STATS 503 - Final Project - Group 5

Group 5: Marco Arriola-Moscoso, Yin Kwong Lee, Martin Zanaj, Joshua Zimmer

Motivation

With the rise to prominence in the United States of socialist figures such as Senator Bernie Sanders, much of the recent political discourse has been centered around wealth inequality. Additionally, with the recent economic downturn caused by COVID-19, this issue has been thrust even more into the center stage. Much has been discussed in the literature regarding the positive associations between poverty and difficulty finding work, worsened medical outcomes, and poorer overall health and well-being. Thus, we decided to investigate the associations between median county income and a variety of economic indicators by county in the United States. It is our hope that the insights gathered from our analysis can be used to better understand what occurs in poorer neighborhoods and hopefully identify suggestions as to what can be done in order to improve those situations.

Introduction

The goal of our project was to identify the factors associated with whether the median incomes of certain counties in the United States are large or small, and examine the models best used to predict that response. Thus, we focused on the following Research Question:

Research Question: What traits most clearly distinguish wealthier neighborhoods from their less wealthy counterparts in the United States, and what models are best used to predict them?

As part of exploring this underlying research question, we selected as our response variable the binary classification problem of whether or not the median income of a county was above (1) or below (0) the United States national average of median incomes by county. In predicting this response, we employed a variety of economic indicators by county (that are detailed in a subsequent section).

With this data, we identified the following sub-questions that we explored as part of our project:

- Which of these economic indicators have the strongest predictive power for whether a neighborhood's income is above or below average — and thus is likely a key attribute of lower income neighborhoods
- Which Machine Learning model performs best in predicting this response — and thus, what is the expected linearity of the decision boundary between higher and lower income areas
- How well does the best of the explored Machine Learning algorithms predict the response — and thus how sufficient are these economic indicators in capturing the variation of the response

Specifically, the hope is that we will be able to identify the factors that are most important in distinguishing high earning counties from lower earning counties.

Data Description

In answering this question, we employed data sourced from American Fact Finder. The predictors used fell into the following categories:

1. **Employment Status (ES)** — percentage of population in the labor force employed/unemployed/in the armed forces/etc.
2. **Commuting to Work (CTW)** — percentage of the population commuting to work via different means (e.g. walking, public transport, carpooling, etc.)
3. **Occupation (OCC)** — percentage of population with various roles such as management, sales, production, transportation, etc.
4. **Industry (IND)** — percentage of population working in various industries such as agriculture, finance, education, health care, etc.
5. **Class of Worker (COW)** — percentage of population classified as self-employed, government employed, privately employed, or as unpaid family workers
6. **Health Insurance Coverage (HIC)** — percentage of population with private, public, or no health insurance including categorized by employment status

(Source: https://tinyurl.com/yxfm658p. Data dictionary: https://tinyurl.com/yd52rw68.)

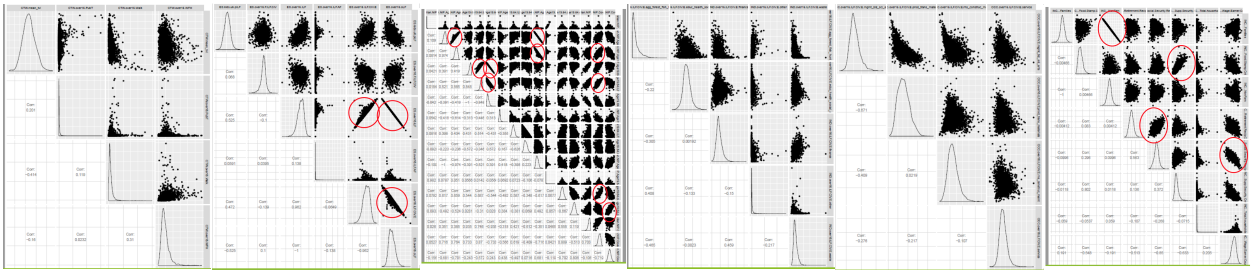
Exploratory Data Analysis

To begin our analysis, we conducted cursory Exploratory Data Analysis to get a better sense of the number of predictors and data points available to us. We observed 3,126 data points (each representing a unique county in the United States) and 82 predictor columns (1 categorical and 81 numeric) in the data set. (Actually, there was a larger number of original columns, however, certain columns such as mean income of the county were so heavily correlated with the response that we removed most of the category centered around Income (INC) so our models wouldn't rely too heavily on predictors that would add very few novel insights on the response variable.)

Next, we examined the summary statistics of the various predictors in order to get a better understanding of the means / medians / standard deviations and skewness / heavy-tailedness of their distributions; however, due to the large number of predictors, these values are not shown here.

Then, we cleaned the data — converting integer columns into percentages of their superset populations other than for items such as county population count. We did this in order to ensure that each predictor constituted an apples-to-apples comparison and the results more easily interpreted by models such as our Random Forest (so that, for example, overall size of a county would not impact all of the predictors simultaneously). Then, we renamed all variables to reflect their semantic meaning, dropped observations with empty attributes, and translated the columns into a table format more easily ingestible and useable by R. Finally, we converted the response variable from median income value into the binary response of (1) if it was above the average median income of U.S. counties and (0) if it was below the average. This conversion to the response resulted in a class split of 55.89% in class 0 and 44.11% in class 1.

Finally, we explored the pairwise plots of the various predictors by category and observed the following.



From left to right, these plots show the pairwise scatterplots of features and their density functions in the categories, CTW, ES, HIC, IND, OCC, and INC. Due to the very large number of features, we were forced to shrink down these plots, which makes them somewhat hard to read. However, they key takeaway here was that we were able to identify some highly correlated features (circled in red) which we decided to remove in the event that they caused $X^T X$ to become singular in the application of our linear models.

Complexity Reduction

Again, due to the large number of features, our exploration of basic summaries and correlation matrices became rather difficult. Thus, for these reasons, the daunting inefficiency, and the difficulties in the interpretability of linear models with a large numbers of features, we decided to shift the EDA process towards trying to reduce to number of predictors to only those that accounted for the largest percentage of variation in the response and focused the rest of the analysis on these most important features.

Feature Importances

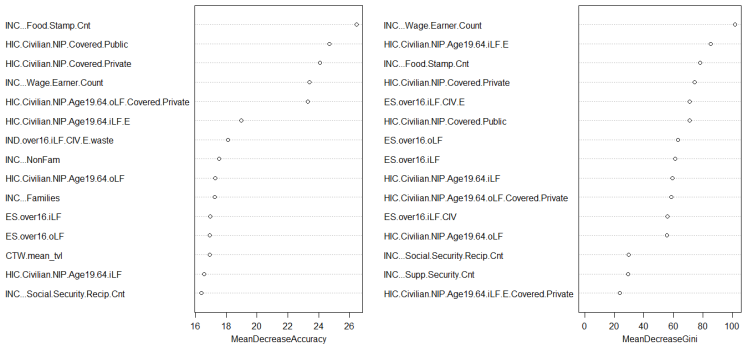
In exploring feature importance for complexity reduction, we employed two methods: Feature Selection via Random Forest Importances and via PCA. Additionally, because the Random Forest showed the best predictive power (as described later in the Random Forest section), we interpreted the associated feature importances of the Random Forest.

Random Forest

To start, the team identified the most important features as identified by the Feature Importances of the Random Forest model in the response variable. Iteratively, the team reduced the number of features to only those identified in the "Top X" of the Random Forest model, both from the "Top X" features from Mean Decrease Accuracy and from the "Top X" features

from Mean Decrease Gini. Ultimately, we identified that the top 43 features from Mean Decrease in Accuracy, resulted in a sufficiently high prediction accuracy of the model for us to be satisfied with this feature reduction. Regardless, predictors on the top 20 from both groups made it in the final subset of predictors.

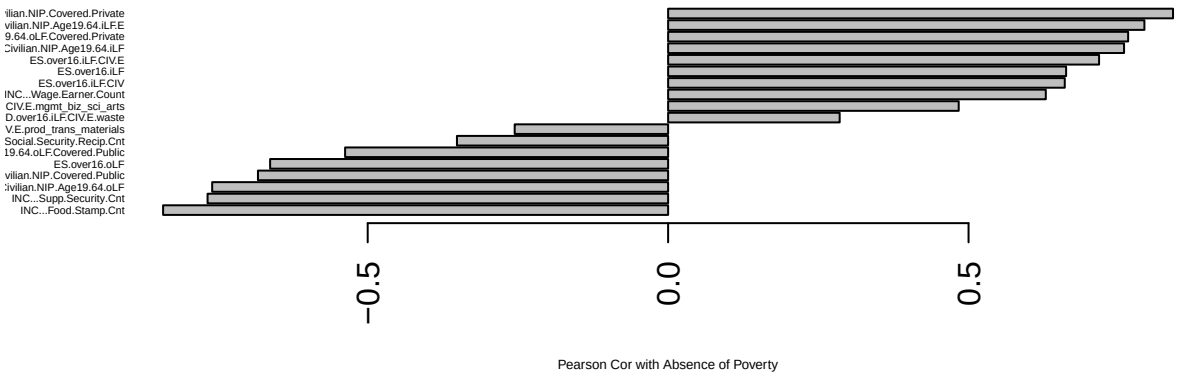
Next, we examined the top 15 Random Forest feature importances with regards to both Mean Decrease Accuracy and Mean Decrease Gini in order to better understand the economic indicators that seemed to have the strongest association with the response.



Examining this plot, we observed that, interestingly, Food Stamp Recipients, Public / Private Health Insurance Coverage, Employment Rates, Employment in the Waste Industry, Employment in the Civilian Sector vs. the Armed Forces, Mean Travel Time to Work, Wage Earner Count, and Recipients of Social Security were the strongest predictors of above or below average median income.

Correlation with Absence of Poverty

Although Random Forest feature importances are non-linear, examining their correlations with variables related to the response can give us some sense of their directionality in association to the response (although the specific numbers should not be taken at face value.) Here, we examine the Pearson Correlation between these feature importances and the percent of the county populations which are not in poverty (i.e. the percent of the population making more than $25,000 per year):



Thus, we see some expected associations and interesting insights between these most important predictors and the absence of poverty. Specifically, we see that there is a positive association between the absence of poverty and private health insurance coverage, percentage employment, private health insurance coverage by those out of the labor force, employment rate, percentage of the population in the labor force, percent of the population working in the civilian sector (vs. the armed forces), count of wage earners, percent of the population working in management, business, science, and the arts, and (very interestingly) the percent of the population working in waste. (Although those last two associations are weakly positive.)

On the other hand, we see negative correlations between the absence of poverty and the count of people on food stamps, the count of people on supplemental security income (a Social Security benefit for those below a certain income level), the percent of people out of the labor force, the percent of people covered by public health insurance, the percent of people out of the labor force covered by public health insurance, the count of people on social security, and the percent of people who work in production, transportation, or materials.

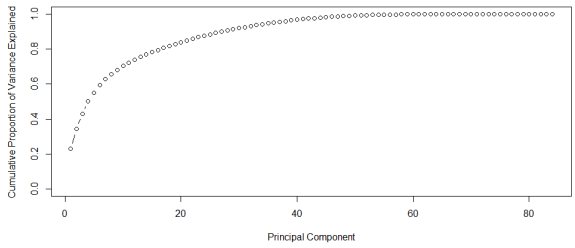
These results are extremely interesting and lead to some possible conclusions about the potential effect of certain government programs on different populations that will be discussed in the conclusion.

Principal Component Analysis

Next, we explored the usage of Principal Component Analysis so as to validate (or disprove) that the most important features found via the Random Forest methodology did indeed account for the largest percentage variation in the response. The key idea, here, was to find a subset of the predictors in our data, that could explain most of the linear variance in the data, such that we accounted for at least 80% of the variance amongst the predictors.

Cumulative Proportion of Variance

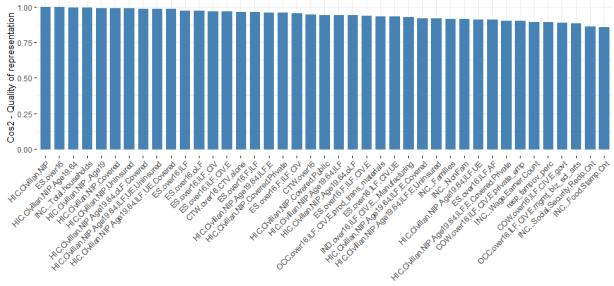
To start, we ran a correlation-based Principal Component Analysis on the predictors since some of the predictors were defined by percentages and other predictors were defined by counts of people. (Had we used the covariance-based approach the vast majority of the variation would have been accounted for by the counts, which would have likely caused issues.)



After examining a scree plot and a plot of the percentage of variance explained (only the latter is shown here for brevity), we observed that the range of 20-40 Principal Components constituted the necessary amount to sufficiently explain at least 80% of the variation in the predictors. Thus, we selected the number of our Principal Components as 20.

Next, in order to maintain the interpretability of our models, rather than simply use these 20 PCs directly, we examined the loadings of all of the underlying predictors and used the square of thier Euclidean norms (as projected onto the vector space spanned by these 20 Principal Components) in order come up with a measure of the underlying feature importances in these 20 Principal Components.

Plotting a bar graph of these values in order to identify the most influential predictors in this 20 PC vector space, we observed the following.

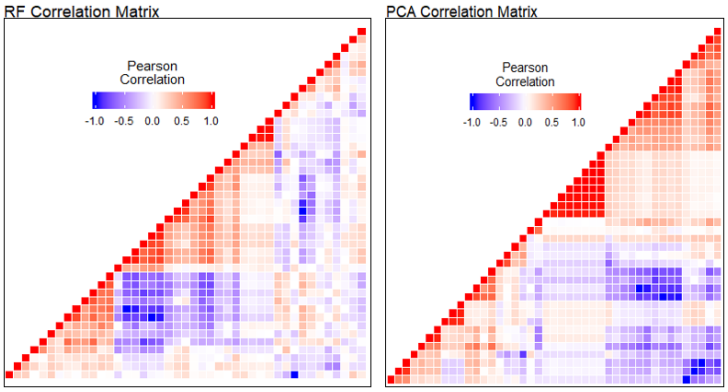


From these results, we selected the top 40 predictors as well. We selected this cut off value partially due to the importance observed of these features compared to the other predictors and partially due to the comparable result of 43 predictors found by the Random Forest.

We observed that, between the two feature sets, many of the selected variables overlapped, which lent credence to both of the methodologies, although moving forward, we decided to primarily use the features selected by the PCA process to simplify the remainder of the analysis.

Correlation & Dataset Creation

In comparing the results from both methods, we generated plots for the correlation matrices of the features selected by both the Random Forest and PCA feature importances, as shown here.



By observation, we see that the two correlation matrices have a similar pattern, which again adds credence the conclusion that they have identified similar features. Next, the correlated variables ($|corr| \geq 70\%$) were simply removed and a final dataset on which we ran the subsequently discussed Machine Learning models with these remaining variables was built.

Final Data Set

As mentioned above, many of the predictors overlapped between those selected by the Random Forest feature importances and the Principal Component Analysis loadings; however, we noticed slightly improved performance amongst some of the models in using the predictors selected by PCA. Thus, although we ran the models using both methodologies, we have omitted the Random Forest feature importance results for the sake of brevity.

Thus, the final feature set identified by Principal Component Analysis for use in the following Machine Learning models had the following properties: **3,126** data points, **40** (39 numerical, 1 categorical) predictors.

(Final Feature Selected Dataset: https://tinyurl.com/y8ylrvdo)

Machine Learning Models

In identifying the machine learning model with the best predictive power, the team considered three categories of Machine Learning models: *linear* (e.g. Logistic Regression), *non-linear* (e.g. Splines), and *highly non-linear* (e.g. Random Forests).

Our reasons for considering these models was as follows:

- **Linear** — The similarity of the PCA feature importances to the Random Forest feature importances indicated that many of the selected features by Random Forest had potentially primarily linear variation. Thus, linear models had strong potential to fit the data without overfitting via excessive model complexity. Additionally, linear models tend to be the most easily interpretable for relationships.
- **Non-Linear** — Similarly to linear models, the non-linear models we considered (i.e. Polynomials and Splines) have strong interpretability (though slightly worse than linear models), while still being able to account for additional non-linear variation in the response. However, the increase in flexibility can increase the risk of overfitting.
- **Highly Non-Linear** — While the highly non-linear models have little-to-no interpretability (other than the feature importances of the Random Forest that were discussed above), their flexibility does allow us to get a better understanding of the true decision boundary between the two classes under consideration, a better understanding of the upper bound for prediction accuracy, and (as a result) a better understanding of what percentage of the variation in the response is truly accounted for by the predictors.

Given these considerations, the team sought out the best model that would minimize the *classification error for the test set*, and, in so doing, find the optimal point of tradeoff between bias and variance. Here, our "goal" (response) was to predict for each county whether their median income was above (1) or below (0) the national average based upon the 40 predictors identified by PCA.

Cross Validation Process

For each of the non-linear models (as well as KNN), we used a process of 10-fold cross-validation and, ultimately, selected the model that minimized the misclassification error (MCE) averaged across each of the 10 folds. Ultimately, the best model that was selected was evaluated on the test data, which resulted in our final MCE for comparison models. This testing MCE was our final basis for the selection of the best, recommended model — Random Forests.

Linear Models

LDA assumes that the distribution of each class is Gaussian with the same (constant) covariance given each class. It is not robust against outliers. From the analysis below, training errors are below 25% and testing errors lie within the 25% borderline. This means that LDA fits the response fairly well and that the population may very well be in line with the normality assumption. LDA performed the best amongst the linear models considered.

QDA is similar to LDA and allows for different covariance matrices between classes. In our actual application of QDA, we were forced to remove additional predictors that suffered from too much collinearity and caused the algorithm to become numerically unstable. Thus, while in theory QDA should always perform better than LDA on the training data due to its additional model flexibility, here, it performed worse since it had fewer of the most important features available for use in prediction.

Naive Bayes assumes that the predictors are independent and that they follow a Gaussian distribution (in the case of non-categorical predictors). While this model has strong underlying assumptions, it actually tends to perform well on examples where the number of predictors is very large. From the observed result, the training and testing errors are comparable to those of LDA and QDA.

Logistic Regression does not assume a Gaussian distribution for each class in order to preserve linearity. While this model is robust against outliers, it tends to perform worse than LDA / QDA when the covariance matrices given the labels are actually normally distributed. The result below shows stable training and testing errors under the logistic regression framework. The misclassification errors for predicting the response are comparable to the results from the previous models.

above_average_median	Train	Test
LDA	0.1471664	0.1268657
QDA	0.2733090	0.2761194
Naive Bayes	0.1832724	0.1876333
Logistic Reg.	0.1997258	0.1908316

Non Linear Models

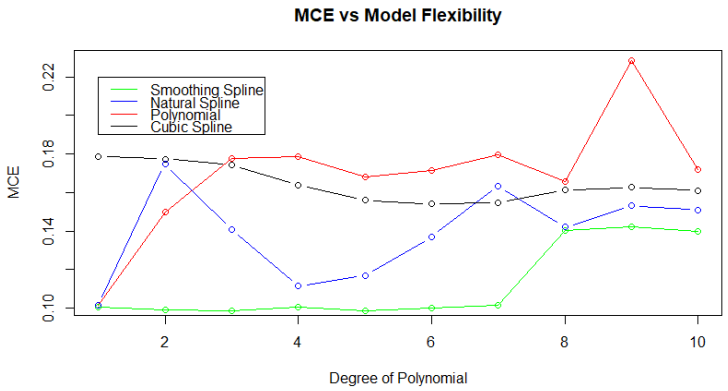
Polynomial models were fit using GLM with the parameter *POLY* in which a series of different degrees of freedom from 1 to 10 were used on the training data. Further, 10 fold CV was applied to find the model with the lowest MCE as presented below.

Cubic Splines were fit using GLM with the parameter *BS* in which a series of different degrees of freedom from 1 to 10 were used on the training data. Further, 10 fold CV was applied to find model with the lowest MCE as presented below.

Natural Splines were fit using GLM with the parameter *NS* in which a series of different degrees of freedom from 1 to 10 were used on the training data. Further, 10 fold CV was applied to find model with the lowest MCE as presented below.

Smoothing Splines were fit using GAM with the parameter *S* in which a series of different degrees of freedom from 1 to 10 were used on the training data. Further, 10 fold CV was applied to find model with the lowest MCE as presented below.

A resulting plot of cross-validation errors is shown here:



Comparing the above models based on their values with the degrees of freedom that minimized their cross-validation errors, we found the following table:

above_average_median	Best DF	Test
Polynomial	1	0.1279318
Cubic Spline	6	0.163113
Natural Spline	1	0.1215352
Smoothing Spline	5	0.1140725

Highly Non Linear Models

KNN predicts each newly observed data point based on a majority vote of the K observed training values that were "closest" to a newly observed data point. This model is fit by finding the appropriate number of neighbors, K, that minimizes the cross-validation error. While we generated a CV / training error plot, for the sake of brevity, we did not include it here and simply present the optimal model that minimized complexity by simultaneously minimizing the CV error, and with it, the bias-variance tradeoff. While the interpretability of a KNN model is poor, we ran this algorithm to identify the upper bound for possible prediction accuracy.

above_average_median	Best K	Test Error
KNN	8	0.1289979

Random Forest models successively build a number of classification trees based on bootstrapped samples from the training data and, at each split, only consider a subset of the predictors available to split upon. Then, the resulting set of trees conducts classification by majority vote. As a result, Random Forests can reduce the variance introduced by Bootstrap Aggregation since only considering a subset of the features at each split allows the model to de-correlate the trees, making them less susceptible to overly relying on one single predictor. One potential drawback to Random Forests is that they can only learn about possible split criteria based on parameter values that have been observed in the training set, making it hard for them to learn about features that fall outside of this range. This was part of the reason that we decided to use percentage features rather than population counts since with a fixed range parameter from 0 to 1, Random Forests will be able to better learn from apples-to-apples comparisons.

While we conducted hyper-parameter tuning on Random Forests by hand, due to the lack of interpretability of their structure, we do not detail the results here, but simply report the testing Misclassification Error of the optimal model for comparison.

above_average_median	Test Error
Random Forest	0.05758684

Neural Networks are built on the interspersing of linear transformations of features with non-linear activation functions. While they do not provide the opportunity for direct interpretability, we considered them here simply for the sake of trying to identify the upper bound on our prediction capabilities.

Again, due to their lack of interpretability, although we conducted a pseudo-cross-validation process by hand, we simply report here the testing error of the optimal model for comparison.

above_average_median	Test Error
Neural Network	0.1108742

Conclusions

Based on the above analysis, we were able to make the following insights and conclusions.

Data Insights

As discussed in Feature Importances, the correlations between many of the most important Random Forest features with the response revealed interesting insights. Many of these results were obvious (such as employment rate being positively associated with the absence of poverty), however, there were some interesting results that have potential implications for government policy. Here, are the insights we drew from those results:

Positively correlated predictors with absence of poverty: Private health insurance coverage of the overall population and of those out of the labor force. Percent working in the civilian sector (not in the Armed Forces). Count of wage earners. Percent working in management / business / science / arts. Percent working in the waste industry.

Negatively correlated predictors with absence of poverty: Count of those on food stamps. Count of those on supplemental security income. Percent out of the labor force. Percent on public health insurance. Percent out of the labor force and on public health insurance. Count on social security. Percent working in production, transportation, and materials.

Noticeably absent features from the set of most important features: Primary Industry (other than waste). This is surprising since locations with many workers in Finance & Technology are generally assumed to be higher income locations such as San Francisco, Chicago, and New York.

Based on these results, it is likely that a lot of the proposed social welfare programs that are designed to assist the poor, do in fact help these underprivileged communities: such as Medicare for All, Veteran's Benefits, Food Stamps, and Labor Laws that support Blue Collar workers.

Machine Learning Insights

Based on the above results, we found that the Random Forest model performs best on the data (which lends credibility to the insights obtained from the feature importances), followed by the Neural Network, Smoothing Spline, and LDA. Respectively, we observed these specific results on the testing set.

Best Models by Category	Test Accuracy
Random Forest	94.20 %
Neural Network	88.91 %
Smoothing Splines	88.60 %
LDA	87.30 %

Random Forest likely performed particularly well, somewhat due to the restriction of most variables to the interval [0,1] which allows the RF classifier to identify patterns between counties that might not otherwise be obvious via an apples-to-apples comparison as described above.

Further, these results are particularly interesting because it seems to indicate that while the decision boundary is highly non-linear, as shown by the strong performance of the Random Forest / Neural Network / Smoothing Spline, the somewhat strong performance of LDA implies that the assumptions of LDA are actually not that far from the true underlying pattern of the data.

Thus, it is likely that given the class labels, a fair number of the predictors follow an approximately multivariate normal distribution; however, in order to be confident in this result, we would have to do further investigation.

Future Research

With more time, we would have liked to explore these LDA results more thoroughly to verify the multivariate normality of the predictors given the labels. Additionally, we would have liked to add regional factors into the prediction process such as the categorical variable of state and the latitude / longitude coordinates of counties to see if physical location had an association with community income.

Additionally, we would have liked to compare the results for all of the models run on the PCA-selected most important features versus the full predictor space in order to ensure the sufficiency of the PCA features in capturing the variance of the response (although we did see that the Random Forest specifically only showed less than a 0.7% improvement on the full predictor space compared to the PCA-selected model, indicating that PCA was at least somewhat sufficient).

Finally, we would have liked to try using interaction terms in some of the linear and non-linear models which would could have allowed us to observe the effect of the absolute counts of predictors on the models rather than simply using the percentage bases of these values.

We hope you enjoyed our report on the prediction of United States incomes by county. Thank you all so much for the course — we thoroughly enjoyed learning from you all, and hope you are doing well and staying safe.

– Marco, John, Martin, and Josh