**Predicting Heart Disease Rates in US counties**

Joseph O’Malley

**Overview/Business Problem:**

The goal is to predict the rate of heart disease (per 100,000 individuals) across the United States at the county-level. This prediction will allow hospital systems and governing bodies in counties with unknown rates of heart disease mortality benchmark their performance based on several socioeconomic, comorbid, and other features included in the dataset. The data is compiled from a wide range of sources and made publicly available by the United States Department of Agriculture Economic Research Service (USDA ERS). The dataset contained 33 variables and 3198 rows (with each row representing a unique county) in this dataset. I will predict if the rate is above the average heart disease mortality rate. I will also predict the continuous heart disease rate using regression models for each row of the test data set.

**Exploratory Data Analysis:**

This appears to be a fairly clean dataset that is should work well with some of the more complex models. It started with 34 columns (including the target) and 3198 rows. I started by looking at the column names/datatypes to find that a majority of the columns were already in percentage terms (i.e. percent adult obesity) for the prevalence of certain comorbid or demographic conditions in that particular county. There were three that were categorical, so I looked at the distribution of the values within each of these to see if/what needed grouped. I left them alone, as each group had a significant amount of observations (<2% of total) and created a unique column for each. Next, I looked for observations with more than half of the columns with missing, but none appeared. I then looked at correlations across the entire dataset to see which columns could be eliminated/grouped. There were only two ('Large-in a metro area with at least 1 million residents or more' and 'Metro counties with at least 1 million residents or more') that were highly correlated (>90% corr), so I removed one. After looking at distribution of the numeric columns, I looked for outliers using a boxplot visualization. Though there were several outliers (outside of the 1.5 IQR), they did not appear to be in error. I decided not to impute them and instead opted for min/max normalize. This normalization (min/max 0-1) compressed several columns, but about half remained unchanged, as they were already in percentage terms. Bringing these all to the same scale will help many of the models run more efficiently/effectively. Lastly, I created split the target ('heart\_disease\_mortality\_per\_100k') as a binary at the mean. This will be used to assess the models that use classification and predict whether counties with unknown rates are above or below average. There is also a copy that keeps the target as a continuous variable for the regression models.

**Models:**

I used 7 different classification model types (Table 1.1). I applied GridSearch on several of these the base models, which allowed me to tune several parameters at once and finds the optimal combination to apply to each of the models. I added adaptive boosting to the Decision Tree, which reweights the training set in iterations to include more of those observations that were misclassified in the previous iteration so the model can "learn from its mistakes". Bagging was the other method I used to fits several decision trees, each on random subsets of the original dataset, and then aggregate their individual predictions (by voting) to form a final prediction. This reduces the variance by introducing randomization into its construction procedure and then making an ensemble out of it.

To predict heart disease mortality rates as a continuous outcome, I used 3 different regression models (Table 1.1). Two apply the same techniques as those run for classification (Extremely Random Forest, ANN). The third model (LASSO) weights the different features by explanatory power, reducing some features to zero.

**Table 1.1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | Accuracy | Recall (of TP) | AUC | R^2 |
| KNN | 0.916 | 0.9 | 0.883 | NA |
| DT | 0.798 | 0.77 | 0.775 | NA |
| DT (w AdaBoost) | 0.899 | 0.9 | 0.897 | NA |
| DT (w bagging) | 0.824 | 0.82 | 0.821 | NA |
| Random Forest | 0.872 | 0.85 | 0.855 | NA |
| SVM Linear | 0.832 | 0.82 | 0.826 | NA |
| SVM (rbf) | 0.924 | 0.91 | 0.923 | NA |
| Extremely Random Forest | 0.894 | 0.89 | 0.893 | NA |
| ANN | 0.886 | 0.89 | 0.887 | NA |
| (Regression) ANN | NA | NA | NA | 0.683 |
| (Regression) Extremely Random Forest | NA | NA | NA | 0.758 |
| LASSO | NA | NA | NA | 0.635 |

**Selection/Recommendation:**

Area Under the Curve explains the tradeoff between capturing all of the true positives and getting more noise (in the form of False Positives). I also used True Positive recall (how many of the total counties over the average rate of heart disease were predicted) as an important measure of model performance because these are the at-risk counties that we are concerned with. Accuracy was my third measure of model performance.

The best performing model across all 3 measures (Accuracy, TP recall, and AUC) was the Support Vector Machine (rbf) using grid search to tune the model to the optimal parameters. The model achieved an accuracy of 92.4% and True positive recall of 91%. Cross validation averaged 0.895 with a range of .871-.910, which a is below that of the model so we may expect this model to perform in this range when exposed to new data. As seen in the chart (above) it also had an AUC of 0.923, which explains the relationship of True Positives to False Positives.

Of the regression models, the Extremely Random Forest explained the most variance of any model with an R^2 of 0.758. However, I would recommend the LASSO regression model for ease of interpretation. It cut the number of features in half (from 56 to 23) and still achieved an R^2 of 0.635. This may also help them focus on patients/habits may be best to focus on (though the true underlying cause should be analyzed).

I recommend using the SVM (rbf), as it had the best performance across all three of the measures. It caught the highest number of True Positives (counties with heart rates above the average). It also performed within a relatively tight range (.871-.910) in cross validation, so we can expect this model to perform consistently when exposed to new data. This model is fairly “black-box” in nature, however. Using the LASSO regression model in combination will give you a prediction of where a county falls along the spectrum of heart rate disease. Additionally, it provides the features that had the most explanatory power. This may help inform hospital and governing bodies in creating a strategy from a clinical/public health perspective.