Analysis of County Poverty Rates

January 2018

Executive Summary

This project focused on the analysis of demographic, regional, health, and economic data for counties in the United States as well as their associated poverty rates. In the United States, poverty rate is defined as the percentage of individuals in a county having an income below the federal poverty threshold. In 2017, this threshold was an income of $12,600 for an individual and $24,600 for a four-person household. The ultimate objective is to use these data to estimate the poverty rates in counties where the rate is currently unknown.

The dataset contained 3,198 samples and 33 features at the county level, describing the economic and demographic makeup of the county as well as its poverty rate. The data were explored initially with the creation of descriptive and summary statistics to gain an initial understanding of the profile of the dataset, as well as identify areas where missing data might be a problem. This first pass was followed up by a more robust visual and statistical analysis of the data to explore the relationships between the features. To expand on the analysis and create a tool that can analyze further counties, a regression model was developed to predict a county's poverty rate based on its various features.

While the analysis yielded a variety of features which can be used to predict the poverty rate in a county, several strongly predictive numerical features were discovered. These fell especially across economic, demographic features:

Civilian Labor Percentage (Pearson Correlation to Poverty Rate 0.67): An economic indicator. The county's civilian labor force as a percentage of the population.

Percent Unemployment (Corr 0.59): A economic indicator. The annual percentage of people in the county who unemployed.

Percent Uninsured Adults (Corr 0.54): An economic indicator. The percentage of adults in a county who do not have insurance.

Percent of Adults Without a High School Diploma (Corr 0.68): A demographic indicator. The percentage of adults in the county who lack a completed high school diploma.

Additionally, two categorical features were determined to be predictive of the poverty rate in a county:

Rural-Urban Continuum Code: A USDA regional descriptor that classifies metropolitan counties by the population size of their metro area and non-metropolitan counties by the degree of urbanization and adjacency to a metro area.

Economic Typology: A USDA categorical economic descriptor that classifies all US counties into six mutually exclusive categories of primary economic drivers (farming, mining, government etc.) and six overlapping policy relevant themes (low education, low employment, retirement destination etc.).

From the exploration and modeling of the data, we can conclude that efforts to improve county high school graduation rates and increase civilian labor percentage may be successful in decreasing poverty rates.

Data Exploration

Descriptive Statistics

The first step in the analysis of the dataset began with the creation of a table of summary and descriptive statistics for the numeric features. The minimum, maximum, mean, median, range, standard deviation, and sample count were calculated across 3,198 samples. While descriptive statistics do not tell the entire story, they are a very useful first step in understanding the features of a dataset.

Features	Minimum	Maximum	Mean	Median	Range	Standard Deviation	Count
Air Pollution Particulate Matter	7.0000	15.0000	11.6262	12.0000	8.0000	1.5523	3,042
Birth Rate Per 1K	3.0000	30.0000	11.6568	11.0000	27.0000	2.6043	3,080
Death Rate Per 1K	1.0000	22.0000	10.1357	10.0000	21.0000	2.6694	3,080
Homicides Per 100K	0.0200	31.4000	5.6015	4.5750	31.3800	4.3861	1,250
Motor Vehicle Crash Deaths Per 100K	2.8000	96.6300	20.4257	18.4000	93.8300	9.9842	2,735
Pct Adult Obesity	0.1170	0.4850	0.3048	0.3070	0.3680	0.0430	3,072
Pct Adult Smoking	0.0280	0.4910	0.2117	0.2050	0.4630	0.0634	2,678
Pct Adults Bachelors Or Higher	0.0655	0.7280	0.2091	0.1873	0.6625	0.0912	3,080
Pct Adults Less Than A High School Diploma	0.0142	0.5413	0.1422	0.1284	0.5271	0.0648	3,080
Pct Adults With High School Diploma	0.0810	0.5267	0.3445	0.3474	0.4457	0.0710	3,080
Pct Adults With Some College	0.1190	0.4807	0.3041	0.3040	0.3617	0.0516	3,080
Pct Aged 65 Years And Older	0.0270	0.5130	0.1690	0.1660	0.4860	0.0438	3,072
Pct American Indian Or Alaskan Native	0.0000	0.8710	0.0185	0.0070	0.8710	0.0570	3,072
Pct Asian	0.0000	0.4450	0.0136	0.0080	0.4450	0.0278	3,072
Pct Below 18 Years Of Age	0.0760	0.4110	0.2277	0.2280	0.3350	0.0332	3,072
Pct Civilian Labor	0.1850	0.9920	0.4745	0.4760	0.8070	0.0696	3,080
Pct Diabetes	0.0290	0.2150	0.1076	0.1070	0.1860	0.0237	3,072
Pct Excessive Drinking	0.0310	0.5630	0.1652	0.1610	0.5320	0.0533	2,220
Pct Female	0.3220	0.5550	0.5006	0.5040	0.2330	0.0213	3,072
Pct Hispanic	0.0000	0.9690	0.0832	0.0370	0.9690	0.1235	3,072
Pct Low Birthweight	0.0220	0.1840	0.0817	0.0790	0.1620	0.0211	2,933
Pct Non Hispanic African American	0.0000	0.8110	0.0866	0.0200	0.8110	0.1395	3,072
Pct Non Hispanic White	0.0230	0.9940	0.7862	0.8500	0.9710	0.1866	3,072
Pct Physical Inacticity	0.0900	0.4510	0.2713	0.2740	0.3610	0.0540	3,072
Pct Unemployment	0.0110	0.2300	0.0582	0.0560	0.2190	0.0214	3,080
Pct Uninsured Adults	0.0360	0.5250	0.2130	0.2110	0.4890	0.0646	3,072
Pct Uninsured Children	0.0050	0.3220	0.0841	0.0755	0.3170	0.0388	3,072
Pop Per Dentist	269.0000	27,249.0000	3,309.0000	2,589.0000	26,980.0000	2,606.0000	2,882
Pop Per Primary Care Physician	340.0000	20,940.0000	2,580.0000	1,910.0000	20,600.0000	2,171.0000	2,896
Poverty Rate	2.5000	47.4000	16.8171	15.8000	44.9000	6.6980	3,198

Figure 1: Descriptive Statistics of Numeric Features

Notably, the counts for all the features, excluding Poverty Rate, are lower than 3,198. This indicates that there are missing vales for some of the county samples which will impact the analysis and modeling of the data. For some features, especially Motor Vehicle Crash Deaths Per 100k with only 1,250 samples, the amount of missing data is significant. Poverty Rate is the key field in the analysis and we can observe that its mean is greater than the median, and that the standard deviation is significant compared to

those values. A visualization of the distribution of the Poverty Rate confirms that the data are right skewed and that there is significant variance. While most counties in the dataset have poverty rates in the teens or low twenties, a visible minority have significantly higher rates.

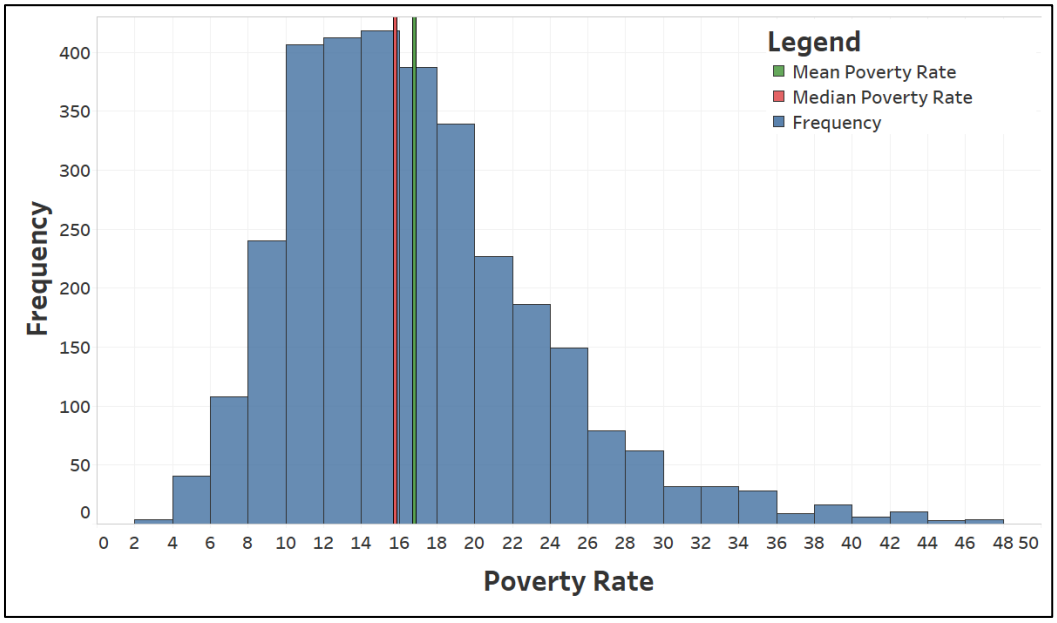

Figure 2: Histogram of Poverty Rate (Mean in Green, Median in Red)

In addition to the numeric features in the dataset there are four categorical features in the dataset:

- **Urban Influence –** The primary urban influence on a county ranging from 'Large-in a metro area with at least 1 million residents or more' to 'Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents.'
- **RUCC –** The Rural Urban Continuum Code that distinguishes metropolitan counties based on their population size and non-metropolitan counties by their level of urbanization and proximity to a metropolitan area. Ranges from 'Metro - Counties in metro areas of 1 million population or more' to 'Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area'.
- **Economic Typology -** Classifies all US counties into six mutually exclusive categories of primary economic drivers (farming, mining, manufacturing, Federal/State government, recreation, and nonspecialized) and six overlapping policy relevant themes (low education, low employment, persistent poverty, persistent child poverty, population loss, and retirement destination).
- **Year –** Contains the values "a" and "b". Typically, year would be an ordinal (ordered) feature, however in this dataset the values only differentiate between two years and from the description of the encodings "a" and "b" cannot be confidently used to place the years in chronological order.

To understand the character of the categorical features, exploratory visualizations were created. These suggested the following relationships:

- Roughly two thirds of counties in the dataset are in Non-Metro RUCC areas.
- Of those Non-Metro RUCC counties, most have an urban population of 2,500-19,999 and are adjacent to a metro area.
- The least common county by RUCC is ‘Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area’.
- The most common Economic Typology for counties is ‘Non-Specialized’ (40% of counties), the least common is ‘Mining Dependent’ (8%).
- The counties are evenly split between years ‘a’ and ‘b’.
- The least common type of Urban Influence is ‘Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents’.

All frequencies for the different members of the categorical features were distributed without significant outliers, especially for members with lower frequency counts. As such it was decided to maintain the members as they existed in the dataset, without combining any to bin the features.

Correlation and Feature Relationships

The next portion of the analysis focused on exploring the nature and strength of the relationships between the features as well as their relationship to the target variable: **poverty rate**.

Numerical Relationships

Since there are twenty-nine numerical features in addition to poverty rate, plotting all of them together to understand the nature of the relationship would be difficult to interpret. The six features with the strongest Pearson Correlation coefficient to poverty rate are presented to give an understanding of the nature of the relationships between those features and to poverty rate.

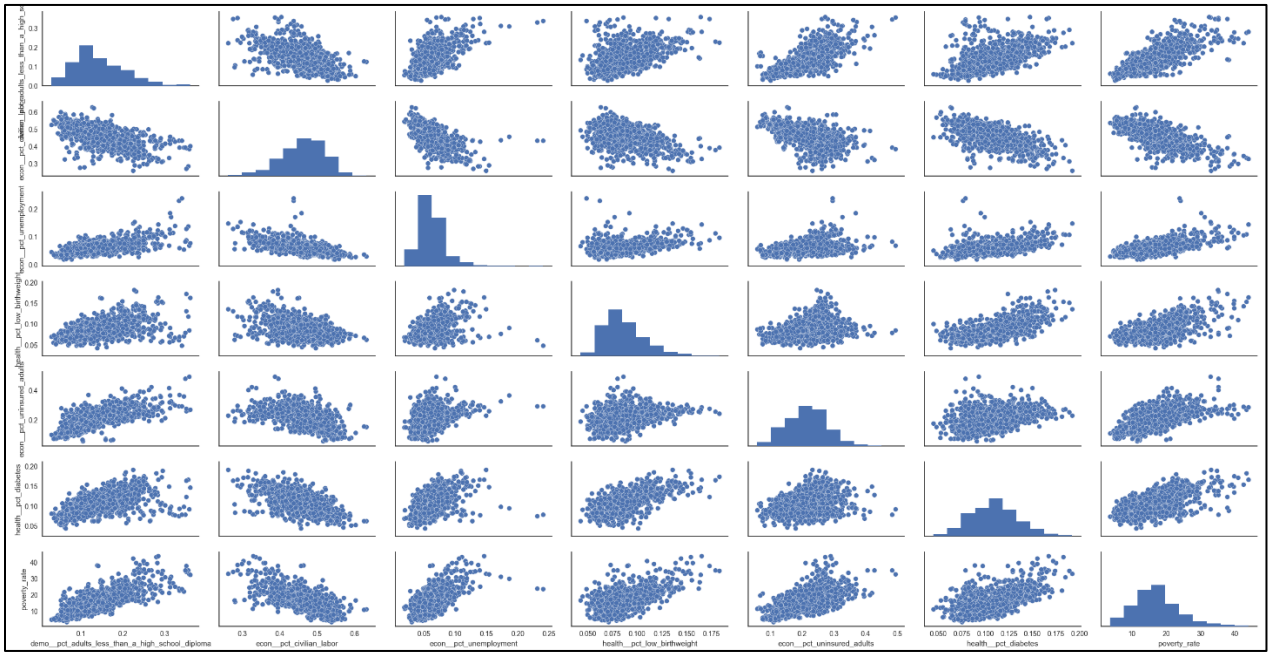

Figure 3: Pairplot of the Top Six Features by Correlation with Poverty Rate

Examining the plots of each feature against poverty rate in the furthest right column, each feature has an apparently linear relationship with poverty rate.

The next step in exploring the relationships between the numerical features of the dataset, as well as poverty rate, was to construct a correlation matrix. The matrix encodes the magnitude of the Pearson Correlation coefficient between features as a color saturation in each cell and the direction of the correlation as hue, with red for positive and teal for negative. From the matrix it is immediately clear that there are some strongly correlated features such as 'birth_rate_per_1k' to 'pct_below_18_years_of_age' and 'pct_uninsured_adults' to 'pct_uninsured_children'. Additionally, some features including 'pct_adults_less_than_a_high_school_diploma' and 'pct_civilian_labor' are noticeably correlated to the 'poverty_rate'.

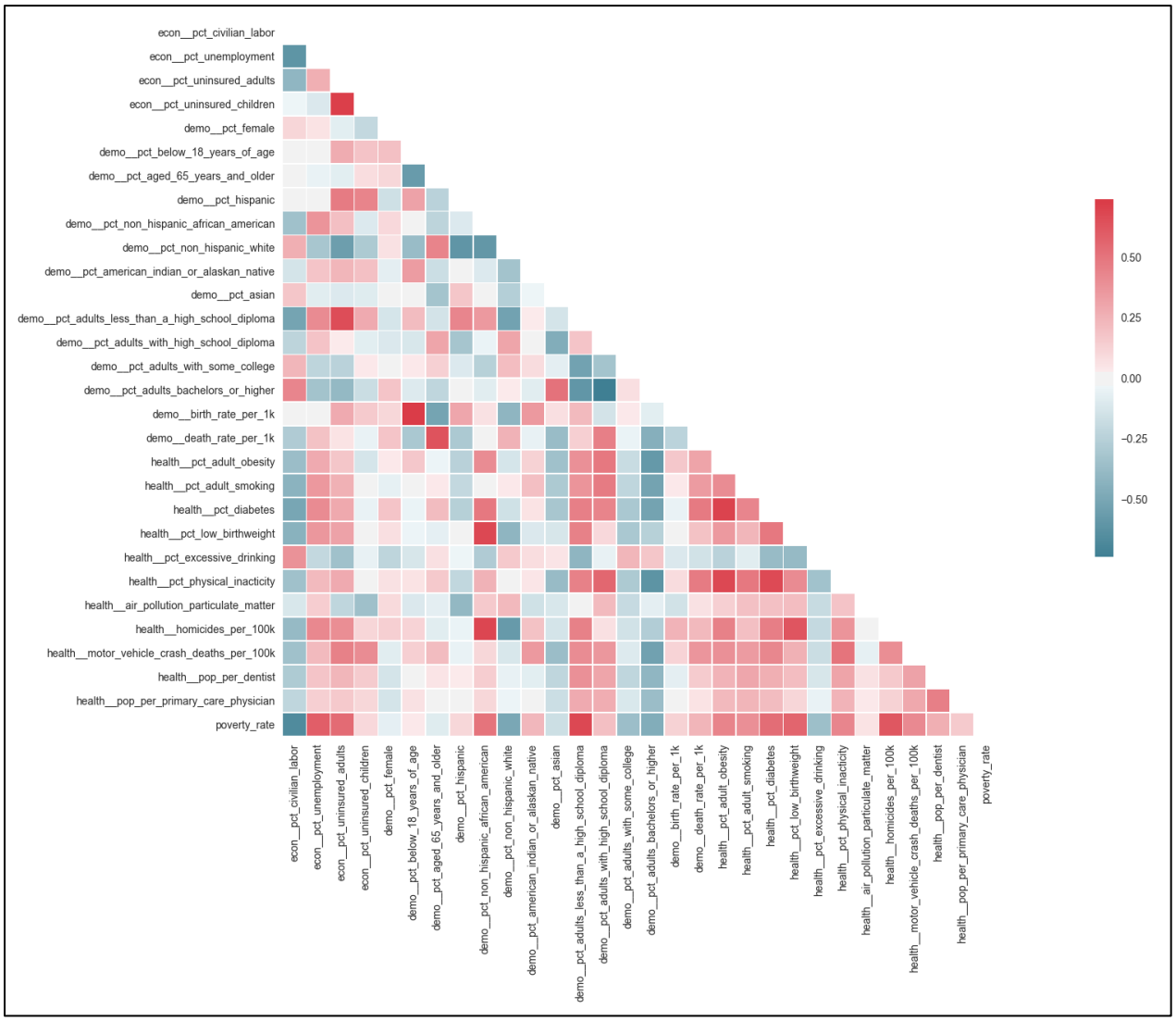

Figure 4: Correlation Matrix of Numerical Features

Interestingly, among the features that might be considered stress induced vices or conditions, all are positively correlated with the poverty except for the rate of excessive drinking in a county which has a negative correlation:

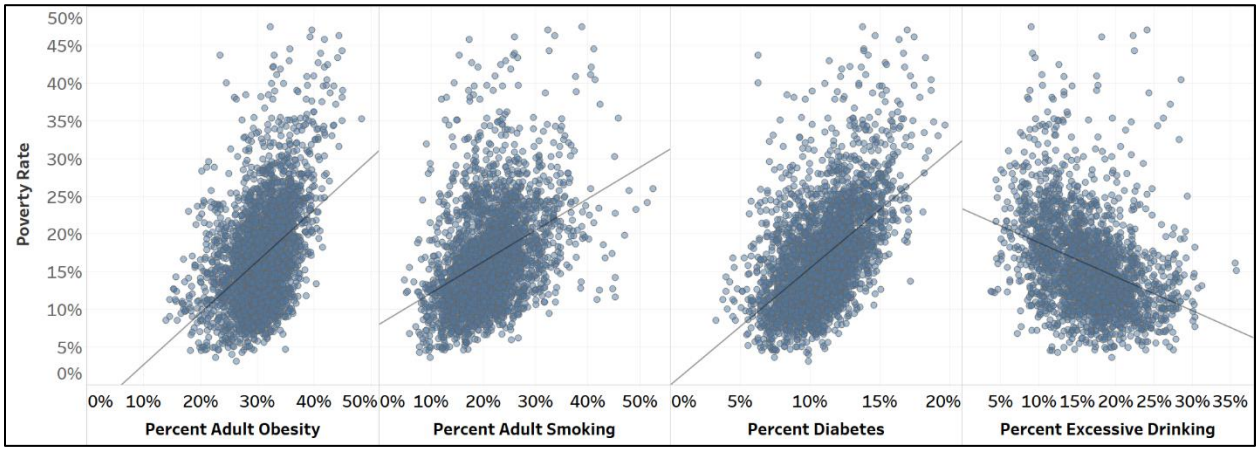

Figure 5: Selected Vices/Conditions Correlated to Poverty Rate

Categorical Relationships

The analysis of the relationships between the numerical features yielded important information on the interactions between them and their relationship to poverty rate. To expand on the numerical analysis, the relationship between the categorical features and poverty rate was explored using boxplots:

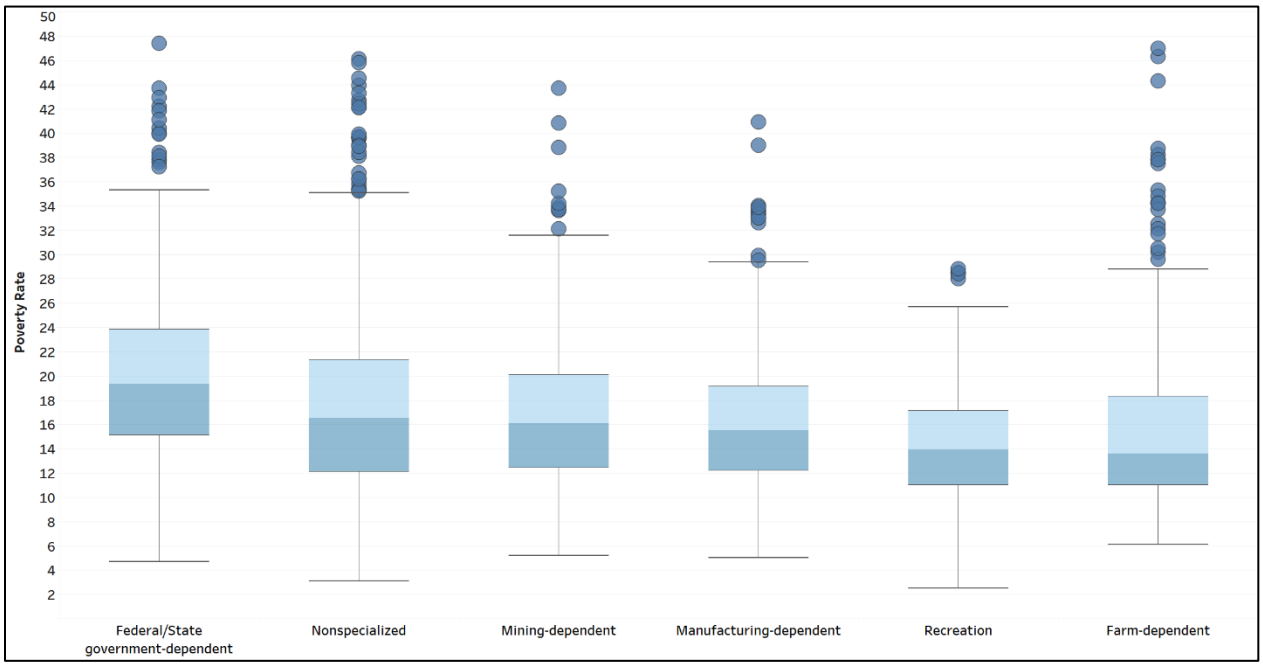

Figure 6: Poverty Rate by Economic Typography

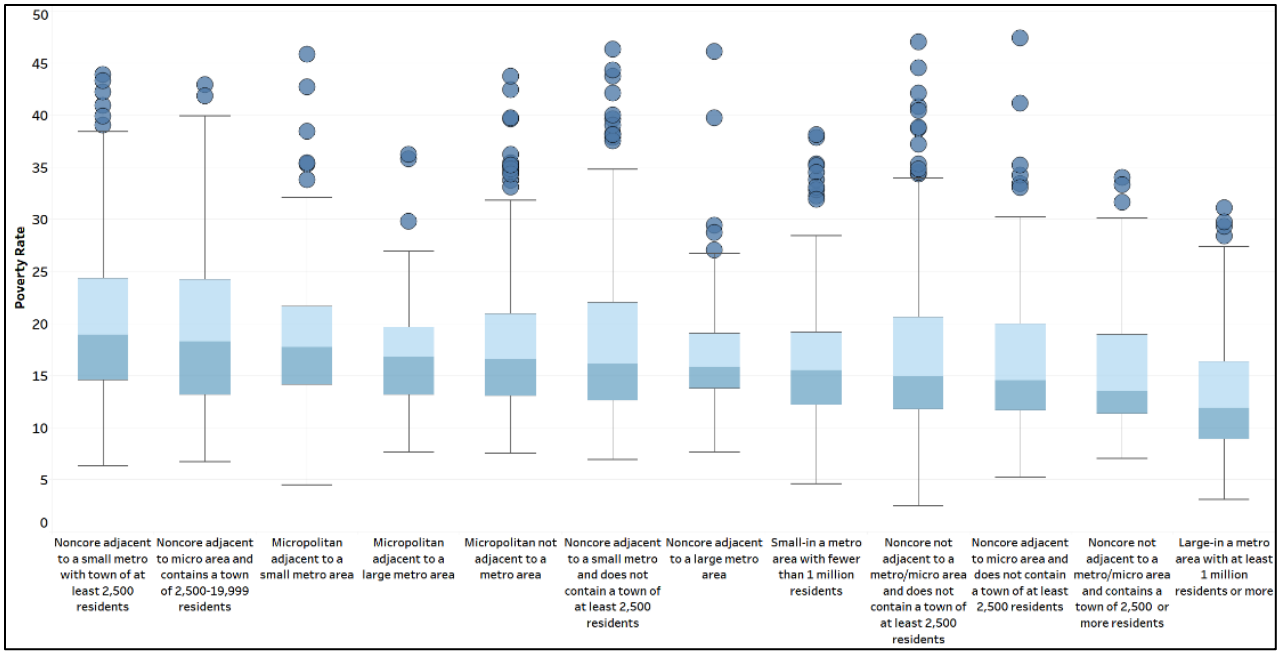

Figure 7: Poverty Rate by Urban Influence

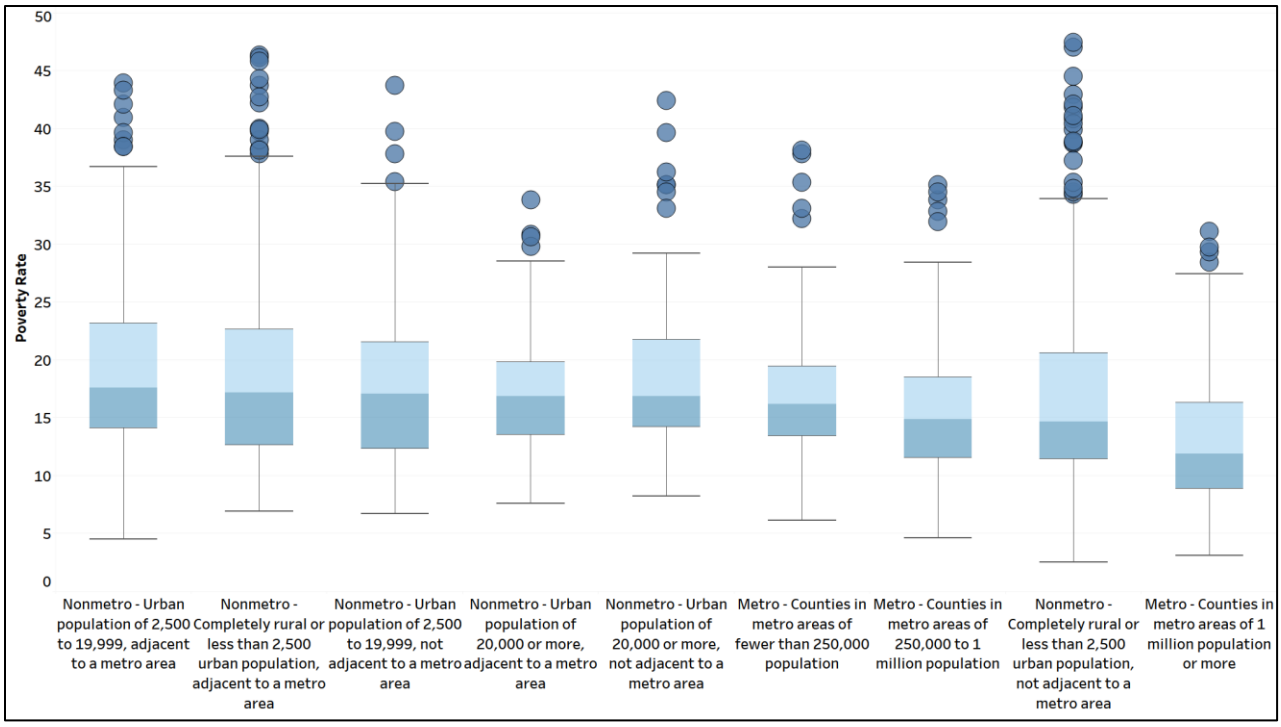

Figure 8: Poverty Rate by RUCC

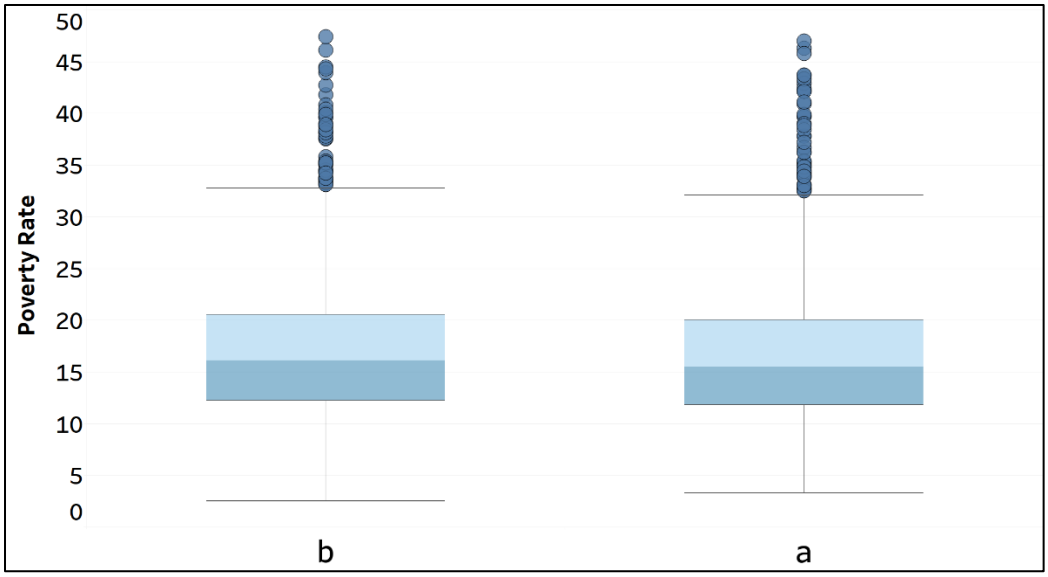

Figure 9: Poverty Rate by Year

For the categorical features Economic Typography, Urban Influence, and RUCC there is a clear variance of median poverty rates between their members, indicating that they can be predictive of poverty rate. This visual indication was supported by a single factor ANOVA of each feature with respect to poverty rate that yielded p-values below 0.01 for all three. However, for Year the medians between years 'a' and 'b', 15.5 and 16.1, were visually indistinguishable. To further explore this feature, and determine whether there was a significant difference between the poverty rates in each year the sample variance in each year was calculated and a two-tailed t-test was performed. The resulting t-statistic was -2.31 with a p-value of 0.02 allowing for the rejection of the null-hypothesis that there was no difference in the mean poverty rates of the two years. As a result, Year was kept in the dataset as a feature of interest for further exploration and modeling.

Additionally, the several observations were made based on the distributions of poverty rate values within the members of the categorical features including:

- Urban Influence, though showing clear differences in the distributions between its members, does not have a clear pattern in median poverty rates between metro/micropolitan and noncore areas.
- For Urban Influence, areas that are adjacent to or in large urban areas have tighter distributions than other areas that are not in or adjacent to or in densely populated areas.
- Counties with an RUCC code that started with Non-Metro typically had higher median poverty rates than those with a Metro prefix.
- Counties with a farm dependent economic typology had the lowest median poverty rate. However, they also had a wide distribution of values that stretched into some of the highest poverty rates in the dataset.
- Though counties primarily dependent on recreation in their economic typography did not have the lowest median poverty rate, they had a much tighter distribution

To further understand the distribution of poverty rates among the members of the Economic Typology, a second distribution visualization was created using histograms broken out by the different typologies:

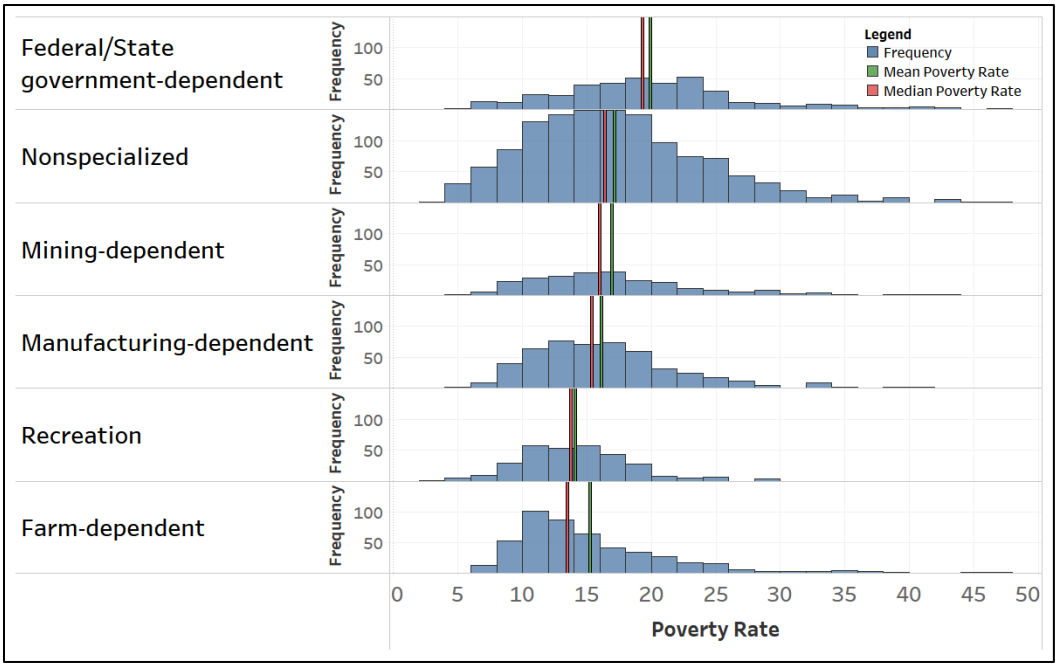

Figure 10: Distribution of Poverty Rates by Economic Typography

From this visualization it can clearly be observed that although 'Farm-dependent' counties had the lowest median poverty rate, the distribution is highly right skewed, with a mean that is significantly larger than the median. This contrasts with 'Recreation' focused counties that have a more normal distribution of poverty rates.

Multi-Dimensional Relationships

Though there are clear relationships between poverty rate and a variety of individual features in the dataset, this may not offer a complete picture of the way that the features drive an outcome. Multiple features may interact together to produce effects that are not observable when looking at the features separately. Some selected multi-variate interactions are recorded below:

Figure 11: Cities with a Metro RUCC Code Have Tighter, Less Extreme Distributions Across Features Strongly Correlated with Poverty Rate

To simplify analysis, counties were broken down into groups based on those in Metro and Non-Metro areas using the RUCC classification. When plotted with the 7 features most correlated to poverty rate, the Metro counties showed a tighter grouping in the correlated feature and less extreme poverty rates. One notable exception to this trend is in homicides per 100k, where the extreme high outliers in this feature are all Metro counties.

As a single feature, Economic Typology's members have clearly differentiated median poverty rates. However, each member except for Recreation shows a high degree of variance with a right skew. Some of this variance is explained when poverty rate is examined through a combination of a county's Economic Typology, Urban Influence and Metro or Nonmetro RUCC codes. The two typologies that exhibit the highest median poverty rate, 'Federal/State government-dependent' and 'Nonspecialized', show a wide range of median poverty rates when broken down by different Urban Influences. The Urban Influence code 'Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents' has poverty rate values that are strongly dependent on Urban Typology. Counties of this type in 'Federal/State government-dependent' economic areas tend to have among the highest poverty rates, while those in 'Farm-dependent' and 'Recreation' areas are among the lowest.

	Metro		Nonmetro									
	Large-in a metro area with at least 1 million residents or more	Small-in a metro area with fewer than 1 million residents	Micropolitan adjacent to a large metro area	Micropolitan adjacent to a small metro area	Micropolitan not adjacent to a metro area	Noncore adjacent to a large metro area	Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents	Noncore adjacent to a small metro with town of at least 2,500 residents	Noncore adjacent to micro area and contains a town of 2,500-19,999 residents	Noncore adjacent to micro area and does not contain a town of at least 2,500 residents	Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents	Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents
Farm-dependent	14.80%	12.35%		15.55%	14.20%	15.95%	13.45%	19.90%	14.40%	13.40%	12.35%	12.95%
Federal/State government-dependent	15.70%	17.25%	19.00%	21.45%	19.20%	15.75%	20.10%	23.35%	22.35%	19.30%	21.10%	22.40%
Manufacturing-dependent	11.90%	15.50%	15.10%	14.10%	14.30%	15.40%	19.00%	17.75%	19.90%	21.25%	18.65%	15.10%
Mining-dependent	14.45%	16.60%	15.65%	18.90%	16.55%	16.00%	15.35%	18.00%	13.75%	12.45%	12.90%	18.75%
Nonspecialized	12.05%	15.60%	17.55%	18.85%	19.20%	15.85%	24.60%	19.95%	23.25%	23.50%	12.85%	21.30%
Recreation	10.45%	13.30%	13.15%	15.40%	13.10%	15.15%	14.00%	14.90%	16.60%	18.10%	12.85%	13.30%

Figure 12: Median Poverty Rate is Strongly Influenced by its Economic Typology and Urban Influence

Finally, there are a range of multi-dimensional interactions between numerical features as well as those between categorical features. One example of this kind of interaction is the relationship between the percentage of adults with less than a high school diploma, the civilian labor percentage, and poverty rate. Counties with a higher percentage of adults without a HS diploma and a lower civilian labor pool as a percentage of population also have higher poverty rates.

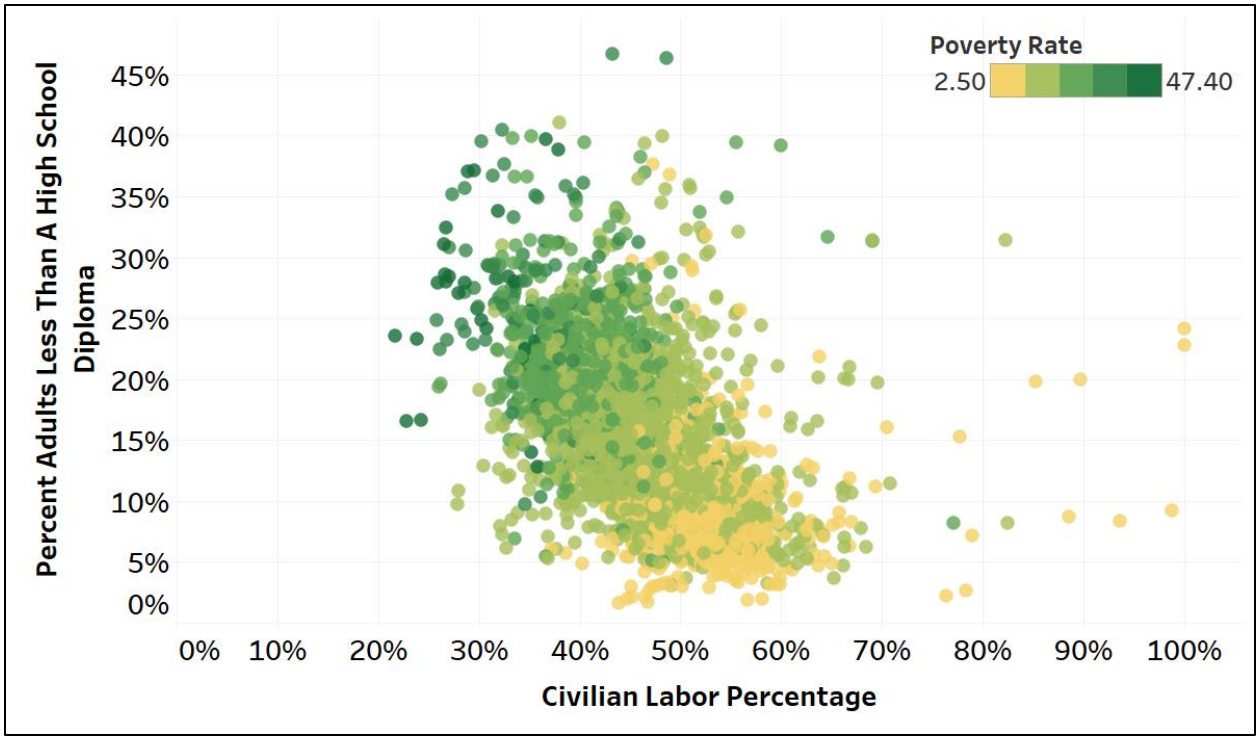

Figure 13: Poverty Rate is Influenced by the Interaction Between the Percentage of Adults Without a HS Diploma and the Civilian Labor Percentage

The fact that many of the features in the dataset are correlated with each other, and exhibit multi-dimensional relationships with poverty rate, has important implications for developing a model to predict poverty rates. Any model will need to handle or eliminate correlated features, and support interactions between features.

Prediction of County Poverty Rates

Based on the analysis of the relationships between the various features and the target variable poverty rate, a model was created to predict the poverty rate in new counties where the rate is currently unknown. This model can also serve as a tool to understand what actions might help alleviate poverty in a county.

Data Cleanup and Preparation

One of the first challenges in preparing the dataset for the model is the number of missing values in the in the 34 features in the data. Of the 3198 samples in the dataset, 2129 are missing at least one value with some missing as many as 22. Given the number of rows missing values, the brute force approach of dropping all rows with missing values would be detrimental to the analysis since it would result in 66% of the available data being lost.

To handle this issue probabilistic principal component analysis was used to fill in missing values. This technique has a significant advantage over other replacement methods including zeros and column mean. While those approaches only return a single value based on the values in each column,

probabilistic PCA models the correlations between features across the entire dataset, so interactions between features are captured in the resulting predicted values.

To prepare the data for modeling, numerical features were normalized to values between 0 and 1 using the MinMax scaling method. Finally, since the dataset contained categorical features, these features were transformed into a binary matrix, with each column representing a member of each feature using the OneHotEncoding method.

Regression

After preparing the data for analysis, several regression models were developed to analyze the best approach to modeling the data to predict the poverty rate in a county. The dataset was split into training and test sets in a 75% to 25% ratio.

As a first approach, a Linear Regression model was applied to the data. Cross validation of this model produced a relatively poor average Root Mean Squared Error value of 3.33 on the test set. Especially at larger target values there is a noticeable increase in variance from the scored labels as depicted in the scatter plot below:

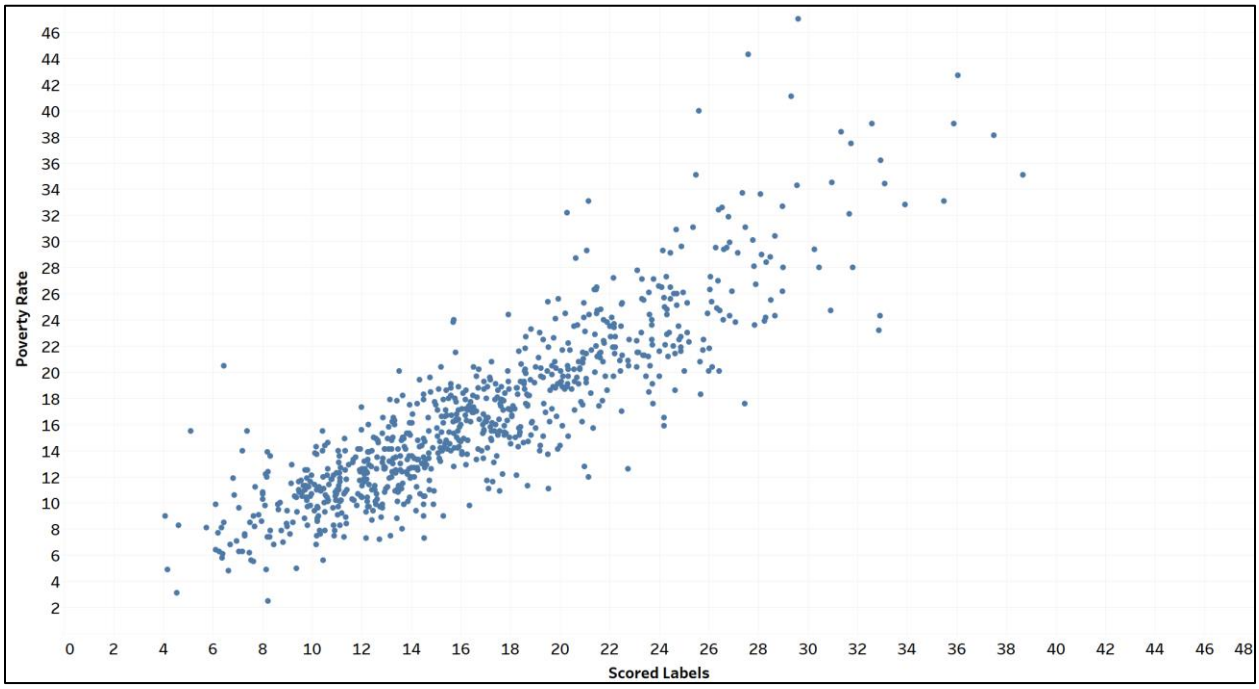

Figure 14: Results of Linear Model Compared to Actual Values

Given the strong indications of non-linear multi-dimensional feature interactions in the data exploration phase, combined with the relatively poor performance of the linear model in predicting poverty rates, it was clear that a model capable of handling non-linear relationships was required. To replace Linear Regression, a Boosted Decision Tree Regression model was selected. Boosted Trees Regression has several other advantages over traditional Linear Regression that were advantageous in this dataset. Earlier analysis had shown that multiple features were highly correlated with each other, an issue which

can complicate a linear model but to which decision trees are much less sensitive. The Boosted Trees Regressor produced a cross-validated average RMSE of 2.69 on the test set.

Comparing the scored labels from the model to the target poverty rate values, the boosted trees model produces estimates that are much closer to the true values, especially at higher and lower test poverty rates. This relationship is depicted in the scatter plot below:

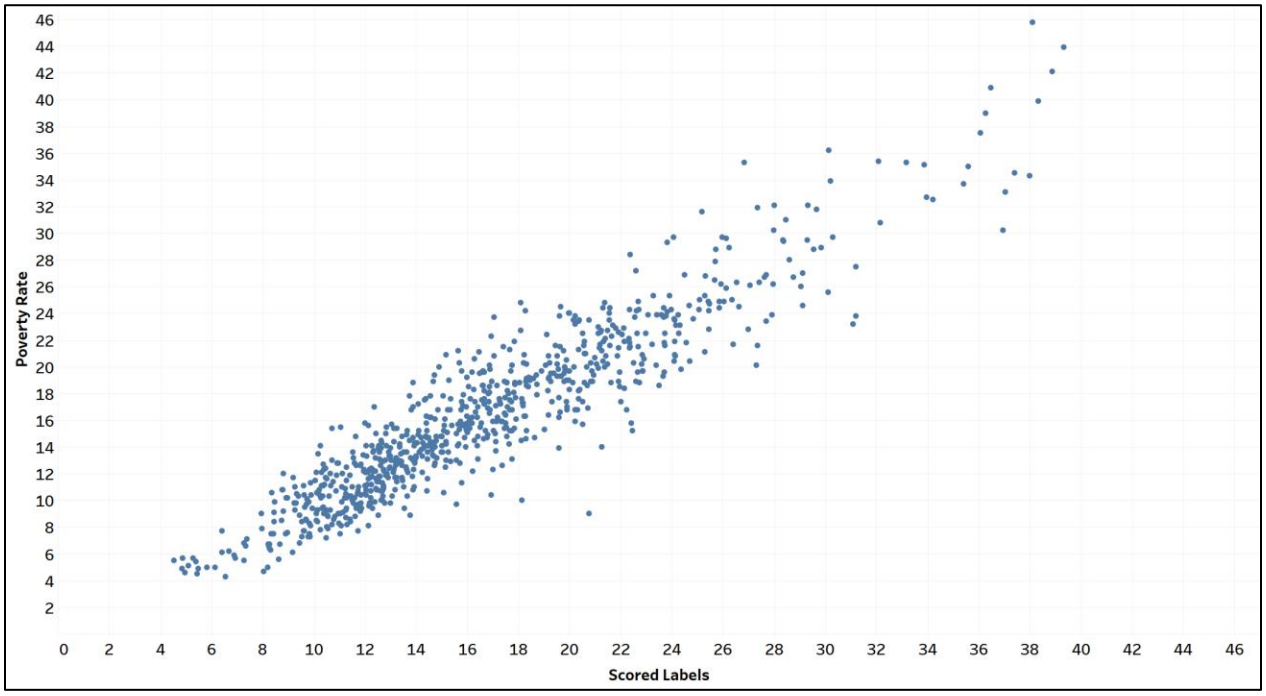

Figure 15: Results of Boosted Trees Model Compared to Actual Values

Based on this model, permutation feature importance analysis supported earlier feature strength findings with 'pct civilian labor' (1.372), 'pct adults less that a high school diploma' (0.940), and 'pct uninsured adults' (0.646) having the strongest importance scores.

Conclusion

Analysis of the economic, demographic, regional, and health factors from a county indicate these features can be confidently used to calculate the county's poverty rate. Particularly, the civilian labor force percentage, the percentage of adults without a high school diploma, the percentage of uninsured adults, and the percent unemployment rate are strongly predictive of the outcome. Additionally, RUCC codes and economic typology are strong categorical predictors that can differentiate counties by poverty rate.

The results of this analysis indicate that efforts to improve high school graduation rates may be the single most effective tool to combat poverty across counties with a range of regional and demographic profiles. At a more systemic level, ensuring that counties have a suitably large percentage of their population able to participate in the civilian labor force appears to be a strong defense against a high poverty rate.