Analysis of Poverty in United States Counties

Joe Auer, January 2018

Executive Summary

As difficult as it may be to believe, the United States - one of the richest countries in the world - has a poverty problem. According to the Brookings Institute, 46.7 million Americans (14.8% of the population) lived in poverty in 2014.

The condition of poverty is nothing new, and the people who make up this population have many different stories to tell regarding the circumstances which led to them into poverty. There are many programs, both inside and outside of government attempting to fight the problem. However, given the recent trend of rising housing prices and devastating natural disasters (e.g., hurricanes, flooding, and wildfires), the battle against poverty is not getting any easier.

The analysis presented here explores recent data at the US County level, and proposes a Machine Learning regression model to predict poverty rate. The source data file contained broad characteristics such as the rural/urban makeup, types of employment available, education levels, health factors, and ethnicity. Unfortunately, other potentially telling data regarding income, housing prices, and housing availability was not provided.

Some of the factors which largely correlated with poverty rate included the unemployment rate, low levels of education, and poor general health. The regression model trained and tested with the source data is a fairly good predictor of the poverty rate.

About the Raw Data

The custom data set consists of two recent years of US County socioeconomic statistics. It was provided by datacapstone.org and compiled from United States Department of Agriculture Economic Research Service (USDA ERS) sources.

Overall, there were 3198 observations of 35 variables.

Data Wrangling Decisions

Unfortunately, the source file was missing some data. Decisions were made to maximize the value of the data present without skewing the results based on incomplete information. If less than 15% of data for a particular feature was missing, the mean value of the available observations was substituted for the missing values. In two cases, there was simply too much data missing for those features to be helpful.

Features with Many Missing Values
Action Taken: Removed

Feature	Percent Missing
Homicides per 100,000 population	62%
Percent of adults engaging in excessive consumption of alcohol	31%

Features with a Small Percentage of Missing Values
Action Taken: Replace Missing Data with Mean Value

Feature	Percent Missing
Percent of adults who smoke	15%
Deaths by motor vehicle crash per 100,000 population	13%
Population per dentist	8%
Population per doctor	7%
Percent of babies born with low birth weight	6%

Data Exploration

This section discusses general data categories and observes correlation with the poverty rate.

1. How Important is Where You Live?

Urban or Rural Setting

One can find areas of poverty in urban centers as well as rural towns. One of the categorical features in the source data was the Rural Urban Continuum Code from the USDA Economic Research Service. This feature divides counties into nine groups based on the degree of urbanization - three Metro codes and six Nonmetro codes. Overall, nonmetro counties have a higher average poverty rate.

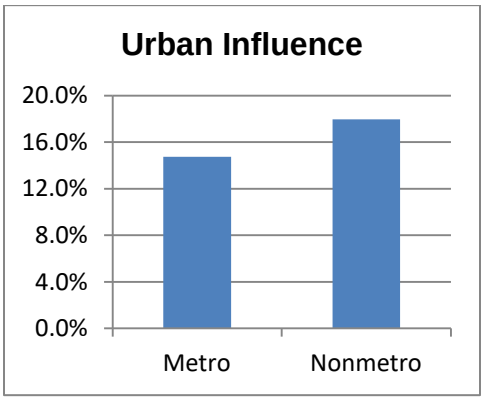

Fig 1: Poverty Rate by Rural Urban Continuum Code
Metro/Nonmetro Grouping

Access to Healthcare

The source file included two features related to the general availability of healthcare providers in the county - population per doctor and population per dentist. Looking at the correlation between these items and poverty rate, the data shows slight positive correlations. In the table below, a value close to one indicates a strong, direct relationship. A number close to zero indicates no correlation.

Feature	Correlation Value
Population per Dentist	0.27
Population per Primary Care Physician	0.16

2. How Important is the Local Job Market?

Unemployment

Since poverty is officially calculated based on income, it's not surprising that there is a strong relationship between the Unemployment Rate and the Poverty Rate. In each county, there should be some form of social safety net to assist the unemployed with alternative income, but it certainly does not sufficiently replace traditional employment income.

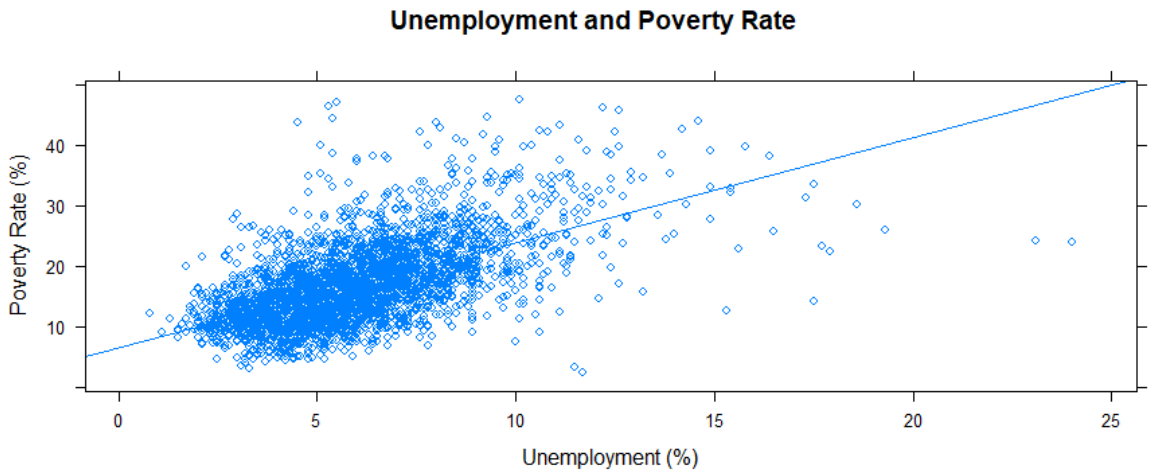

Fig 2: Poverty Rate by Unemployment Rate

Type of Work Available

Another categorical feature from the source data is the "County Topology Code", provided by the USDA Economic Research Service. It lists a county's economic dependence by industry. See Fig 3 below.

One possible explanation to the ordering in the chart could be barriers of entry for a particular type of work. For example, Recreation work may not require a high level of skills or training compared to other areas. Additionally, there may be stricter controls related to federal and state government work.

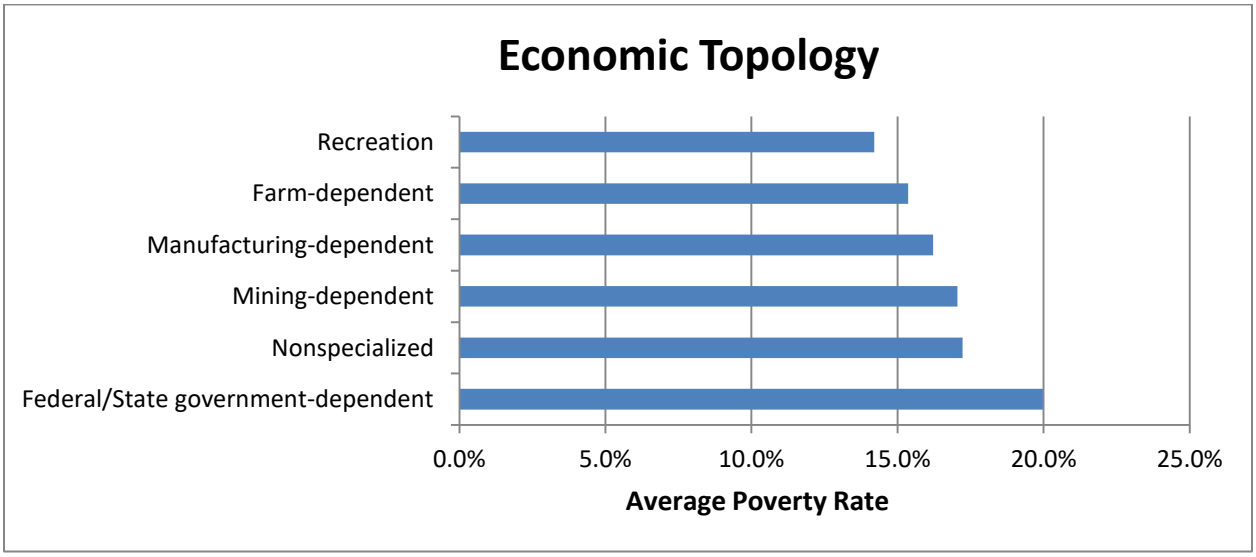

Fig 3: Poverty Rate by Economic Topology Code

Expanding on the notion that there is higher poverty for counties where the primary topology code is government work, one can also see this in the percent of residents employed in the civilian labor force. As more residents work in civilian labor positions, the poverty rate is lower.

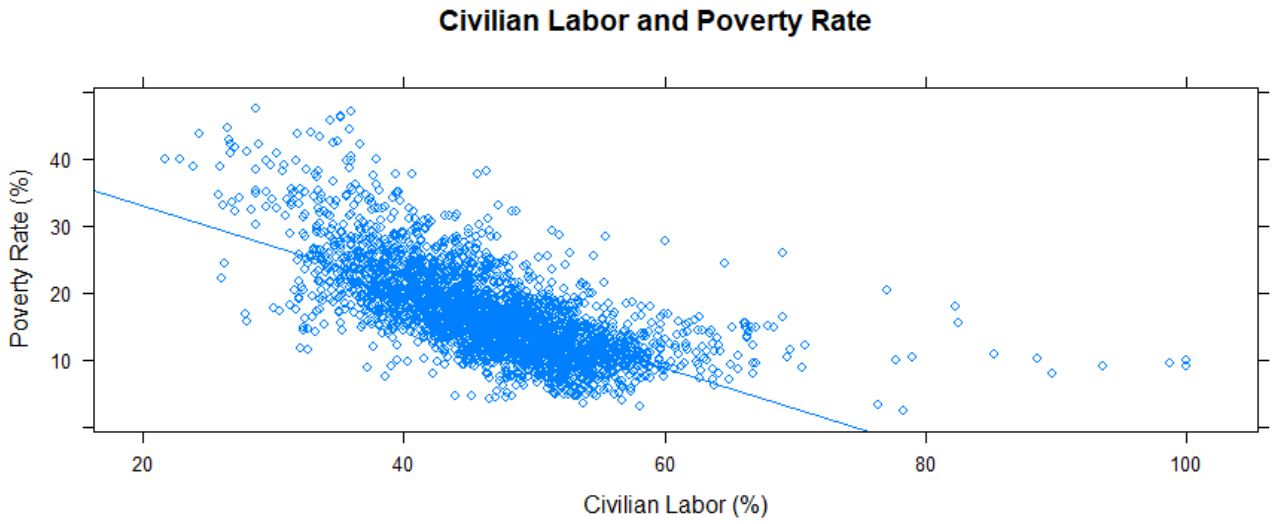

Fig 4: Poverty Rate by Civilian Labor Force

3. How Important is Demographic Makeup?

The source data also included several demographic features including gender, young and old age brackets, and ethnicity.

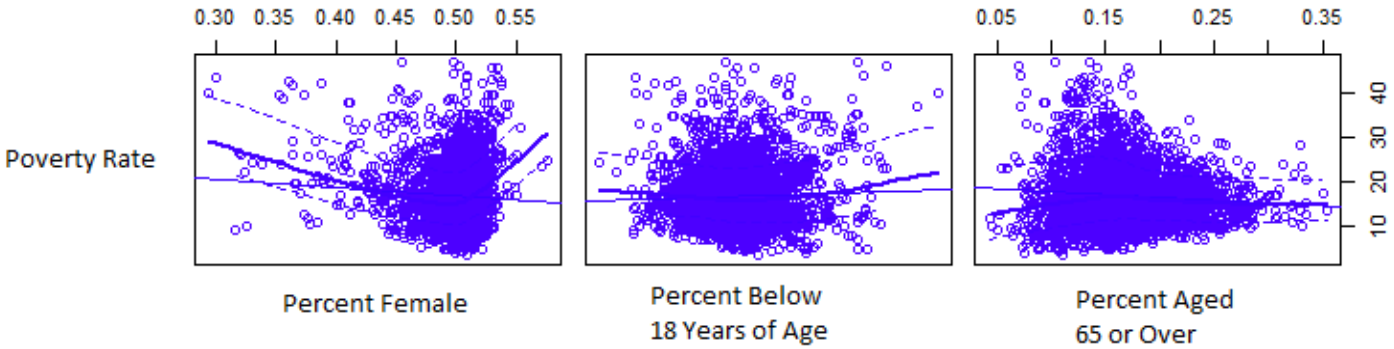

Fig 5: Poverty Rate by Gender and Age Demographics

Gender

Poverty appears to show no particular preference for male or female residents

Outer Age Buckets (Under 18, 65 and Over)

There appears to be no correlation of poverty with the percent of residents who are under the age of 18 or over the age of 65.

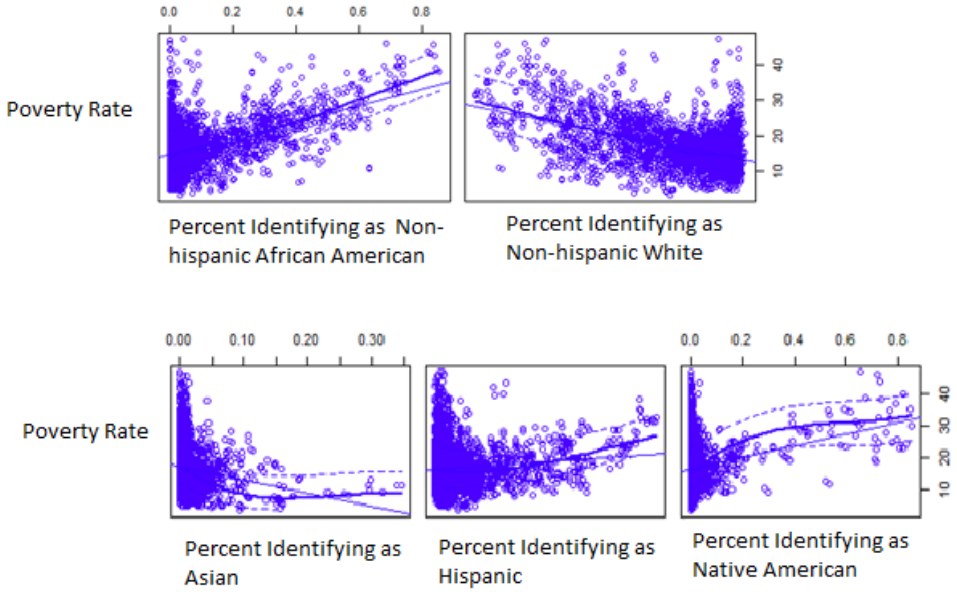

Fig 6: Poverty Rate by Ethnicity

Ethnicity

Five ethnic groups were identified in the data. There appears to be a positive correlation of poverty rate with the percentage of the population identifying as African Americans, and a negative correlation between poverty and the population identifying as Non-Hispanic White. The percent identifying as Native Americans showed a slight positive correlation, while the population identifying as Hispanic and the population identifying as Asian showed little to no relationship with the poverty rate.

4. How Important is Education?

Data on four education levels were provided. The evidence below supports the goal at the heart of the "Stay in School" campaigns. There is a strong association of poverty with those who do not have a High School diploma. There is a better chance of staying out of poverty with a High School diploma, and it improves with the amount of college attended.

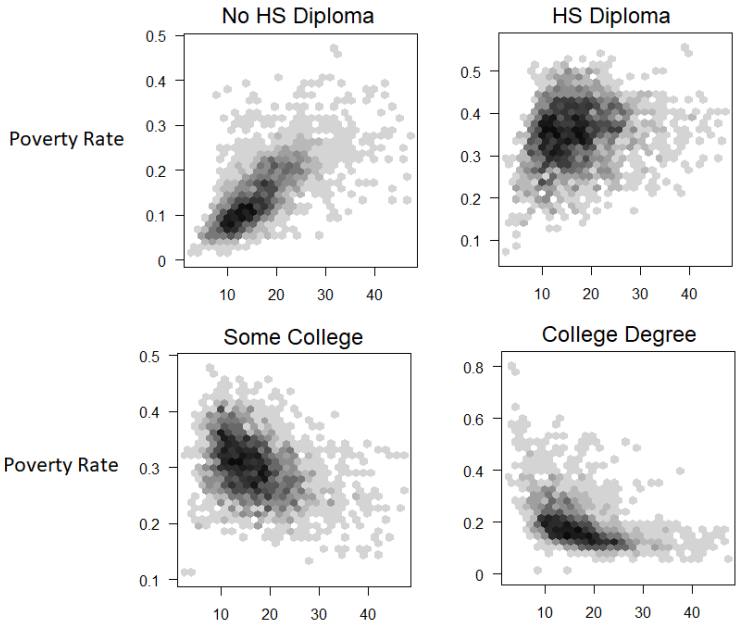

Fig 7: Poverty Rate by Education Level

Unfortunately, the source data only considers college as the post-High School educational option. It may also be useful to consider trade schools or even the availability of high-speed internet access to support online learning. Furthermore, as Trevor Noah describes below, sometimes providing tools in addition to education can make all the difference.

> "People love to say, "Give a man a fish, and he'll eat for a day. Teach a man to fish, and he'll eat for a lifetime." What they don't say is, "And it would be nice if you gave him a fishing rod." That's the part of the analogy that's missing."
>
> — Trevor Noah, *Born a Crime: Stories From a South African Childhood*

5. How Important are Healthy Residents?

Several features about resident health were included in the data. The three features below (Obesity, Smoking, and Diabetes) all tend to increase with the poverty rate.

Something else to consider is the concept of a "food desert", a neighborhood where healthy food is difficult to find. Following a poor diet can lead to obesity. Perhaps data regarding healthy food availability could also be helpful.

There tend to be more smokers in a higher poverty area. On the other hand, higher cigarette taxes have been shown to decrease the amount of smoking. It would be interesting to compare the average price of a pack of cigarettes to the other feature data.

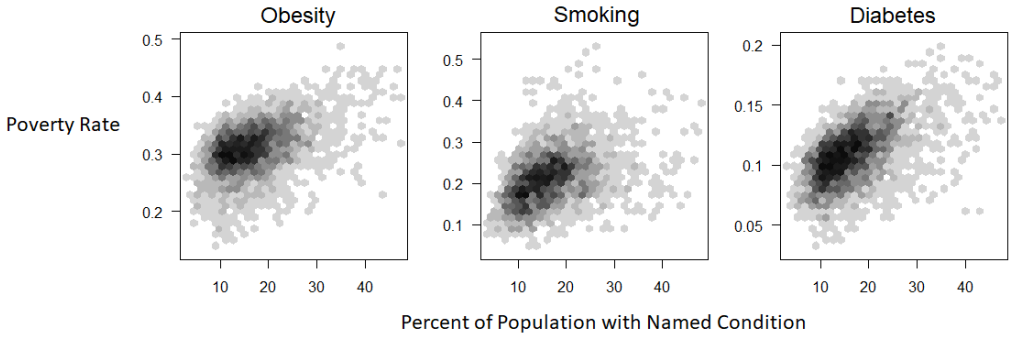

Fig 8: Poverty Rate for Several Health Factors

Regression Model

Using the observations described above, a Machine Learning regression model was created to predict poverty rate based on a similar set of features.

Feature Engineering

In order to avoid over-fitting the model, and in an attempt to create a simpler model, some features were removed.

Features Excluded from the Model due to Low Correlation with Poverty Rate

Category	Feature
Economic	Percent of children without health insurance
Demographic	Percent of population that is female
	Percent of population that is below 18 years of age
	Percent of population that is aged 65 years or older
	Percent of population that identifies as Hispanic
	Percent of population that identifies as Asian
	Percent of adult population which has a high school diploma as highest level of education achieved
	Births per 1,000 of population
Health	Air pollution - Fine particulate matter

Choosing the Regression Model Algorithm

Of the many regression models available, several were considered. The set included Linear, Boosted Decision Tree, Decision Forest, and Neural Network. Using Root-Mean-Squared Error (RMSE) as the evaluation mechanism, the Boosted Decision Tree performed the best with a RMSE of 2.77 and a Coefficient of Determination of 0.83. This model was selected for the predictive web service.

Conclusion

While there is no “silver bullet” solution to solving poverty, perhaps with a deeper understanding of the factors which directly contribute to the condition, resources and money can be directed in more effective ways.

The exploratory analysis above identified several features which correlate to the poverty rate, and some that do not. Factors like urban influence and economic topology would be difficult to change. However, a few factors appear to have low barriers to making a positive change. These are education, obesity, and diabetes. Perhaps programs to improve job skills and promote healthy eating and exercise would be a good next step.

It would also be beneficial to build upon the discoveries here and evaluate additional features such as financial data, housing density, and non-university post High School education.

Additional Sources

- Rural Urban Continuum Code (USDA Economic Research Service, https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/)
- County Typology Codes (USDA Economic Research Service, https://www.ers.usda.gov/data-products/county-typology-codes.aspx)
- Relationship between cigarette tax and Smoking (https://www.tobaccofreekids.org/assets/factsheets/0146.pdf)