Regression Analysis of Murder Rate by US State

Data Science Final Project

Jeremy McCormick

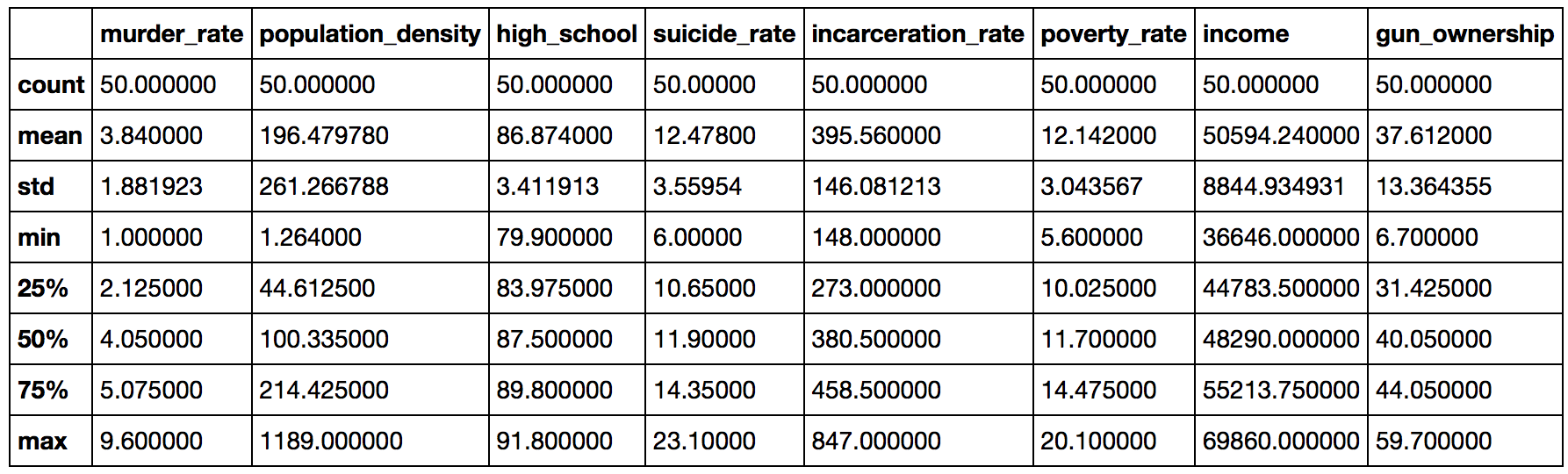
The purpose of this project is examining possible correlations between the murder rate in US states and a set of input statistical variables. These input variables include:

* Population density (number of people / square mile)
* Median Household Income ($)
* Poverty Rate (percentage of households below the federal poverty level)
* Percentage of the population that graduated High School
* Incarceration rate (persons incarcerated / 100,000)
* Suicide Rate (suicides / 100,000)
* Gun Ownership (percentage of households claiming to own one or more firearms)

The primary techniques used will be Linear and Polynomial Regression on a single input variable. Random Forest Regression will be used to effectively perform regression using all input variables in a single model. Leave One Out (LOO) cross validation will be used to test the predictive accuracy of the model.

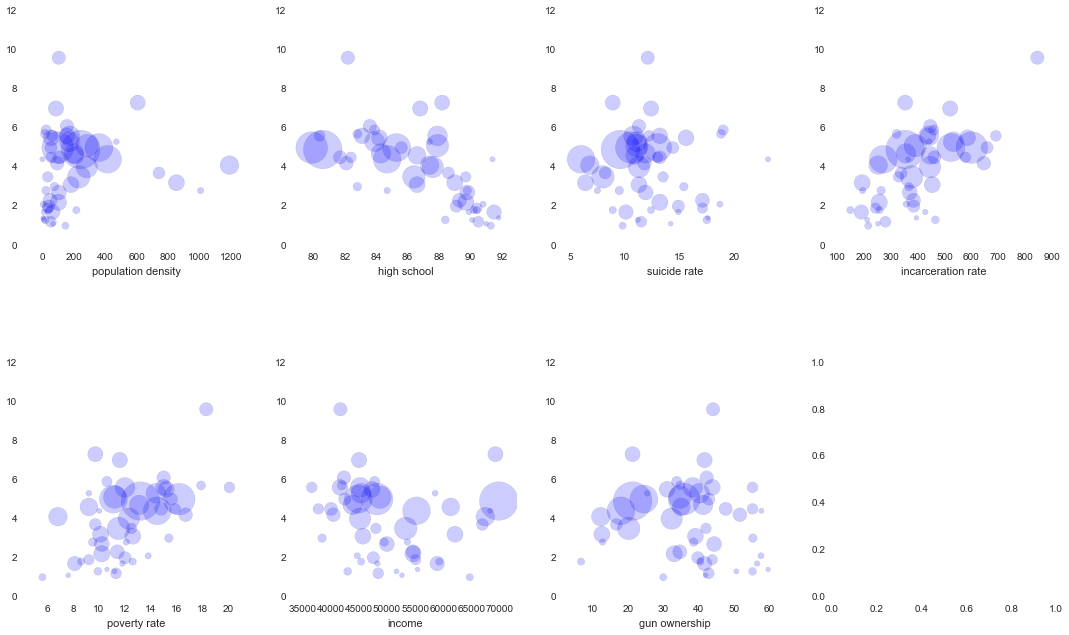
Data was obtained for individual features by “scraping” tables from online sources, primarily Wikipedia. Then this data was combined into a single Pandas DataFrame using the *merge* function. This resulted in a 50-row table with each row mapping the features and the murder rate onto a US state. Washington D.C. was excluded from the analysis.

The following table shows some of the statistical features of this dataset.

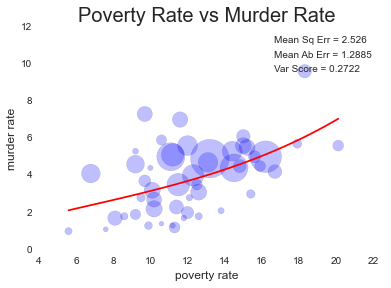


A set of Python functions was written to perform the regression and plot the results, as well as run LOO cross validation on each of the statistical models. These functions also computed the Mean Squared Error, Mean Absolute Error and Variance Score using the predictions and the true values. These values were used to benchmark the predictive accuracy of the models and the effects on the models of using cross-validation.

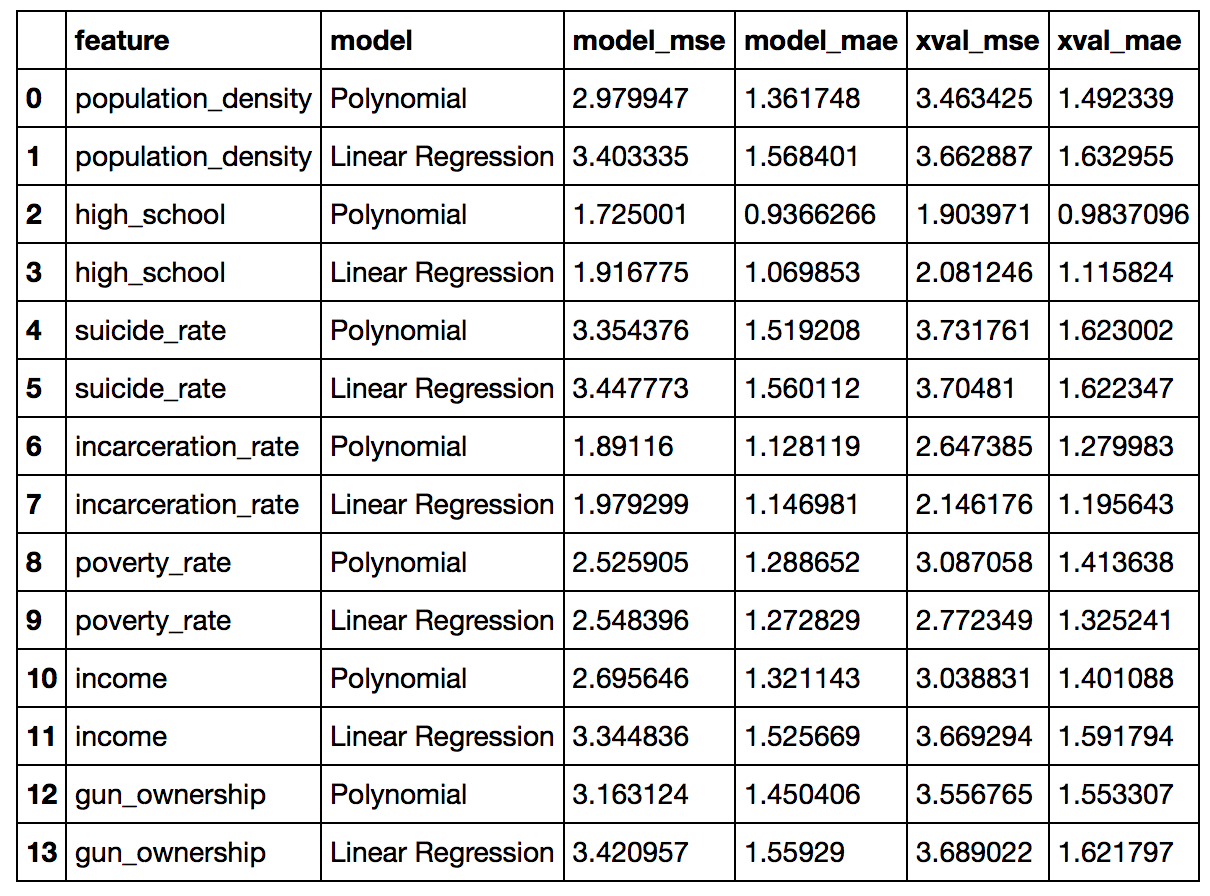
Initial data exploration was performed to identify features that appeared to have correlations. In the figure below, each input statistical feature is plotted against the output (murder rate) with the state’s population used to scale the size of the points.



Linear regression and polynomial regression were applied to each input variable separately. Below is an example showing the scatter plot of a feature against murder rate with the polynomial regression line in red.

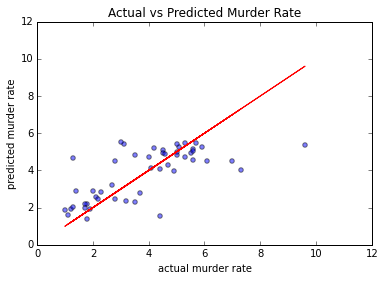


After performing linear and polynomial regression on each feature, cross validation was performed using the LOO algorithm. This fits the data using all but one of the data points and then predicts the value for that point. This can be used to predict the stability and accuracy of the model. The results of performing the cross validation are summarized below.



The mean deviation of the murder rate in the dataset was 1.5624. The table indicates that the most strongly correlated features have smaller Mean Absolute Errors than this, meaning the regression models have some predictive power in comparison to the mean, at least for these specific features.

Finally, the Random Forest Regression technique was used in an attempt to use all of the features in the model. This resulted in the smallest errors, with a Mean Absolute Error of 0.948.



In summary, correlations were found between murder rate and poverty rate, incarceration rate, and percentage of the population without a high school education. Both linear (line fit) regression and polynomial regression using a LASSO constraint resulted in Variance Scores and Mean Absolute Errors that suggested a statistical dependence. All statistical features were used in a Random Forest Regression to yield the lowest errors.

Certain features of the input variables hindered the analysis:

* Low number of data points (50) which meant that performing a LOO cross validation resulted in 2% of the entire data set being removed just from leaving a single data point out. Low statistics also meant that many machine algorithms would potentially perform poorly in their fit functions or would possibly over-fit the data especially when using larger degree polynomials.
* Strong correlations in the input variables, so multivariate analysis was not effective when using standard regression techniques.

In retrospect, this type of analysis of crime data might best be done on smaller geographical units such as cities or counties. Due to murder rates in these areas being quite small (zero for many), alternately violence crime rate could be analyzed instead, which has a strong statistical correlation with murder rate.