Analysis of poverty rate

Seema G, Jan 2018

Executive Summary

The goal of this analysis is to predict poverty across US at the county-level from socioeconomic indicators. The dependent variable for this analysis is poverty_rate which represents percentage of a county's population meeting the criteria for poverty as defined. Socioeconomic indicators considered in this analysis include metro/non-metro county type, county's economic dependence types including farming, mining, manufacturing, Federal/State government, or recreation, percentage of civilian labor force, unemployment percentage, health indicators, age and ethnicity classification.

https://www.datasciencecapstone.org/competitions/3/county-poverty/page/10/ has detailed information about the data. Labeled data contains 3198 rows and 34 columns.

Initial step involved exploring data, gathering summary statistics and creating visualizations of data to identify potential relationships. After exploring the data, a regression model was created to predict poverty rate.

After performing the analysis, in conclusion, while many factors can help predict poverty rate, significant features found in this analysis were:

Civilian labor force and unemployment — Counties with higher percent of civilian labor force had lower poverty rates. Counties with lower percent of unemployed adults had lower poverty rates.

Education – Counties with lower percent of adult population that did not have a high school diploma had lower poverty rates.

Economic dependence - Poverty rate tends to be higher for Federal/State government-dependent counties, and lower for farm and recreation dependent counties.

Rural/Urban area - Large-metro area with at least 1 million residents or more counties tend to have lowest poverty rate. Poverty rate for non-metro counties tend to be higher than metro counties. There was more variance in the poverty rate for non-metro counties than for metro counties

Health indicators – Counties with lower percent of adult smokers, lower percentage of adults with diabetes, and lower percentage of adult obesity had lower poverty rates.

Overview of steps

Following steps were performed in this analysis.

1. Data exploration

2. Data preparation and handling missing data

3. Feature selection, scaling, and editing Meta data

4. Regression modeling

5. Permutation Feature importance and feature engineering

6. Model evaluation and updates

7. Tuning model hyper parameters

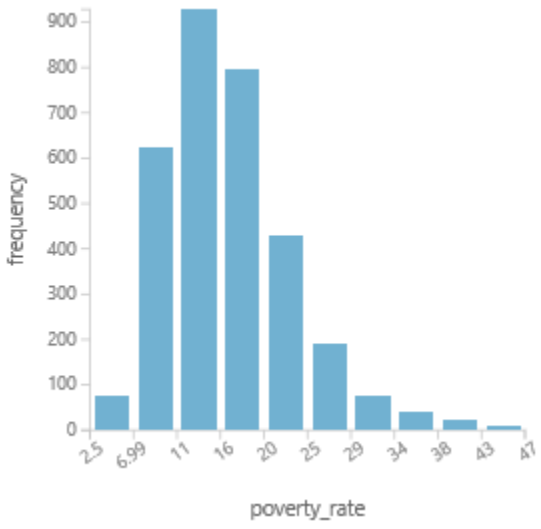
8. Predicting poverty rate for private test data

Initial Data Exploration

Initial data exploration began with summary and descriptive statistics. Summary statistics for poverty rate is as follows.

Mean	16.8171
Median	15.8
Min	2.5
Max	47.4
Standard Deviation	6.698
Unique Values	341
Missing Values	0
Feature Type	Numeric Feature

Following histogram of poverty rate indicates that it has a uniform right tailed distribution.

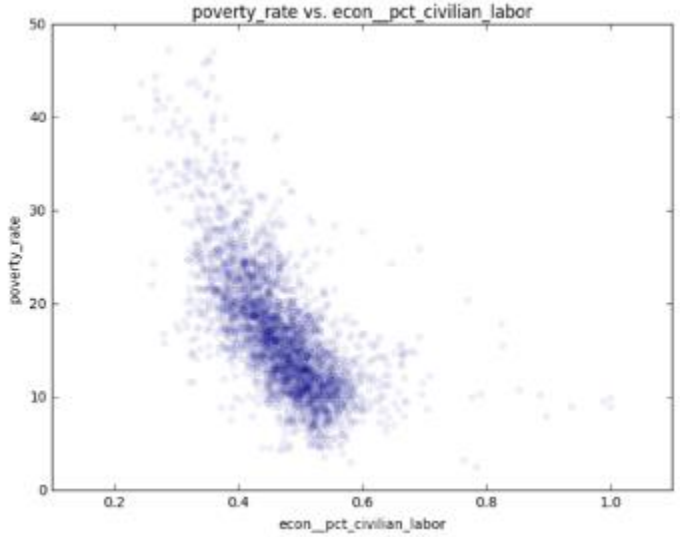


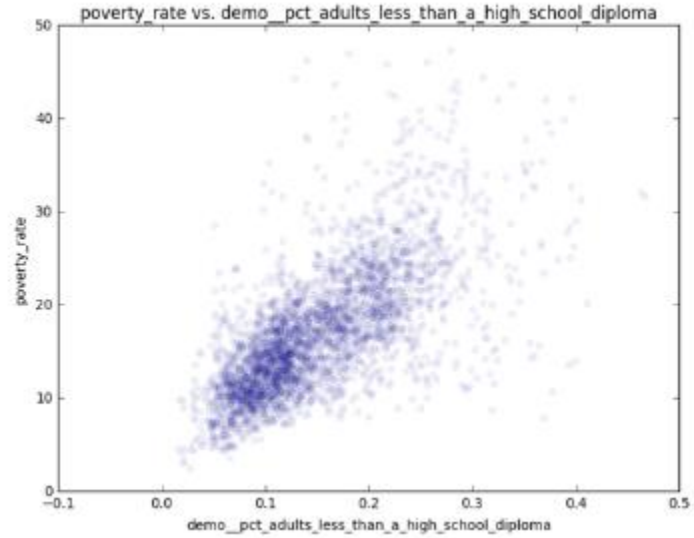
Relationships in data

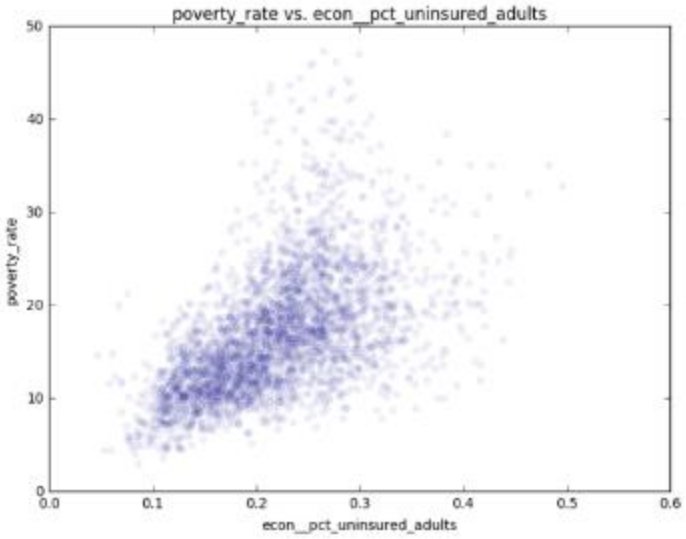
Relationship with numeric features

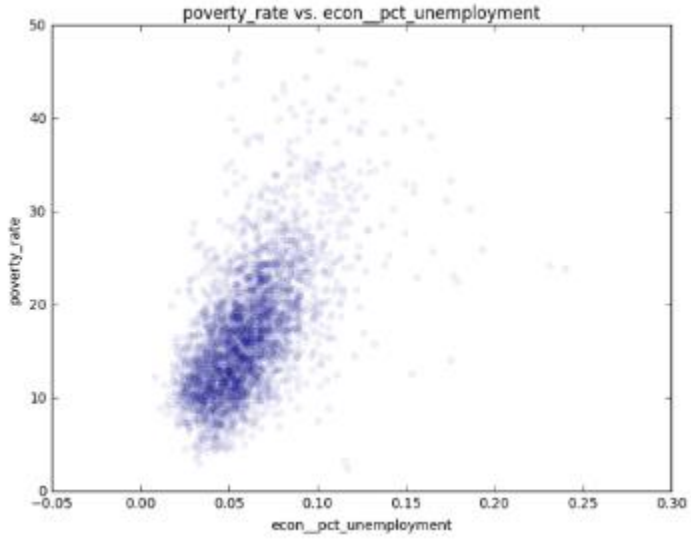
Poverty rate was analyzed against numerical features like civilian labor force, percentage of unemployed adults, percentage of adults without health insurance and percentage of adults with less than a high school diploma. Following correlation coefficients indicate that higher values of (unemployment, percent adults without high school diploma and uninsured health) relate to higher poverty rates. Poverty rate is negatively correlated to civilian labor force. Higher values of civilian labor force result in lower poverty rates.

	poverty_rate
poverty_rate	1.000000
econ__pct_civilian_labor	-0.670417
demo__pct_adults_less_than_a_high_school_diploma	0.680360
econ__pct_uninsured_adults	0.541712
econ__pct_unemployment	0.592022







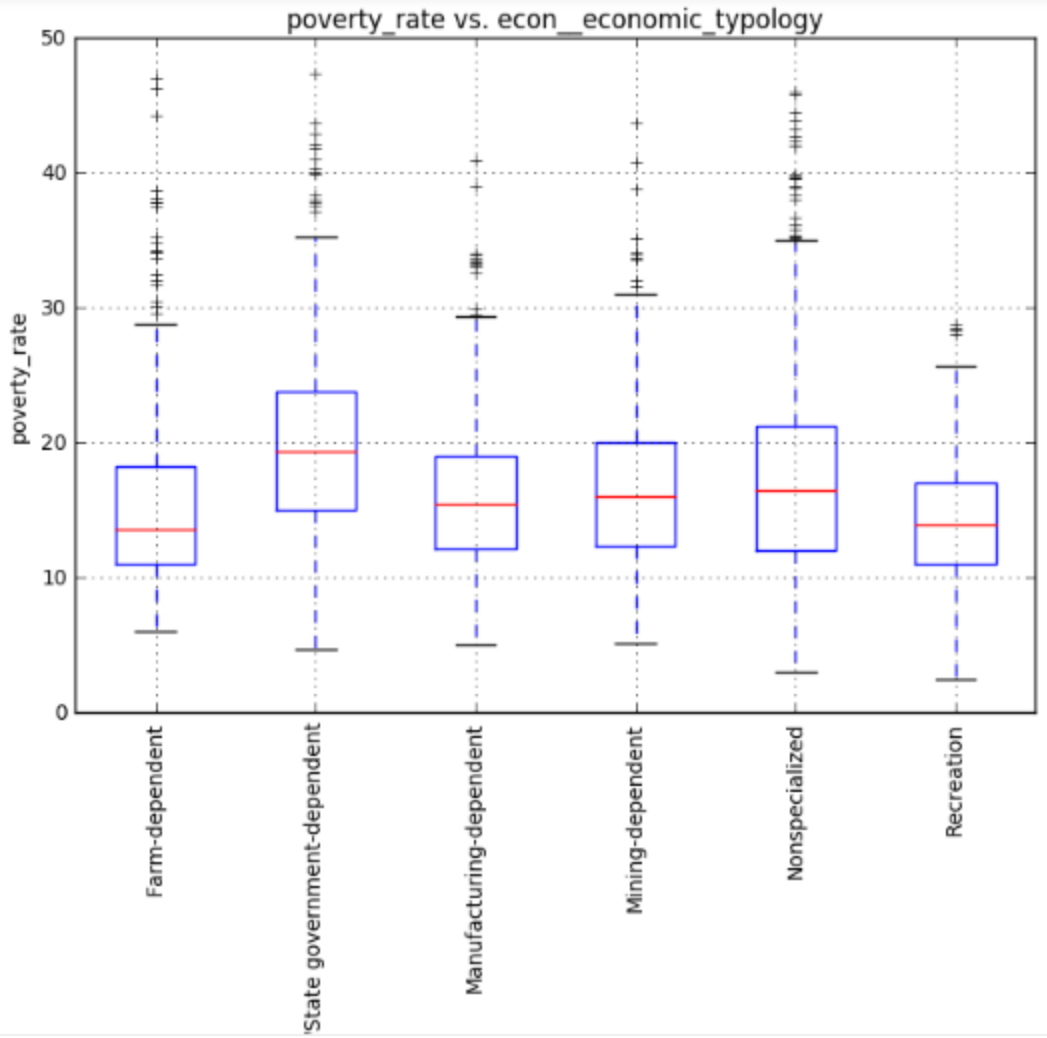


Following matrix shows poverty rate's correlation to age and health indicators. Poverty rate is positively correlated to percentage of adult smokers, percentage of population who are diabetic and percentage of adult obesity.

	poverty_rate
poverty_rate	1.000000
health__pct_adult_obesity	0.444293
health__pct_adult_smoking	0.395457
health__pct_diabetes	0.537038
health__pct_excessive_drinking	-0.353254
demo__pct_below_18_years_of_age	0.039237
demo__pct_aged_65_years_and_older	-0.088123

Relationship with categorical features

Economic_typology of a county represents the county's economic dependence type amongst farming, mining, manufacturing, Federal/State government, recreation or unspecified. Median poverty rate was highest for Federal/State government-dependent counties, and lower for farm and recreation dependent counties.

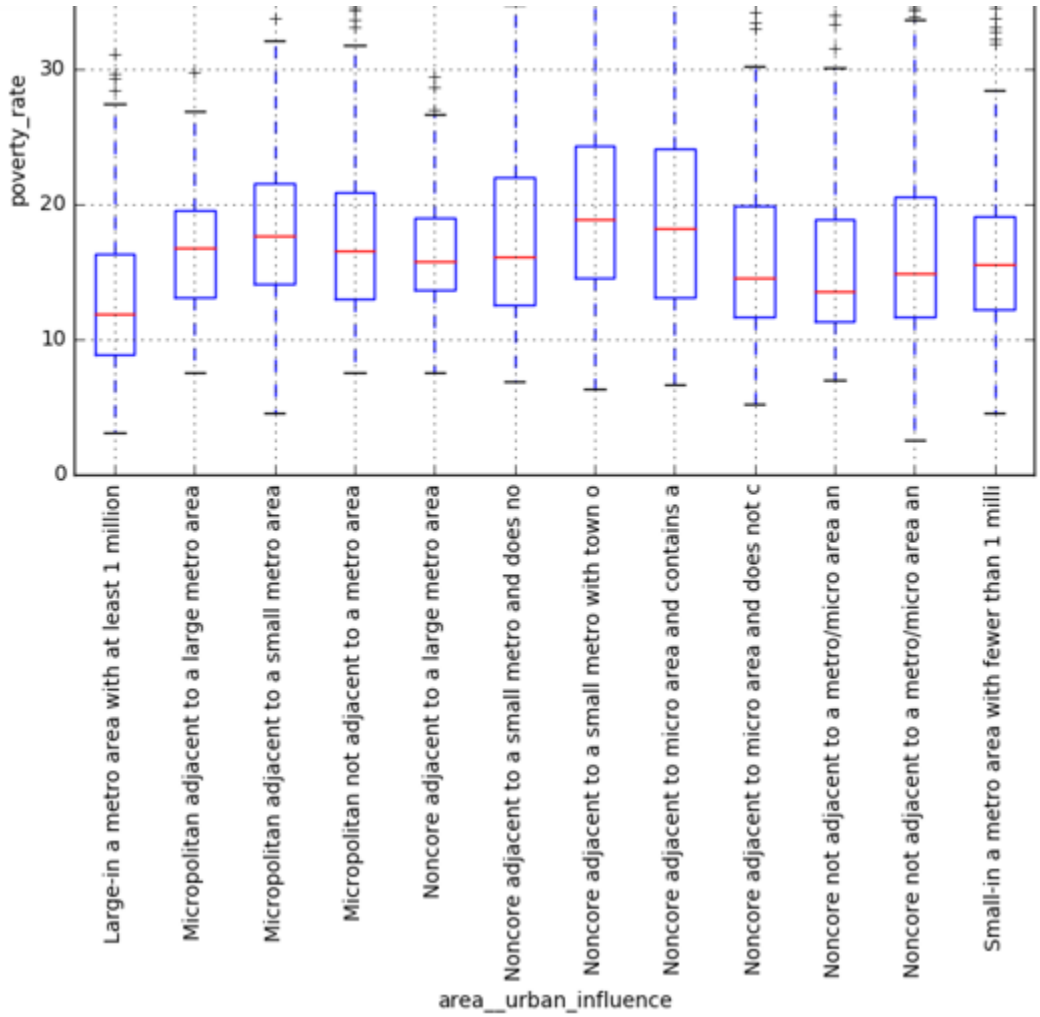


area__urban_influence — Urban Influence Codes "form a classification scheme that distinguishes metropolitan counties by population size of their metro area, and nonmetropolitan counties by size of the largest city or town and proximity to metro and micropolitan areas." This column has following unique values.

```
df.area__urban_influence.unique()
```

```
array(['Noncore adjacent to a large metro area',
       'Micropolitan adjacent to a large metro area',
       'Noncore adjacent to micro area and contains a town of 2,500-19,999 residents',
       'Large-in a metro area with at least 1 million residents or more',
       'Micropolitan not adjacent to a metro area',
       'Noncore not adjacent to a metro/micro area 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',
       'Small-in a metro area with fewer than 1 million 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 adjacent to a small metro and does not contain a town of at least 2,500 residents',
       'Micropolitan adjacent to a small metro area'], dtype=object)
```

Following plot illustrates that Large-in a metro area with at least 1 million residents or more counties have the lowest median poverty rate.

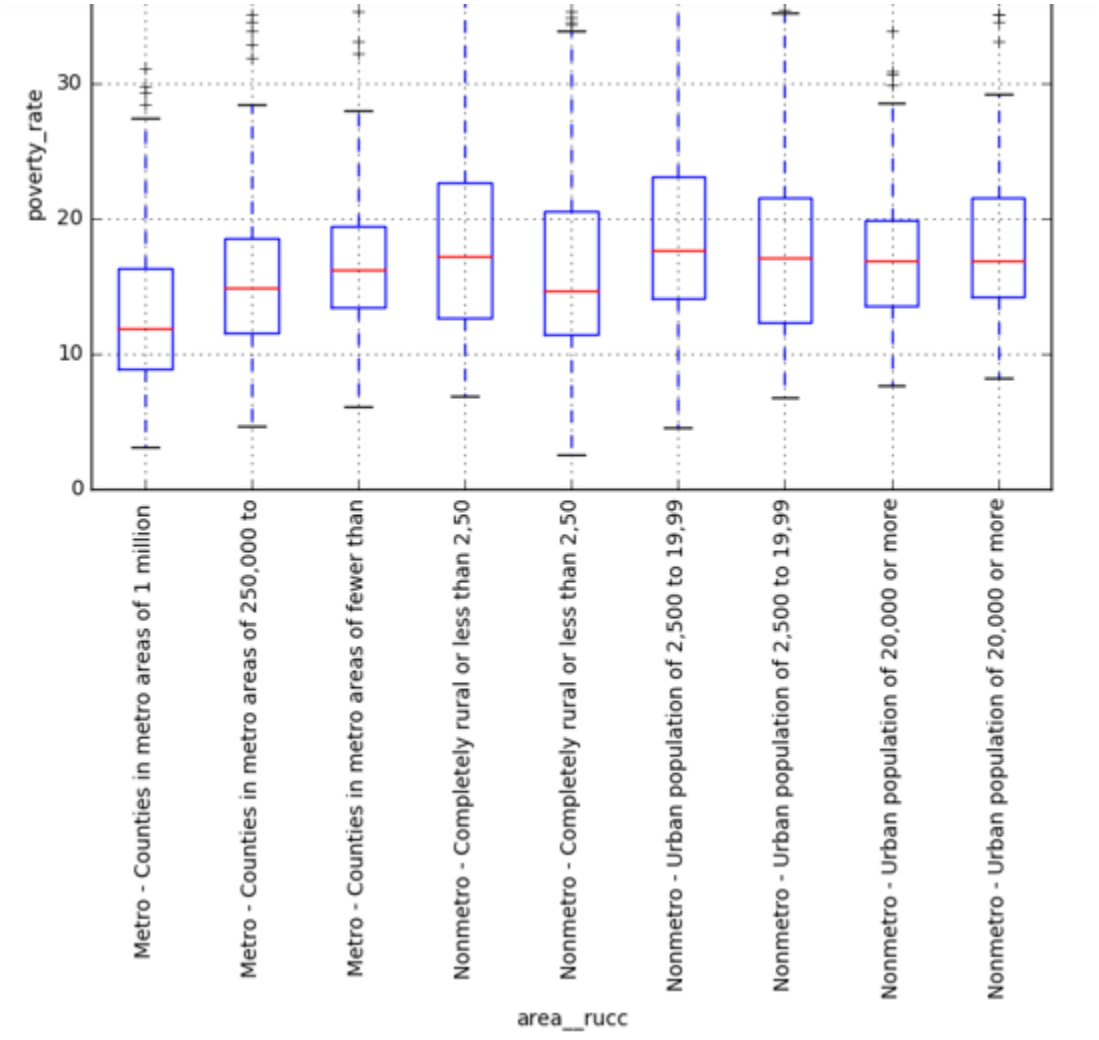


area__rucc — Rural-Urban Continuum Codes "form a classification scheme that distinguishes metropolitan counties by the population size of their metro area, and nonmetropolitan counties by degree of urbanization and adjacency to a metro area. A county is defined as "metro" if the rural-urban continuum code for that country begins with the text "Metro". This column has following unique values.

```
df.area__rucc.unique()
```

```
array([ 'Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area',
       'Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area',
       'Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area',
       'Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area',
       'Metro - Counties in metro areas of 1 million population or more',
       'Metro - Counties in metro areas of 250,000 to 1 million population',
       'Nonmetro - Urban population of 20,000 or more, adjacent to a metro area',
       'Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area',
       'Metro - Counties in metro areas of fewer than 250,000 population'], dtype=object)
```

Following plot shows that the median poverty rate for non-metro counties is higher than for metro counties. There is more variance in the poverty rate for non-metro counties than for metro counties.



Getting data ready for machine learning

Raw labeled-data has 3198 rows and 33 features. After filling missing data with zeros, converting string columns to categorical and using z-score transformation for numeric columns, data was prepared for machine learning.

Private unlabeled test data

Unlabeled test data has 3080 rows for which poverty rate needed to be predicted.

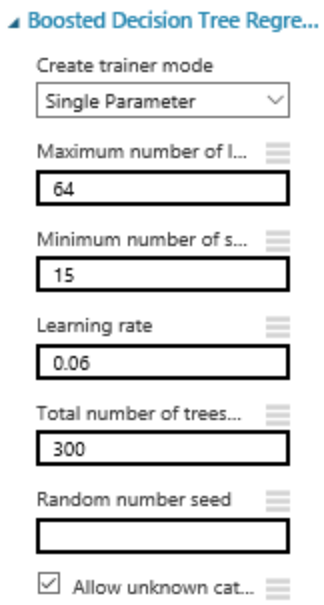
Permutation Feature Importance

Azure ML's permutation feature importance module was used to identify important features that are more sensitive to the shuffling process, resulting in ranking feature variables in order of permutation importance scores. Following are the top 10 features identified by permutation feature importance and the corresponding scores,

Feature	Score
econ__pct_civilian_labor	1.187074
demo__pct_adults_less_than_a_high_school_diploma	0.9467
econ__pct_uninsured_adults	0.605735
demo__pct_non_hispanic_white	0.514892
demo__pct_below_18_years_of_age	0.403315
econ__pct_unemployment	0.253349
econ__economic_typology	0.180773
area__rucc	0.150831
demo__pct_aged_65_years_and_older	0.126074
health__pct_diabetes	0.125984

Regression model to predict poverty rate

Labeled data was split into 90% training data and 10% test data. Model was trained using Boosted decision tree regression with the following parameters. These parameters were identified using Tune model hyper parameters module and doing a sweep across a range of parameters.

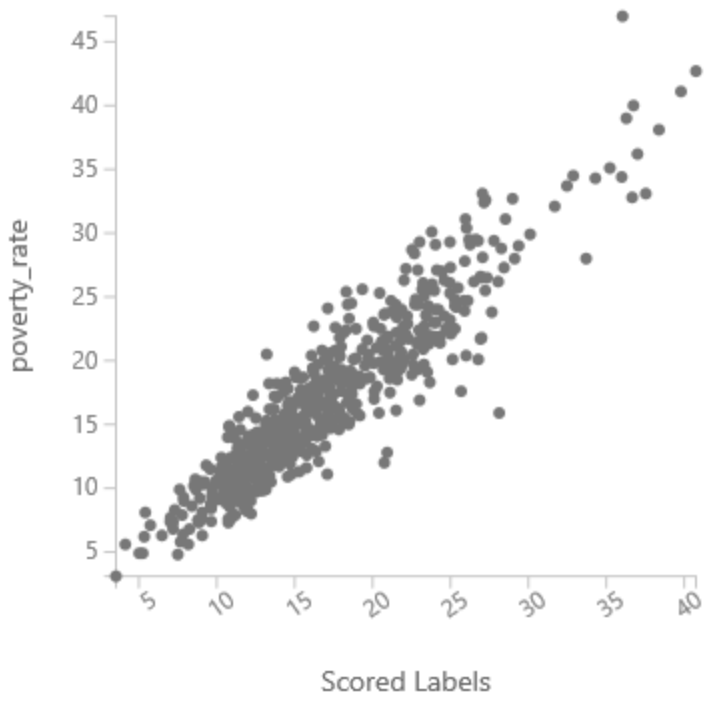


- maximum number of leaves per tree was set to 64.
- minimum number of samples per leaf node was set to 15.
- learning rate was set to 0.06.
- Total number of trees constructed was set to 300.

The modeled was scored against the 10% labeled test data. It yielded Root Mean Square Error (RMSE) of 2.26 and co-efficient of determination of 0.867 as follows.

Mean Absolute Error	1.721415
Root Mean Squared Error	2.265452
Relative Absolute Error	0.351333
Relative Squared Error	0.132027
Coefficient of Determination	0.867973

Following is a scatter plot of poverty rate versus predicted/scored labels. Based on these metrics the model was accepted for predicting labels for the private test data.



Predicting poverty rate for private unlabeled test data

Then the trained model was used to predict poverty rates for unlabeled test data. The Root Mean Square Error (RMSE) metric after the submission was 2.89.

Conclusion

This analysis has shown that poverty rate can be predicted from socioeconomic indicators like percentage of civilian labor force, unemployment percentage, uninsured health percentage, percent of adults without high school diploma, economic dependence, metro/non-metro area and health indicators of the population.

Appendix A.

https://www.datasciencecapstone.org/competitions/3/county-poverty/page/10/ has detailed information about the data.