Final Project – Years of Potential Life Lost

DS 740 August 8, 2019 Mark E. Riley

Executive Summary

Introduction

The purpose of the analysis is to determine if we can accurately predict the Years of Potential Life Lost (YPLL) rate per 100,000 people in United States (US) Counties and the District of Columbia (DC) using the predictor variables in the dataset. This analysis will also assess which of the variables in our data are the most important in predicting the YPLL rate. The YPLL rate is a measure of premature mortality that gives death at younger ages, which may be preventable, greater weight.

The YPLL rate is of interest to the US Department of Health and Human Services (HHS) because it encompasses several other US government agencies such as the Food and Drug Administration (FDA), the Centers for Medicare & Medicaid Services (CMS), and the Centers for Disease Control and Prevention (CDC). Using this analysis, HHS could provide direction to its agencies to conduct research in US Counties with high YPLL rates to determine if it is possible to reduce the rate of premature deaths.

Years of Potential Lives Lost is calculated by subtracting the age at death from 75 for those persons whose age at death was less than 75. For example, a person who dies at age 25 has an YPLL of 50. Those persons who are aged 75 or older at death have an YPLL of zero.

Data Set Analysis

The data set consists of twenty variables and 3,192 rows representing each US County and DC, and summary numbers for each State and DC. We removed the 51 summary (State and DC) rows, as well as 291 rows where the YPLL rate value was either missing or indicated as unreliable, leaving 2,850 observations.

We removed variables that did not have predictive value, e.g. unique identifier and unreliable indicator, leaving 15 predictor variables and one response variable, YPLL.Rate. All of the remaining predictor variables and the response variable have continuous values. The predictor variables measure aspects about each US County in the dataset that may contribute to the YPLL rate.

Four of the predictor variables have missing values, including HIV.rate (629 observations), PctFreeLunch (15 observations), PctChildLiteracy (2 observations), and Rural (1 observation). We imputed values for the missing data points using the Random Forest algorithm.

The variables in the data set show low levels of correlation between each other. The highest correlations are between Physical.Inactivity and PctDiabetes (0.762), and PctFreeLunch and the YPLL.Rate (0.712).

Nearly all the variables in the data set were not normally distributed and contained outlier values when plotted in histograms and boxplots.

Data Mining Techniques

The first data mining technique we employed in the analysis is robust regression. We chose this method because of the number of outliers seen across the variables in the data set. Robust regression puts less

weight on outliers in the data without excluding them from the model. We used both the Tukey Bisquare and Huber methods for weighting outliers, two common forms of robust regression. To add breadth to the robust regression we considered various subsets of variables based on their importance as calculated by the bagging algorithm as seen in Figure 1. We started with a single predictor variable, PctFreeLunch, and continued to add one variable until there were ten models including the variables up to HIV.rate. The final model included all predictor variables.

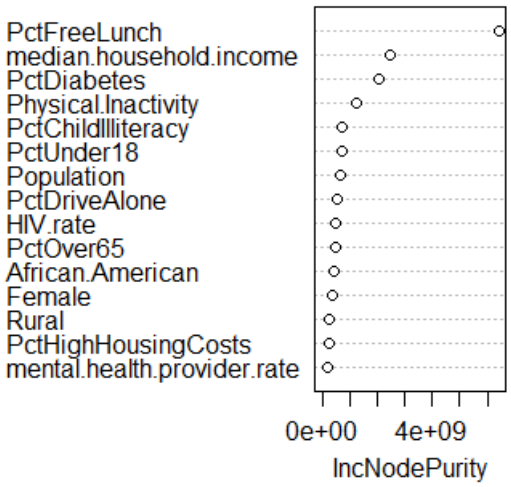

Figure 1

The second data mining method chosen was decision trees for regression, including the bagging, boosting, and random forest ensemble methods. These methods are a good fit for our regression analysis because our data has minimal predictor interactions.

Our bagging model uses the default value of 500 trees because analysis of the error rate shows that the out-of-bag error rate levels out at approximately 75 trees. We used all predictor variables because there is low correlation between the predictors. The boosting method is included evaluate if it is able to better fit the YPLL data than bagging. Our boosted model uses two tuning parameters during cross validation. The first is interaction depth where we use values between one and four. Because we have a fairly large number of observations we can go above an interaction depth of three. The second tuning parameter is the number of trees, of which we will try 2,000, 4,000, 6,000 and 8,000. Our shrinkage parameter is 0.0005 so we will need at least 2,000 trees.

We included random forests because, as seen in Figure 1, a handful of the predictor variables are more informative about the response variable than the others, such as PctFreeLunch, median.household.income, and PctDiabetes. The random forest model included five predictor variables per tree as the recommended number of predictors is the total number of predictor variables (15) divided by three.

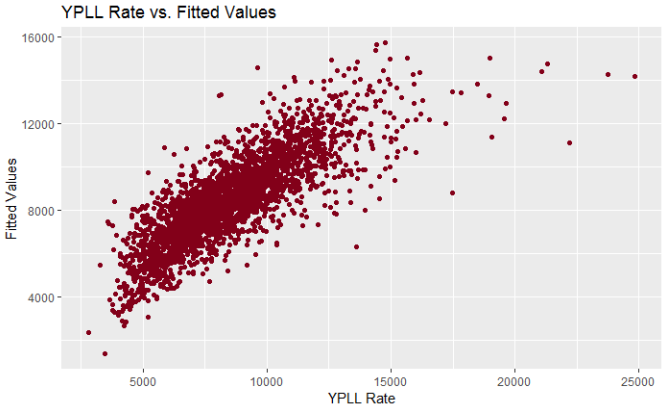

Figure 2

We chose to include a third data mining method of artificial neural networks for regression. As we can see in Figure 2, a plot of the fitted values from a linear regression against the response variable, the relationship is not exactly linear, which is a good fit artificial neural networks. We chose two tuning parameters during cross validation. The first is the number of hidden nodes, for which we used values between one and five to avoid too much variance in the model. The second tuning parameter is the decay weight, for which we used values between 0.2 and 3.0, incremented by 0.2 for a total of 15 decay rates.

Final Project – Years of Potential Life Lost

DS 740 August 8, 2019 Mark E. Riley

Conclusion

The model that performed best using single-level cross validation was random forest decision tree with the number of predictor variables equal to five at each node. This model used all predictor variables. The mean squared error (MSE) value was 1,764,288 and the coefficient of determination (R^2), or the amount or variation explained by the model is 71.24%.

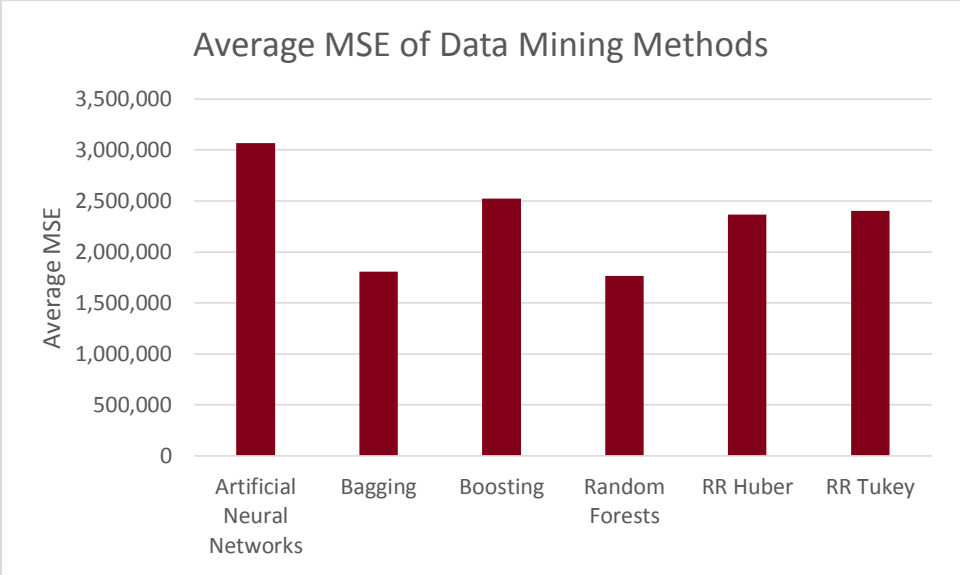

Figure 3

As we can see in Figure 3, the next best performing data mining method was bagging (as measured by the average MSE during single cross validation). The overall worst performing data mining method for predicting the YPLL rate was artificial neural networks.

When using double cross validation to assess the performance of all the selected models our MSE value is 1,768,827 with an R^2 value of 71.16%. These values are very close to the assessment values for single cross validation. During each outer fold of the double cross validation, the random forest data mining method had the lowest MSE value every time. This would explain the relatively close assessment values between the single and double cross validation assessment methods.

The US Department of Health and Human Services should direct its agencies to use the results of this analysis focus their research on the causes of premature death in US Counties and Washington, DC. As seen in Table 1, the variables with the highest importance for predicting the YPLL rate should be the primary focus to determine if there are opportunities for interventions to reduce premature death, especially in areas with high YPLL rates. Factors such as prevalence of diabetes, inactivity, and child literacy are potential areas where an intervention could be applied while continuing to monitor the long term effect on the YPLL rate.

Further analysis to determine if it is possible to more accurately predict the YPLL rate could include performing random forest regression using different numbers of predictor variables. HHS may also want to gather additional predictor variables, including social, financial, and physical factors such as tobacco use, sexual activity, air quality, water quality, and unemployment rate.

Table 1

Predictor Variable	% Inc. MSE
Percentage of children enrolled in public schools that are eligible for free or reduced price lunch	32.14146
Median household income	31.57641
Percentage of adults aged 20 and above with diagnosed diabetes	29.73947
Percentage of adults age 20 and over reporting no leisure-time physical activity	28.97209
Percentage of child illiteracy	23.34065
Percentage of population under age 18	21.40747
Percentage of households that spend 50% or more of their household income on housing	20.43357
Population	20.22905
Percentage of population African American	18.34953
Percentage of population over age 65	18.00786
Percentage of population living in rural areas	17.4943
Ratio of population to mental health providers	17.38126
Percentage of population that is female	16.9978
Rate of people aged 13 years and older living with a diagnosis of HIV per 100,000	16.18297
Percentage of the workforce that drives alone to work	12.6371

References

1. Kaggle 2017, US County Premature Mortality Rate, accessed July 30, 2019, <https://www.kaggle.com/royxss/us-county-premature-mortality-rate>.