Prediction of United States' Counties Poverty Rates

Mark Vervuurt
25-01-2018

Executive Summary

This document presents the results of the regression analysis to predict poverty rates of United States' Counties. The CRISP-DM methodology was used throughout the regression analysis. The result of this regression analysis is the creation of a regression model with an RMSE of 2.7847.

In the data understanding phase was discovered that the features shown in the table here under play a significant role in predicting United States' counties poverty rates. They have moderate positive and negative Pearson correlation coefficients. Furthermore, the Boxplots of the categorical variables show enough variation and separation of the data with respect to the target variable poverty_rate. Moreover recursive function elimination in combination with cross validation was ultimately used to select the optimum number of features for the lowest RMSE score.

Significant Features	**Short Description**
area__rucc	Rural urban continuum code of county
area__urban_influence	Urban influence continuum code of county
econ__economic_typology	Economic dependence type of county
demo__pct_adults_less_than_a_high_school_diploma	Percentage of adults with less than high school diploma per county
health__homicides_per_100k	Homicides per 100k inhabitants per county
econ__pct_unemployment	Percentage of unemployment per county
health__pct_low_birthweight	Percentage of low birth weight per county
econ__pct_uninsured_adults	Percentage of uninsured adults per county
health__pct_diabetes	Percentage of diabetes per county
demo__pct_non_hispanic_african_american	Percentage of African Americans per county
econ__pct_civilian_labor	Percentage of civilian labor per county
demo__pct_non_hispanic_white	Percentage of Hispanics and whites per county
demo__pct_adults_bachelors_or_higher	Percentage of Adults with bachelor degree or higher per county

Data Science methodology

The CRISP-DM Methodology was used in order to create an accurate regression model:

- **Business Understanding**: read through the 'Rural Poverty & Well-being' online report to better understand the circumstances of poverty.
- **Data Understanding**: explore the quantitative and categorical variables that play a key role in predicting poverty rates. Create new, better and informative features.
- **Data Preparation**: drop redundant and uninformative features, fill missing values, etc.
- **Modeling**: create and select the best regression model.
- **Evaluation**: compare and evaluate the regression models using nested cross validation.
- **Deployment**: the deployment of the regression model is not strictly applicable here. However presenting the results of the regression analysis with this report can be considered as the deployment step.

Business Understanding

As described in the online report the 'Rural Poverty & Well-being': "Concentrated poverty contributes to poor housing and health conditions, higher crime and school dropout rates, as well as employment dislocations". With this information the data will be explored to see how health, crime, education and employment related factors contribute to poverty.

Another important feature of poverty is time. An area that doesn't have a high level of poverty in two following years is likely better off than an area that has a high level of poverty in both years. It will not be possible to construct a feature with this information because we cannot compare the state's poverty rate over year 'a' and 'b' within this dataset. We don't have a unique key to identify counties.

Counties are generally compared by their Non-Metro and Metro status. There is more poverty in Non-Metro areas than Metro areas. Poverty is also higher under certain ages and ethnicities. Here also the data will be explored on the basis of this information.

Data Understanding

In order to build this regression model and determine its most significant features, the data was explored to understand the relationship between poverty rates and other features.

Initial Data Exploration

The dataset consists of 3198 records about United States' counties. Each record contains socioeconomic indicators about a United States' county for a given year. The dataset contains 32 features about socioeconomic indicators besides the 'row_id', 'yr' feature and the target feature 'poverty_rate',.

Individual Feature Statistics

Here are the summary statistics for all the socioeconomic features:

- summary statistics of categorical variables: the total count (count), number of unique elements (unique), most frequent element (top) and the frequency of the most frequent element (frequent)
- summary statistics of quantitative variables: the mean, the standard deviation (std), the minimum value (min), 25% percentile, 50% percentile (median), 75% percentile and the maximum value (max).

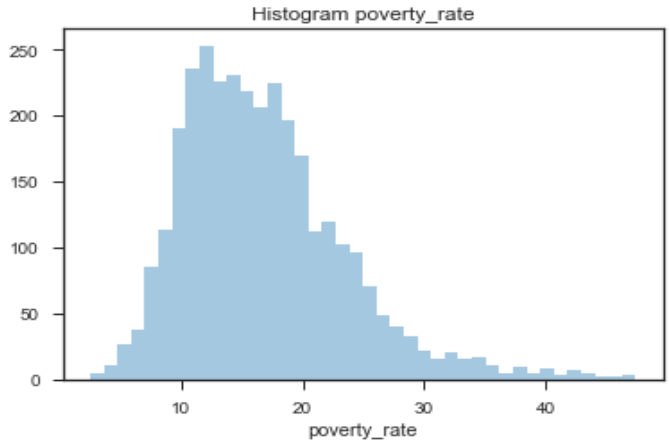
feature	count	mean	std	min	25%	50%	75%	max
econ__pct_civilian_labor	3198.0	0.467071	0.074541	0.217000	0.420000	0.467000	0.514000	1.000.000
econ__pct_unemployment	3198.0	0.059610	0.022850	0.008000	0.044000	0.057000	0.071000	0.240000
econ__pct_uninsured_adults	3196.0	0.217534	0.067372	0.046000	0.166000	0.216000	0.262000	0.495000
econ__pct_uninsured_children	3196.0	0.085920	0.040005	0.009000	0.057000	0.077000	0.105000	0.285000
demo__pct_female	3196.0	0.498781	0.024251	0.294000	0.493000	0.503000	0.512000	0.576000
demo__pct_below_18_years_of_age	3196.0	0.227763	0.034291	0.098000	0.207000	0.226000	0.245250	0.417000
demo__pct_aged_65_years_and_older	3196.0	0.170137	0.043594	0.043000	0.142000	0.167000	0.194000	0.355000
demo__pct_hispanic	3196.0	0.090233	0.142707	0.000000	0.019000	0.035000	0.088000	0.945000
demo__pct_non_hispanic_african_american	3196.0	0.091117	0.147104	0.000000	0.006000	0.022000	0.096250	0.855000
demo__pct_non_hispanic_white	3196.0	0.770207	0.207903	0.060000	0.648000	0.854000	0.936000	0.998000
demo__pct_american_indian_or_alaskan_native	3196.0	0.024659	0.084634	0.000000	0.002000	0.007000	0.014000	0.852000
demo__pct_asian	3196.0	0.013304	0.025366	0.000000	0.003000	0.007000	0.013000	0.346000
demo__pct_adults_less_than_a_high_school_diploma	3198.0	0.148794	0.068255	0.016129	0.097468	0.133501	0.195171	0.466867
demo__pct_adults_with_high_school_diploma	3198.0	0.350300	0.070534	0.072821	0.305915	0.355701	0.399197	0.551689
demo__pct_adults_with_some_college	3198.0	0.301366	0.052498	0.112821	0.265362	0.301595	0.335972	0.474216
demo__pct_adults_bachelors_or_higher	3198.0	0.199540	0.089158	0.013986	0.138840	0.177247	0.233258	0.794872
demo__birth_rate_per_1k	3198.0	11.676.986	2.739.516	4.000.000	10.000.000	11.000.000	13.000.000	29.000.000
demo__death_rate_per_1k	3198.0	10.301.126	2.786.143	0.000000	8.000.000	10.000.000	12.000.000	27.000.000
health__pct_adult_obesity	3196.0	0.307599	0.043404	0.140000	0.284000	0.309000	0.334000	0.484000
health__pct_adult_smoking	2734.0	0.213519	0.063090	0.050000	0.171000	0.211000	0.249750	0.526000
health__pct_diabetes	3196.0	0.109287	0.023197	0.033000	0.094000	0.109000	0.124000	0.197000
health__pct_low_birthweight	3016.0	0.083534	0.022382	0.025000	0.068000	0.080000	0.095000	0.232000
health__pct_excessive_drinking	2220.0	0.164832	0.050232	0.038000	0.129000	0.164000	0.196000	0.358000
health__pct_physical_inacticity	3196.0	0.277309	0.052947	0.097000	0.243000	0.280000	0.313000	0.443000
health__air_pollution_particulate_matter	3170.0	11.626.498	1.544.928	7.000.000	10.000.000	12.000.000	13.000.000	15.000.000
health__homicides_per_100k	1231.0	5.950.747	5.063.374	-0.390000	2.660.000	4.840.000	7.825.000	51.490.000
health__motor_vehicle_crash_deaths_per_100k	2781.0	21.116.077	10.516.984	3.090.000	13.460.000	19.630.000	26.470.000	110.450.000
health__pop_per_dentist	2954.0	3.431.442.789	2.569.444.414	339.000.000	1.812.250.000	2.690.000.000	4.089.750.000	28.129.000.000
health__pop_per_primary_care_physician	2968.0	2.551.349.730	2.100.475.931	189.000.000	1.419.000.000	1.999.000.000	2.859.000.000	23.400.000.000

Here are the summary statistics for the target variable which is quantitative:

feature	count	mean	std	min	25%	50%	75%	max
poverty_rate	3198.0	16.817136	6.697969	2.5	12.0	15.8	20.3	47.4

Poverty rates are right or positively skewed with a skew value of 1.048357. We can recognize a slight bell curve in the data. The mean and median are relatively close to each other and the standard deviation is relatively low which indicates low variability in the poverty rates.

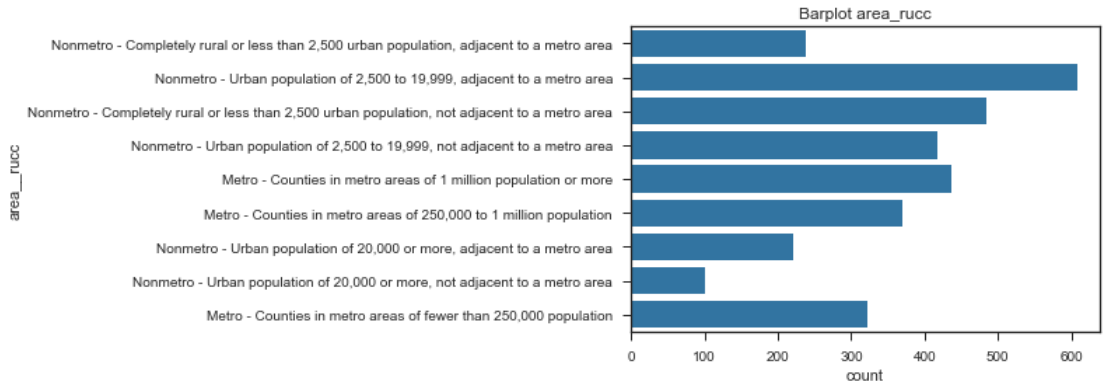
Most United States' counties have a poverty_rate between 10% and 20% poverty as shown in the histogram here under:

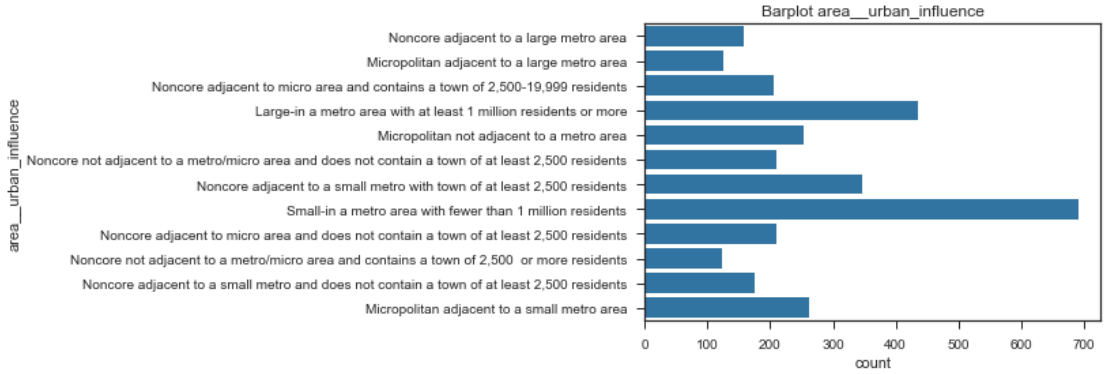


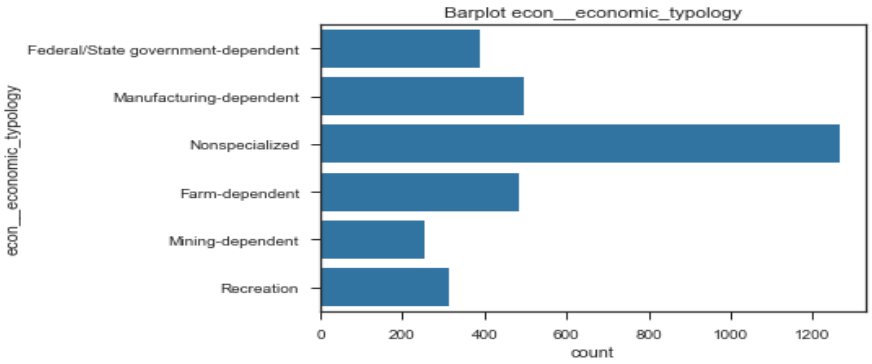
There are four categorical variables included in the dataset:

- ‘area__rucc’ with 9 categories:
  - 'Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area' counties are most frequent with 608 counties.
  - 'Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area' counties are most infrequent with 100 counties.
- ‘area__urban_influence‘ with 12 categories:
  - 'Small-in a metro area with fewer than 1 million residents' counties are most frequent with 692 counties.
  - 'Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents' counties are most infrequent with 122.
- ‘econ__economic_typology’ with 6 categories:
  - 'Non specialized' economic typology counties are most frequent with 1266 counties.
  - 'Mining-dependent' economic typology counties are most infrequent with 254 counties.
- ‘yr’ with 2 categories

The respective horizontal bar plots for ‘area__rucc’, ’ area__urban_influence’ and ‘econ__economic_typology’ are shown here under.







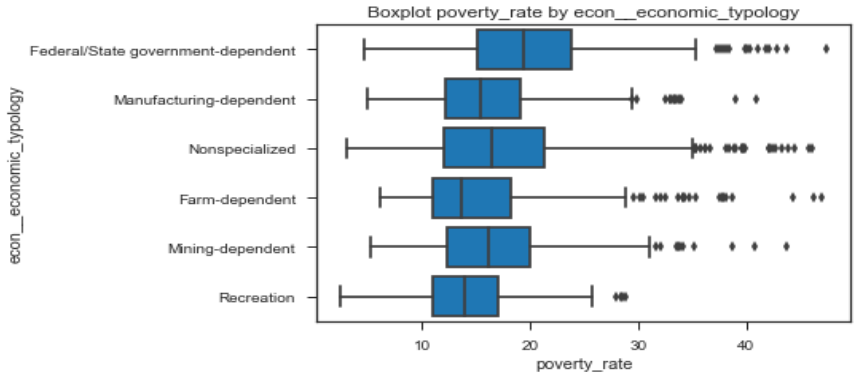
Data Exploration and Visualization of Categorical Relationships

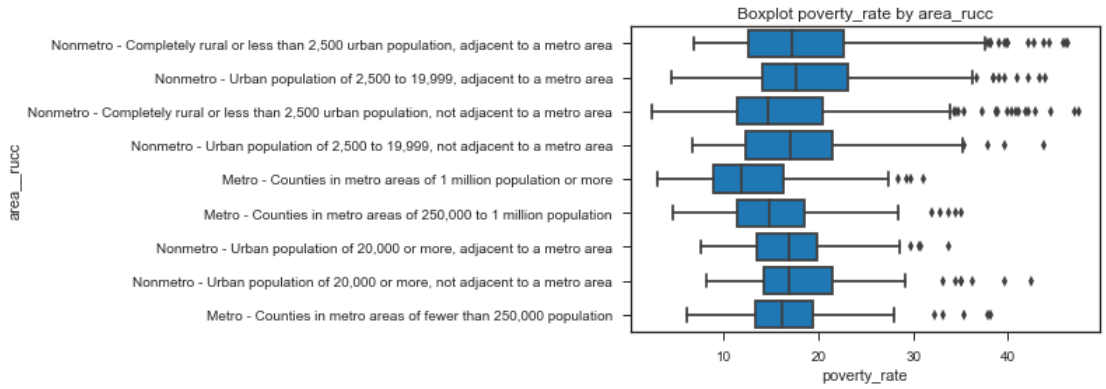
The box plots of the categorical variables ‘econ__economic_typology’, ‘area__urban_influence’ and ‘area__rucc’ show interesting variation with respect to the target feature poverty_rate:

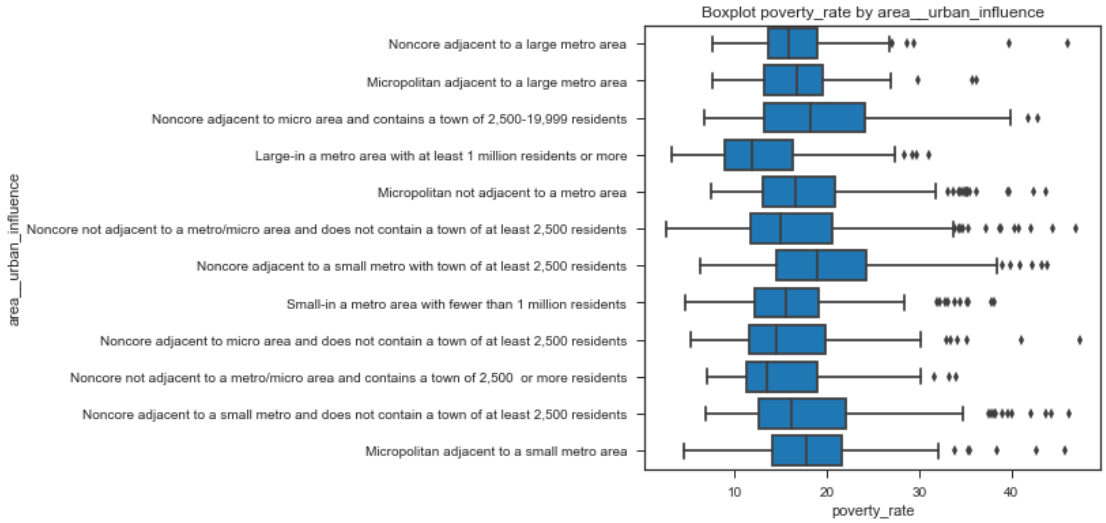
- 'Farm-dependent' counties have the lowest poverty rates and 'Federal/State government-dependent' counties have the highest poverty rates.
- 'Metro - Counties in metro areas with 1 million population or more' counties have the lowest poverty rates.

- 'Large-in a metro area with at least 1 million residents or more' counties have the lowest poverty rates.

The difference in poverty over year 'a' and 'b' is really minimal. This feature will therefore be dropped at the cleaning stage.







Data Exploration and Visualization of Quantitative Relationships

The correlation matrix is computed first followed for the quantitative variables followed by their respective scatter plot matrices.

Correlation Matrix

After reading the 'Rural Poverty & Well-being' online report it is clear that education, demography and health related issues play an important role in predicting poverty. In this dataset are also added economic indicators of United States' counties. The correlation matrix shows that the following features correlate moderately positively or negatively with the target feature poverty_rate:

Feature	**Socioeconomic indicator**
demo__pct_adults_less_than_a_high_school_diploma demo__pct_adults_bachelors_or_higher	Education
health__homicides_per_100k health__pct_low_birthweight econ__pct_uninsured_adults health__pct_diabetes	Health related
econ__pct_unemployment econ__pct_civilian_labor	Economy
demo__pct_non_hispanic_african_american demo__pct_non_hispanic_white	Demography

The whole correlation matrix of interest is shown here under. Only the Pearson correlation coefficients between the socioeconomic features and the target feature poverty_rate are shown.

Features	**Correlation coefficient with poverty_rate**
econ__pct_civilian_labor	-0.670417
demo__pct_non_hispanic_white	-0.499974
demo__pct_adults_bachelors_or_higher	-0.467134
demo__pct_adults_with_some_college	-0.363875
health__pct_excessive_drinking	-0.353254
demo__pct_asian	-0.163033
demo__pct_aged_65_years_and_older	-0.088123
demo__pct_female	-0.068065
demo__pct_below_18_years_of_age	0.039237
health__air_pollution_particulate_matter	0.058582
econ__pct_uninsured_children	0.098882
demo__pct_hispanic	0.105574
demo__birth_rate_per_1k	0.127506
health__pop_per_primary_care_physician	0.156942
demo__pct_adults_with_high_school_diploma	0.202928

demo__pct_american_indian_or_alaskan_native	0.236508
demo__death_rate_per_1k	0.244093
health__pop_per_dentist	0.268996
health__pct_adult_smoking	0.395457
health__motor_vehicle_crash_deaths_per_100k	0.420348
health__pct_physical_inacticity	0.437680
health__pct_adult_obesity	0.444293
demo__pct_non_hispanic_african_american	0.507048
health__pct_diabetes	0.537038
econ__pct_uninsured_adults	0.541712
health__pct_low_birthweight	0.565456
econ__pct_unemployment	0.592022
health__homicides_per_100k	0.621399
demo__pct_adults_less_than_a_high_school_diploma	0.680360

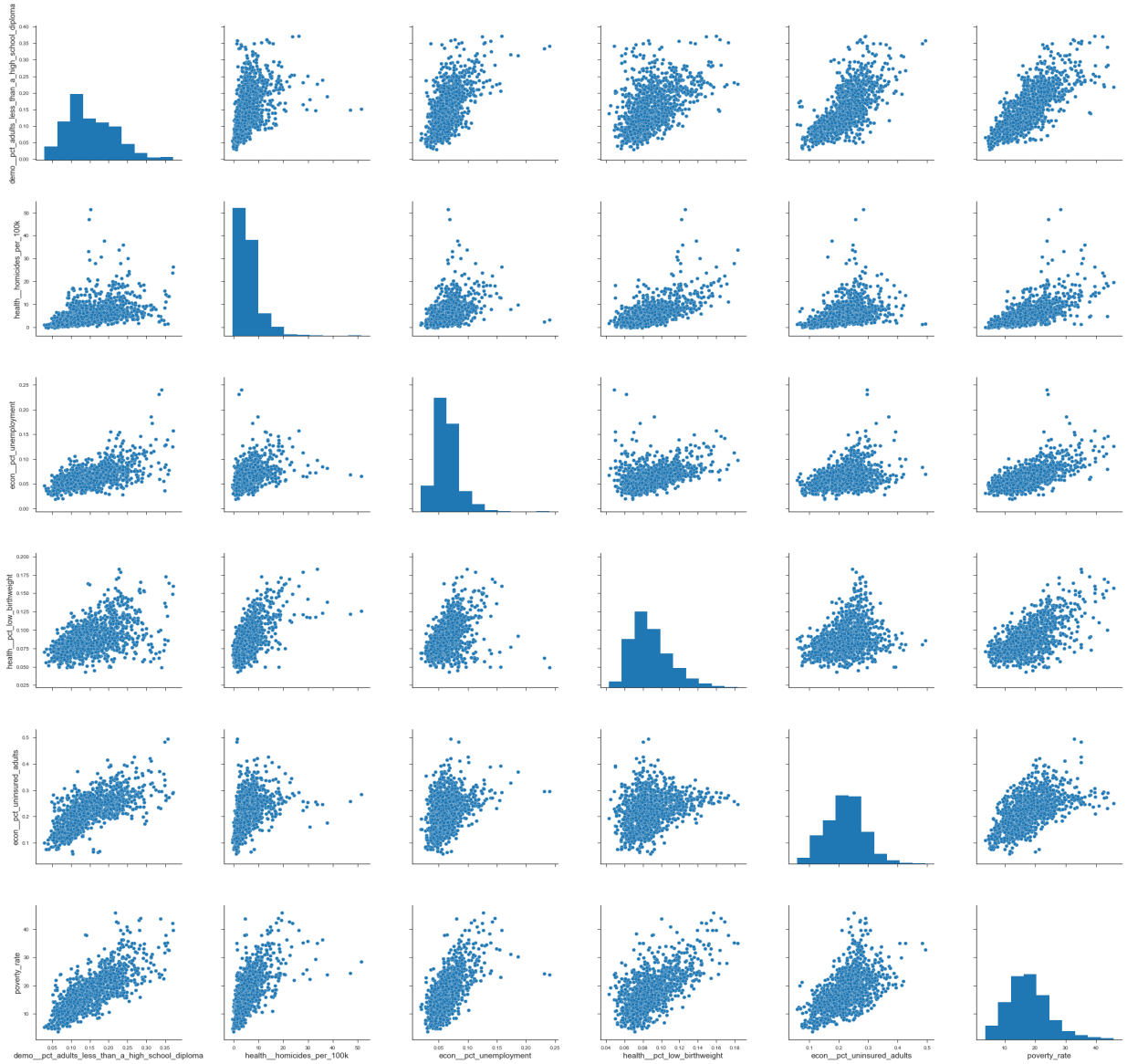
Scatter Plot Matrices

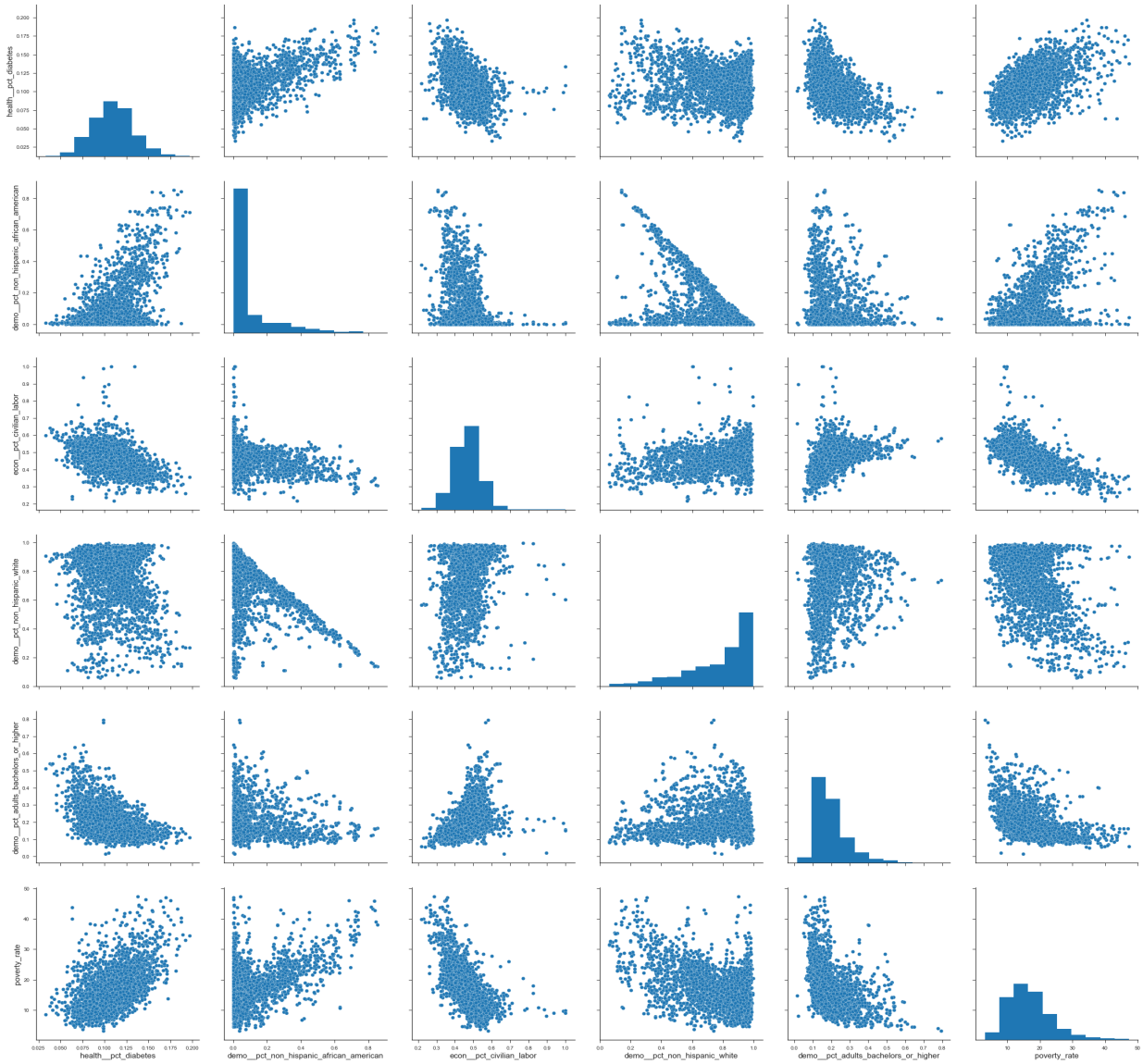
The correlation matrices and scatter plot matrices visually confirm that the variables that correlate moderately strong with the target feature 'poverty_rate' seem to have a linear relationship.

NB: linear statistical transformations (sqrt, square, exponential, etc) were also applied to the target variable 'poverty_rate' but they did not improve substantially the correlation coefficients or linear relationships in the scatter plot matrices.

Moderately Strong Correlating Features Scatter Plot Matrices

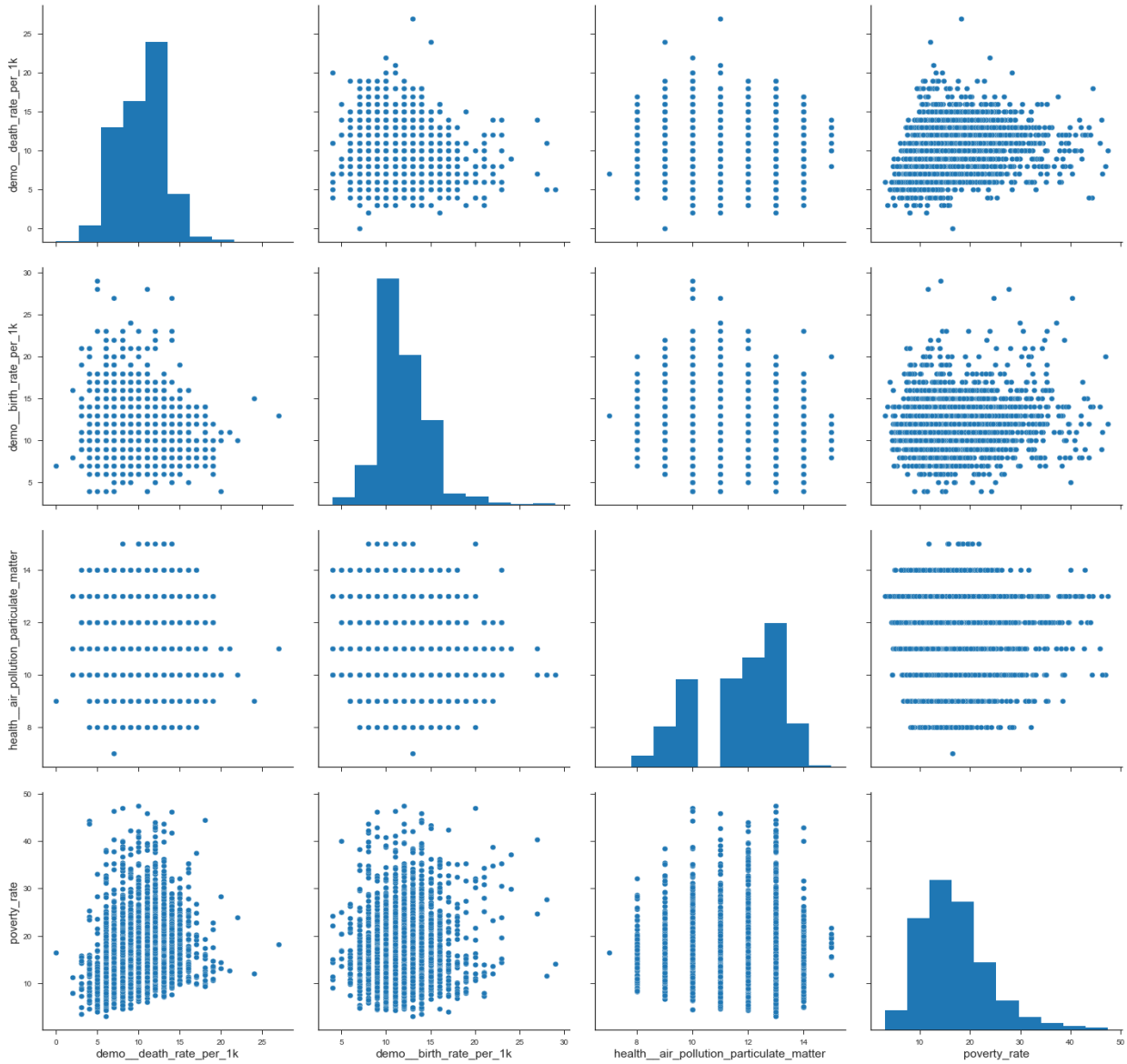
The scatter plot matrices are limited to the variables that correlate moderately strong with the target feature 'poverty_rate'. There are more features of interest, however this would result in too many scatter plot matrices to display in this report.




health__pct_diabetes
demo__pct_non_hispanic_african_american
econ__pct_civilian_labor
demo__pct_non_hispanic_white
demo__pct_adults_bachelors_or_higher
poverty_rate

Categorical Features Scatter Plot Matrices

These quantitative features behave like categorical features with respect to the target feature poverty_rate.



Creation and Visualization of New Categorical Variables¶

From observing the scatter plot matrices of the features 'demo__death_rate_per_1k', 'demo__birth_rate_per_1k' and 'health__air_pollution_particulate_matter' is clear that these quantitative variables behave like categorical variables. They will be transformed into categorical features by binning them.

Data Preparation

This phase involves mostly the cleaning, scaling and one hot encoding of features:

- Dropping redundant features that been transformed into categorical features and uninformative features such as ‘yr’.
- Converting features to the right type.
- Missing values are replaced by the respective median value of the feature. The median is preferred over the mean because it is less sensible to skewed data and gives a better measure of centrality.
- Features are scaled to have the same scale. The MinMaxScaler is applied to the features ‘health__homicides_per_100k', 'health__motor_vehicle_crash_deaths_per_100k', 'health__pop_per_dentist' and 'health__pop_per_primary_care_physician' to scale them the same way as other quantitative variables that are in percentages between 0 and 1.
- One hot encoding of the categorical variables is performed.

The result of this phase is a dataset of 3198 records with 82 features.

Modeling and Evaluation

In this phase two regression models are compared with each other using the RMSE evaluation metric:

- Least Square Linear Model after applying recursive feature selection to create a linear model with the most important features
- An AdaBoostRegressor which is an ensemble learning model of decision trees.

The best RMSE scores obtained by these two models are show in the table here under. Both models are compared with each other using nested cross validation.

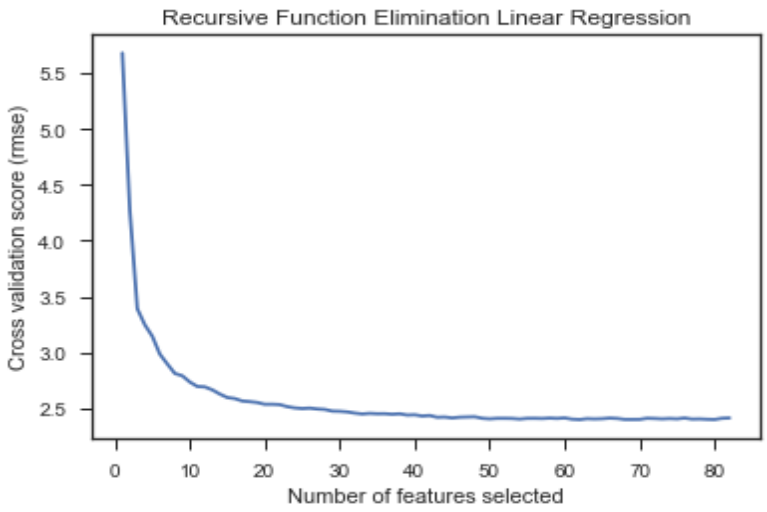
Model	**RMSE Through Online Submission**
AdaBoostRegressor	2.7847
Linear Regression	2.9297

The AdaBoostRegressor happens to be more precise than the Least Squares Linear Model because it can handle non-linear relationships. The AdaBoostRegressor is chosen as the regression model to predict poverty rates for United States’ counties.

NB: By using (nested) cross validation the whole dataset can be used for training and evaluation.

Recursive Feature Selection Linear Regression

After applying recursive feature selection or backwards selection the optimal number of features obtained for a linear regression model is 73 out of 82 features. The model is evaluated using cross validation and the RMSE evaluation metric.

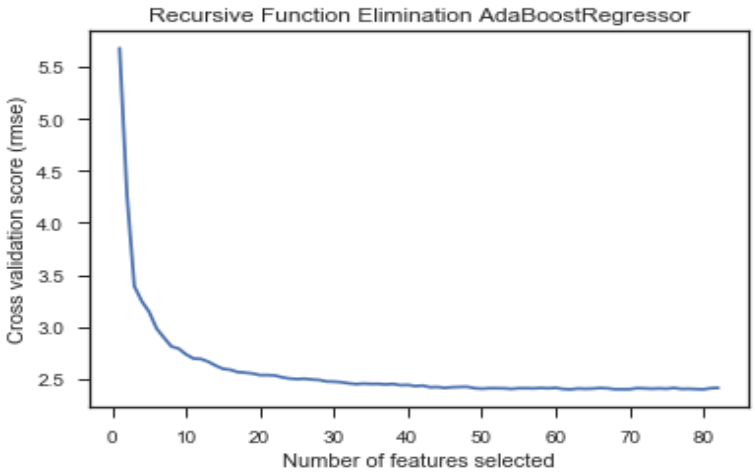


Nested Cross Validation AdaBoostRegressor Vs Linear Regression

Using nested cross validation, the AdaBoostRegressor is compared with the linear regression model. The lowest RMSE score is obtained with the AdaBoostRegressor.

Recursive Feature Selection AdaBoostRegressor

To improve the AdaBoostRegressor even more, the best features are selected using recursive feature elimination or backwards elimination. The model is evaluated using cross validation and the RMSE evaluation metric. The optimal number of features obtained is 62 out of 82 total features.

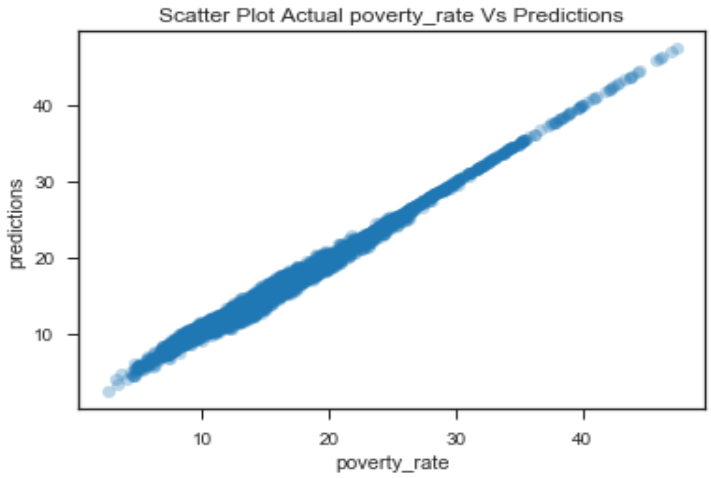


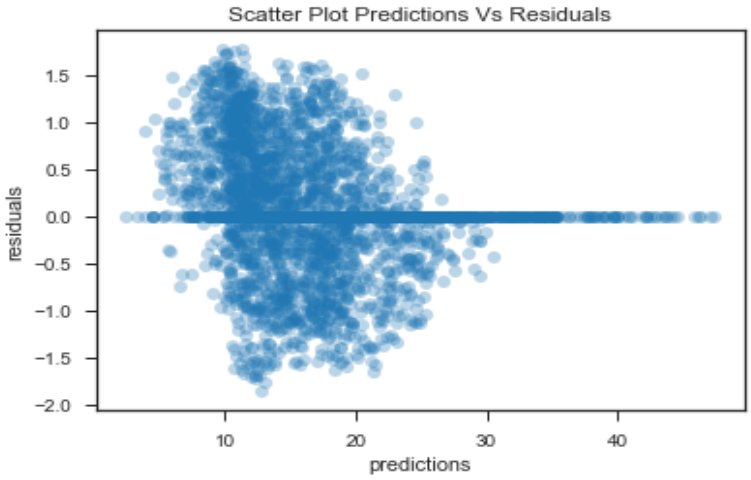
Analysis of Predictions and Residuals

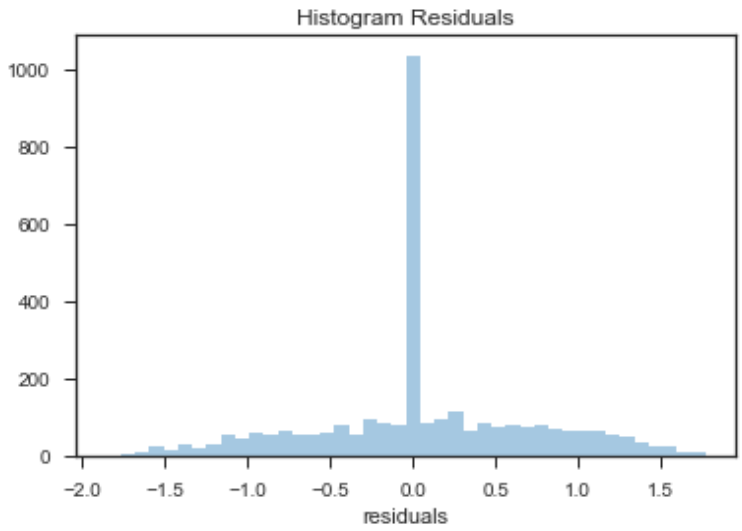
The Analysis of the quality of the predictions and residuals shows that the accuracy of the AdaBoostRegressor is high:

- The scatter plot 'Actual poverty_rate Vs Predictions' displays a nearly straight line.
- In the scatter plot 'Predictions Vs Residuals' and histogram 'Residuals' most residuals are near zero. However an effort can still be made to analyze the predictions with residuals non equal to zero.

However the AdaBoostRegressor probably slightly overfits the data because most of the residuals are equal to zero.







Conclusion

The Regression analysis shows that it is possible to build an accurate regression model to predict poverty rates of United States' counties with an RMSE of 2.7847. From the data exploration phase, it is clear that economical, educational, demographical and health related factors play an important role in predicting poverty. Finally, recursive function elimination combined with cross validation is used to determine the optimal number of features that leads to the best prediction.