to see OOB error stabilize:

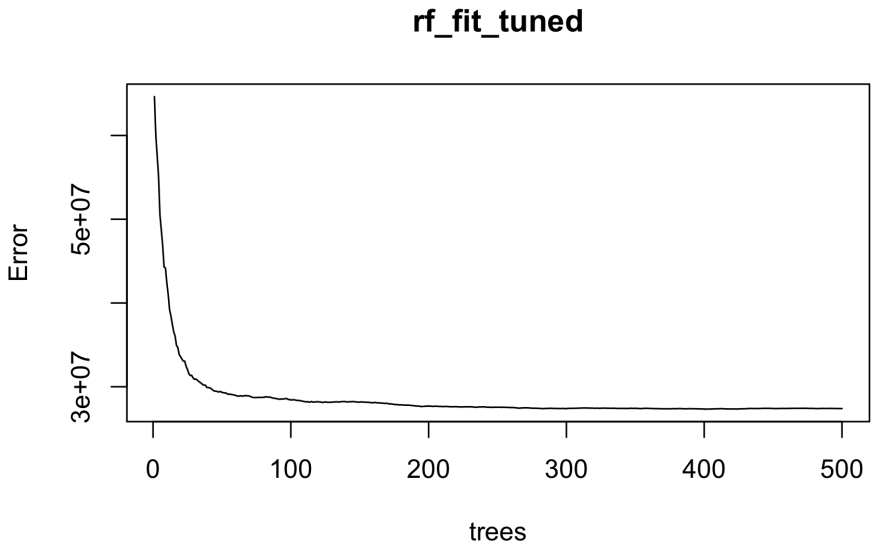

Figure 20: OOB Error vs. Number of Trees for Tuned Forest

Now that we have tuned and extracted the optimal random forest, we can obtain variable importances using two measures: purity-based and OOB-based variable importance:

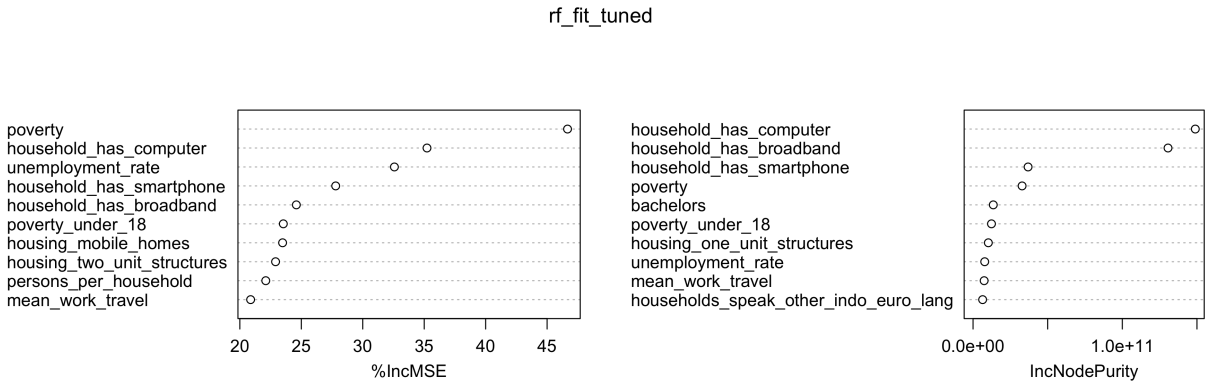

Figure 21: Variable Importances from Tuned Random Forest

Figure 21 shows considerable overlap in the most important features across both measures of importance. Furthermore, these features also corroborate those chosen by the regression tree, lasso regression, and ridge regression. As such, the notable features that appear in both measures have been explored in detail previously. As common to the other models, the percentage of a county owning technology such as computers, smartphones, and broadband internet is incredibly valuable in predicting median household income, because it provides information about a household's discretionary income. Another feature we see represented is `unemployment_rate`. This makes sense, because as a county's unemployment rate increases, fewer members of each household are earning income, which thus lowers the median household income. Thus, unemployment rate, together with measures of poverty, are all important in predicting median household income. This is also consistent with exploratory data analysis, which showed that these measures were among the most negatively-correlated with median household income. Finally, another feature we see is `mean_work_travel`, the average time it takes to travel to work. The relationship between this feature and income is less clear, and this feature likely plays a more nuanced role in the model. It is common that low-income workers have to travel long distances to work, especially in public transportation, which can be arduous and slow. At the same time, many high-earners live in suburbs and commute to a nearby metropolitan city for work, which can also result in longer average travel times. Although this feature appears towards the bottom of both plots, it is nevertheless important and interesting for further analysis.

4.2.3 Boosted Model

Finally, we trained a boosted model, another state-of-the-art method for aggregating regression trees to improve performance. While random forests grow independent deep trees in parallel, boosting grows shallow decision trees sequentially. By starting with a weak learner and then iteratively applying it to the residuals, boosting can gradually learn the trend in the data. The main parameters to tune are how many trees to fit and the interaction depth of each tree. To start, we train three boosted models, each with 1000 trees, a shrinkage factor of 0.1, and interaction depths of 1, 2, and 3. With 5-fold cross-validation, we then extract and plot the CV errors for the three models below:

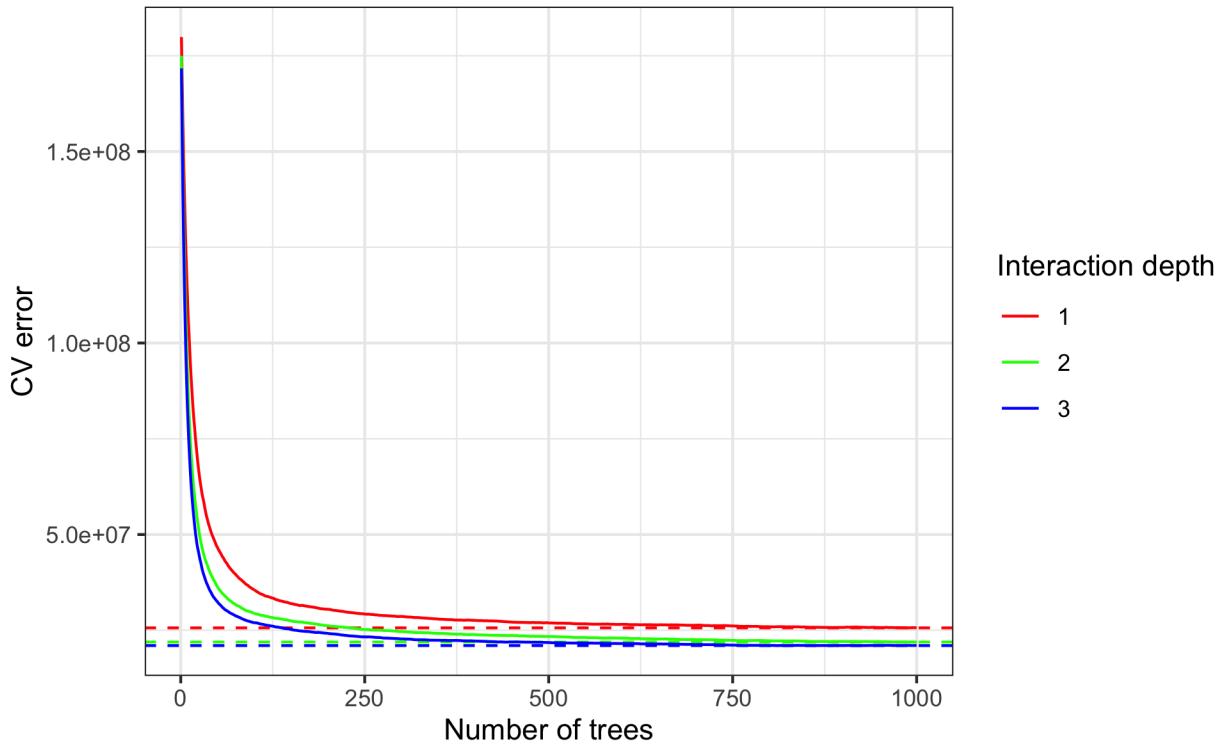

Figure 22: Boosted Model CV Plot

As we can see from Figure 22, the optimal interaction depth that minimizes CV error is 3. This occurs at 810 trees. That is, using an interaction depth of 3 with 810 trees leads to the global minimal CV error of all combinations of interaction depths and number of trees we tested above.

Once we used this information to extract the optimal boosted model, we can output feature importances.

Feature	Relative Influence
household_has_computer	27.18
household_has_broadband	20.07
household_has_smartphone	10.51
poverty	9.32
housing_one_unit_structures	2.86
unemployment_rate	2.56
mean_work_travel	2.51
bachelors	2.37
uninsured	1.97
persons_per_household	1.59

Table 8: Boosted Model Variable Importances

As we can see from Table 8, the boosted model reaffirms the most important features in predicting median household income. Once again, access to technology (computers, broadband internet, and smartphones) is

crucial in predicting household income. Other measures such as poverty, unemployment, and lack of insurance also are important in predicting household income. All of these variables have appeared in previous models: see previous sections for in-depth analyses of these variables.

Beyond the feature importances, boosted models allow us to output partial dependence plots. Shown below are the partial dependence plots for the four most important features:

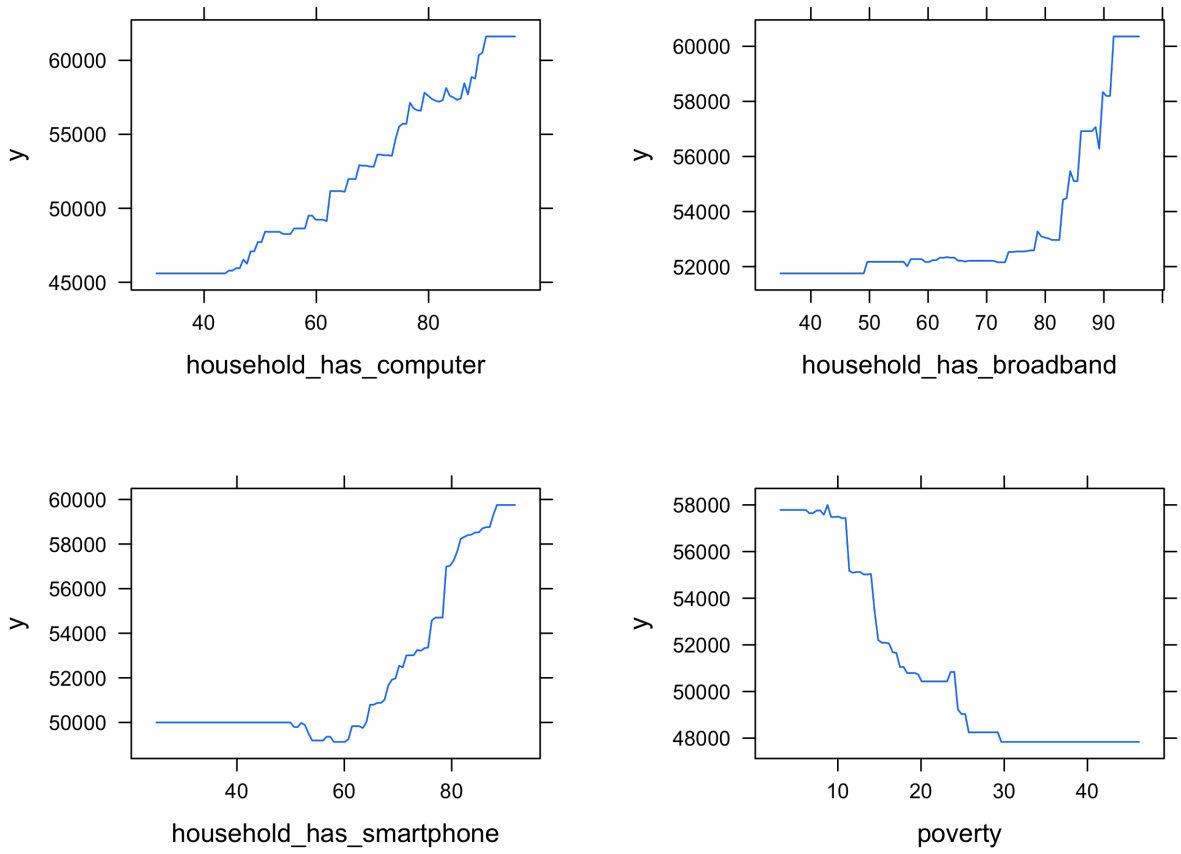

Figure 23: Boosted Model Partial Dependence Plots

As we can see from Figure 23, the relationships follow what we expect, and have seen with other models. Because we are using trees with an interaction depth of 3, these partial dependency plots are not exact representations of how one variable impacts the response; however, they are still useful in examining and confirming the relationships between important features and household income. The first three features and plots are about owning technology. As such, these have positive relationships with median household income. On the other hand, an increase in poverty levels leads to a fall in median household income.

As we explored earlier in this paper (see **Random Forest**), `mean_work_travel` (average time it takes to travel to work) is of particular curiosity, because it does not seem to have a clear relationship on median household income. Here, `mean_work_travel` appears again as one of the most important features in the

boosted model. This allows us to take advantage of the partial dependency plots to examine how this feature impacts the response more closely.

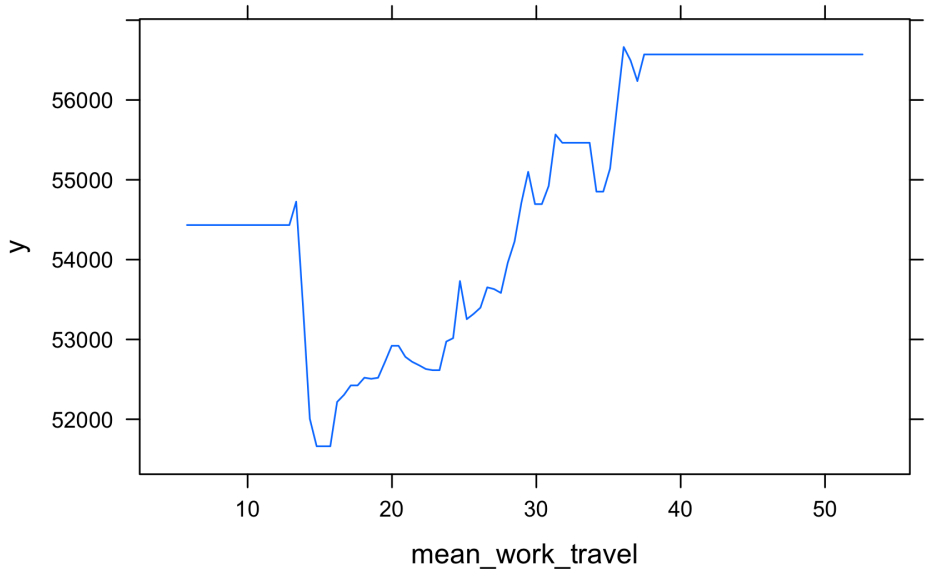

Figure 24: Boosted Model Travel-to-Work Partial Dependence Plots

Figure 24 confirms this hypothesis: the partial dependency plot is more U-shaped than linear, and this falls in line with the previous analysis. On the lefthand side, when travel-time to work is the shortest, these are likely indicative of households where people can work-from-home more often, or have the luxury to live very close to the office. As such, these are more aligned with middle-income jobs that allow for these accommodations. Then, as travel-time to work increases, median household income falls. This is likely indicative of households where people have to commute to-and-from work and do not have the luxury of working from home or living close to the office. This falls more in line with lower-paying jobs. Then, on the right-hand side, as travel-time to work increases even more to almost an hour, median household income rises the highest. This is likely indicative of wealthier people who reside in suburbs and commute to metropolitan areas for work, or even go on business trips. Such a long travel-time falls more in line with higher-paying jobs. As such, we get a parabolic shape, where medium travel-times are predictive of lower-paying jobs, short travel-times are predictive of medium-paying jobs, and long travel-times are predictive of higher-paying jobs.

Overall, from the evaluation of the selected features and the partial dependency plots, it appears that the boosted model effectively learned the signal in the data. The partial dependency plots show clear relationships with the response that one might expect. As such, we expect this model to perform among the best on the test data.

5 Conclusions

5.1 Method Comparison

Below, we display the test RMSEs for all methods, along with the test RMSE of the intercept-only model, in decreasing order of test RMSE.

Model	Test RMSE
Intercept-Only	15109.78
Tuned Regression Tree	8167.43
Default Regression Tree	8156.14
Ridge Regression	5765.52
Lasso Regression	5552.67
Elastic Net Regression	5552.67
Ordinary-Least-Squares Regression	5526.98
Default Random Forest	5505.63
Tuned Random Forest	5466.62
Boosted Model	4922.06

Table 9: Model Performances

As we can see from Table 9, all of our models substantially outperform the intercept-only model. The boosted model performs the best, followed by the tuned random forest. Their test RMSEs are also attractively low, standing at around $5,000 error in median household income. Compared to the 50-th percentile median household income of $51,745, the RMSE of the boosted model represents a 9.5% error, a relatively small fraction of the response (and much smaller than that of the intercept-only model, whose RMSE is almost 30% of the median response).

Furthermore, the ordering of the models also makes sense. The tuned and default regression tree perform the worst – this is likely due to high variance, as the single regression trees perform significantly worse than the other tree-based methods further down the list. The regression-based methods (ridge, lasso, elastic net, and ordinary-least-squares) lie squarely in the middle. As discussed before (see **Regression Methods**), we expect ordinary-least-squares to perform well, and we did not expect to see penalized regressions add significant value. That hypothesis is reflected here, as ordinary-least-squares actually outperforms the penalized versions. Note that the elastic net regression RMSE is the same as the lasso regression, because our tuned elastic net reduces to lasso given the optimal value of $\alpha = 1$. Finally, the random forests and the boosted model perform the best, as expected. Given their ability to capture explanatory power in the features while also keeping variance low (by decorrelating trees, shrinkage, and other methods) through extensive parameter tuning, these models are the best at minimizing test error.

Overall, regression models are effective in explaining most of the variance in the response, but tree-based models are superior given their ability to capture non-linear relationships (as we saw with `mean_work_travel`, for example). Ultimately, not only do the models perform well, they also have significant overlap in selecting the important features that predict median household income.

5.2 Takeaways

Given the overlap between the features selected as important across multiple models, there are real world implications in predicting (and changing) median household income.

Owning technology: we have seen that in every model, owning technology (computers, smartphones, and broadband internet) was one of the most important positive predictors of median household income. We suspect that there is a two-way causal relationship here: it is true that wealthier households have enough discretionary income to purchase these technologies. However, more importantly, having access to these technologies within one's own household seems to lead to a higher median household income. This is because owning computers and smartphones allow the members of the household to (1) continue their education, which is another positive predictor of income, explored later, (2) advance in their careers, such as by working-from-home, staying connected with managers/employees, or even by searching for new and higher-paying job opportunities, and (3) accomplish all of these tasks and more without spending time seeking out these resources (such as by going to a public library to use a computer, for example). This means that, for stakeholders seeking to increase median household income in struggling counties, access to technology should not be overlooked. Policymakers should focus on making technology more accessible to both students and professionals, such as by increasing funding to provide laptops to students or supporting digital literacy to teach computer skills and convey the importance of technology to working professionals.

Economic factors: another set of features that appeared multiple times throughout our models were economic factors, namely the presence of poverty and unemployment and the lack of insurance, all of which were negative predictors of household income. While many of these features seem to have intuitive and direct relationships with household income, the causal relationships are more nuanced. The presence of poverty/unemployment or lack of insurance is not simply due to low household income. Rather, they arise because of many other factors, including inequality, low job security, poor education, lack of infrastructure, and other structural barriers (we saw this reflected in other important features, such as the percentage of people living in mobile homes). As such, policymakers should focus on the root causes for these features if they wish to increase median household income. Increasing access to important resources, from education to job security, would go a long way in boosting household income for many struggling counties.

Education: a third set of features that most of our models selected as important was education-related features, namely having a bachelor's degree and being a high-school graduate. Education is vastly important in both personal and professional success, and should be a main point of focus for stakeholders. Many jobs require/strongly prefer bachelor's degrees, and these jobs tend to be higher-paying as well, so education greatly expands the opportunity set available. Furthermore, education teaches necessary skills to both acquire and manage wealth. Policymakers should thus focus on bolstering educational resources in struggling counties: this includes increasing funding to schools and improving infrastructure (such as busing or daycare) so that all households can have access to education.

Other factors: finally, there are a few miscellaneous features that appeared multiple times. For example, the percentage of households speaking Indo-European languages was a positive predictor of household income.

This may not be directly a cause of higher household income, but it may be useful for stakeholders to examine communities with a high prevalence of Indo-European language speakers to evaluate if there is something else that could be driving the higher household income. For example, English is an Indo-European language: knowing English is crucial for obtaining and succeeding in many jobs in the United States. As a result, improving education for English-as-a-second-language learners may help boost income in those counties. Finally, an interesting feature that we explored previously was the average travel time to work, which had a nonlinear relationship with household income. It appeared that a medium travel time was associated with lower median household incomes, with higher-paying jobs either having the luxury to work from home or live far from their jobs. As such, investing more in public transportation to make it easier for people to commute to and from work could bolster job security and household income for struggling counties.

5.3 Limitations

One limitation that may arise is that the data is pre-pandemic. While we chose to use 2019 data to remove any effects of the pandemic and reflect more "normal" circumstances, it also has the implicit assumption that, if these models are used for prediction on future data, the future will be more closely aligned with the pre-pandemic world. However, it is undeniable that the pandemic has also drastically changed the U.S. COVID-19 exacerbated inequality, as struggling counties had fewer resources to combat the spread of the deadly virus while richer counties got even wealthier due to capital gains during a period of record-low interest rates, as explored in this Wall Street Journal article. As a result, it is possible that the features that predict median household income may have changed since 2019. Ultimately, because we wanted to model household income during more normal circumstances, and assumed that the U.S. will eventually return more to a pre-COVID world, this was a tradeoff we were willing to make.

Other limitations may arise from the nature of the data. Some of the features are correlated with one another, which may lead to unstable models or poor interpretability, as features may be associated with, but not causally related to, the response. Likewise, there may be reverse-causal relationships where the response causally determines one of the features and not the other way around. However, the correlation plot only shows that a certain subset of features are correlated (i.e. different measures of poverty, or different measures of uninsurance), and robust models (that use penalization, for example) should be able to address the relatively limited multicollinearity. Furthermore, while there may be reverse-causal relationships, the results are nevertheless interpretable and the models are effective in prediction.

Lastly, a few limitations may arise during the data cleaning process. Because there were many null values, we chose to impute the missing values rather than drop either the features or observations, in order to preserve more of the dataset for training/validation/testing. However, imputation is not necessarily accurate, because states are not homogenous. As described in **Data Cleaning**, we imputed missing values as the mean for that feature for the same state. This may not be accurate for states with significant heterogeneity in these features.

5.4 Follow-Ups

To follow-up on this project, we would recommend a few next-steps. First, filling in the missing values in the existing dataset would allow for more accurate training. More importantly, the collection of more features would be helpful. While our models performed well in explaining most of the variance in the response, there is the potential to have other valuable explanatory variables that were not reflected in our data. For example, more sociopolitical features, such as data about government officials, policies, voting patterns, and number/types of schools nearby may be insightful to larger trends that affect household income. Likewise, more commerce-related features, such as data about the presence of local restaurants, shopping centers, airports, and shipping ports may be useful. Counties with abundant commerce attract more business and jobs and may result in higher household income. Finally, business-related features may be valuable, such as features related to the number/types of office buildings, the presence of large company headquarters, etc. Counties with a large business-presence tend to have higher-paying jobs and attract wealthier households. Having even more features may lead to more variance, but careful model selection and tuning should combat this.

On the analysis and modeling side, we would recommend exploring more models. For example, implementing deep learning methods, namely neural networks, may be valuable in learning derived features. These models may have even greater prediction performance, but may be less interpretable than the models explored in this paper. The tradeoff between potentially improved performance with worse interpretability and higher computational costs should be explored in future iterations.

Finally, in order to address the first limitation mentioned in relation to using pre-COVID data, it would be valuable to run the same analyses on data drawn during COVID (during 2021, the first full year affected by COVID). Then, one could examine if new features are selected as important and more quantitatively compare the pre- and post-pandemic world.