Nathaniel Kent & Brian Greenberg
Data Mining
Project 2
April 28, 2017

Using Risk Factors and Access to Care to Predict Likelihood of Diabetes by County

I. Introduction:

The goal of this project is to choose our own dataset and use the tools we learned in class to investigate the data. In this project, we will investigate a dataset of common risk factors for disease and attempt to make predictive determinations as to whether a county in the United States has a higher than average population with diabetes. The classification approaches we will use in the experiment are Neural Nets and Random Forests.

II. Data:

Our data comes from the catalog.data.gov/ database. It holds information about the 3000+ counties in the United States. For each county, this dataset holds various health information including infant mortality rate, premature birth rates, and obesity rates by county. The data also includes information on the ethnic and racial diversity of the county and its location and zip code. In this dataset there were over 30 attributes per county. Some of the attributes were not very useful, so while preprocessing the data, we eliminated any attributes that were not useful to our experiment. Now our experiment uses 23 attributes instead of over 30 (Attributes such as zip code, town ID, and town name were among the attributes removed).

While preprocessing our data, we use the imputer to fill missing values. There are many counties that did not fill out all of the information. For this reason, we use the imputer to fill via ‘median strategy’ so that we can still make use of the data. This however, added noise into our dataset and led to somewhat lower predictive accuracy. Our project2.Part2.py file however, uses a dataset where we simply removed any example with incomplete information. We decided to do this in order to see if the accuracy of our output improves without the examples with missing data. In project3.Part3.py, we use the part2 dataset, but remove some more of the attributes and left only the 6 non-class attributes that we deemed to be most important based on the results of the random forests. In this experiment, we use diabetes rates as our class attribute. The class attribute is a binary 1/0 value, representing whether the county has an above rate of diabetes among their citizens. If the rate per county was over the national average (9.3 percent) we would give that county a value of 1. Otherwise it would get a 0 for below average rates.

III. Methodology:

The methods we will use in our experiments are Neural Nets and Random Forests. The reason why we chose to use Neural Nets and Random Forests is because we thought Random Forests would be useful in isolating the attributes that are helpful. We chose to work with Neural Nets because they are very useful in working with large sets of data because they learn as more inputs are added.

The default setup used for the Neural Nets uses logistic activation with maximum iterations set to 100. The variations for the Neural Nets model used differing activation methods to see if the results changed. There were three variations used. ‘Relu’ Activation, adaptive activation, and inverse scaling learning rate all used differing activation methods while the other variation exploited ‘early stopping’ mechanism of ANNs.

The default setup used for the Random Forest had only 5 estimators, had the random state set to zero, and used entropy as criteria for the model. In variations for the Random Forests, we decided to vary the number of estimators to see if it would produce more accurate results. The two variations used for random forests set the number of trees from 5 to 50 to 100. We also tested the predictive accuracy of a single decision tree.

IV. Results:

The results of this experiment show that Neural Nets have the highest accuracy scores when using the default logistic function and max iterations set to 100 (roughly the default parameters). The highest accuracy recorded by this method was 74(85 in part 2-3) percent, while the variations all ranked lower. The least successful activation method used the ‘adaptive learning rate’ with the lowest individual score of 55 percent. The p-values for the Neural Nets show that p-values are significant when comparing the logistic function to both the ‘relu’ and ‘adaptive’ learning rates. The ‘relu’ activation was significantly different from the method using an early stopping. Early stopping was also significantly different from the adaptive learning rate. These p-values all show that there is a very clear difference in result when using these two methods. The confidence interval for the Neural Nets show that the first approach using the logistic function has the highest mean value and also has the smallest confidence interval. This means that it will be more consistent in its results.

The results of the experiment for Random Forests show that the method that had the best results was the random forest with the number of trees set to 100. This approach had an average score that was slightly less than 75 percent (86 in part 2-3). The p-values for the Random Forest approaches showed that there were only significant values against the Random Forest that used 100 trees. It was significant against every other approach used, which indicates a difference in value between this approach and the other ones. The confidence intervals show that the approach that sets the number of trees to 100 has the highest average score. It also has the lowest standard of error within the confidence intervals. This indicates that it will obtain its results more consistently across many runs.

Our experiment had three parts to it. The first part was described above in the Data and Results. The second part of our experiment used all of the data within the .csv file that did not have any missing values. This was done in order to see if you could achieve more accurate results for the data. After running part 2, which eliminated missing data entries, the accuracy of all methods improved significantly. Part 3 takes part 2 a one step further by removing more irrelevant attributes from the data. This approach obtained relatively similar results to Part 2, which may suggest that those attributes (which included number of dentists and some other variables) were not relevant to the experiment.

V. Observations:

Based on the results, the default method of using the logistic approach was objectively better than the other approaches. This could be because the other approaches did not match to the dataset as well as the logistical approach. The relu approach is meant to be used on ‘linear units’. Because not all of our attributes have ‘linear’ values, this may cause an inaccuracy in the results. Changing the learning rate from constant to adaptive or inverse may cause more errors in the data because these approaches decrease the learning rate. This can make these approaches take longer to reach the same level of accuracy.

The Random Forest had the best results for the method that used 100 trees because there were, essentially, more variations to choose from. Having fewer trees means that the model has less data to use and ‘learn’ from. Our model that used the most trees was most successful because of this idea.

This experiment shows that the Random Forest Classifier is better for these types of datasets than the Neural Nets. The Random Forest obtained on average higher accuracy scores with lower margins of error around those scores. This indicates that Random Forests are more effective with varying data types and integers that vary in size.

VI. Next Steps:

There are still many experiments that should be run on this dataset. We only used diabetes as the class attribute in the example, but there were many other attributes that would have been equally interesting as the class attribute. This dataset contained information on other diseases and illnesses as well. A further continuation of this experiment could focus on predicting the likelihood of the other diseases by using them as class attributes instead.

Additionally, we could expand our learning model to attempt predicting the exact percentage of diabetes within the population, rather than just telling us if the percentage will be above average.