**Data Mining Project 2**

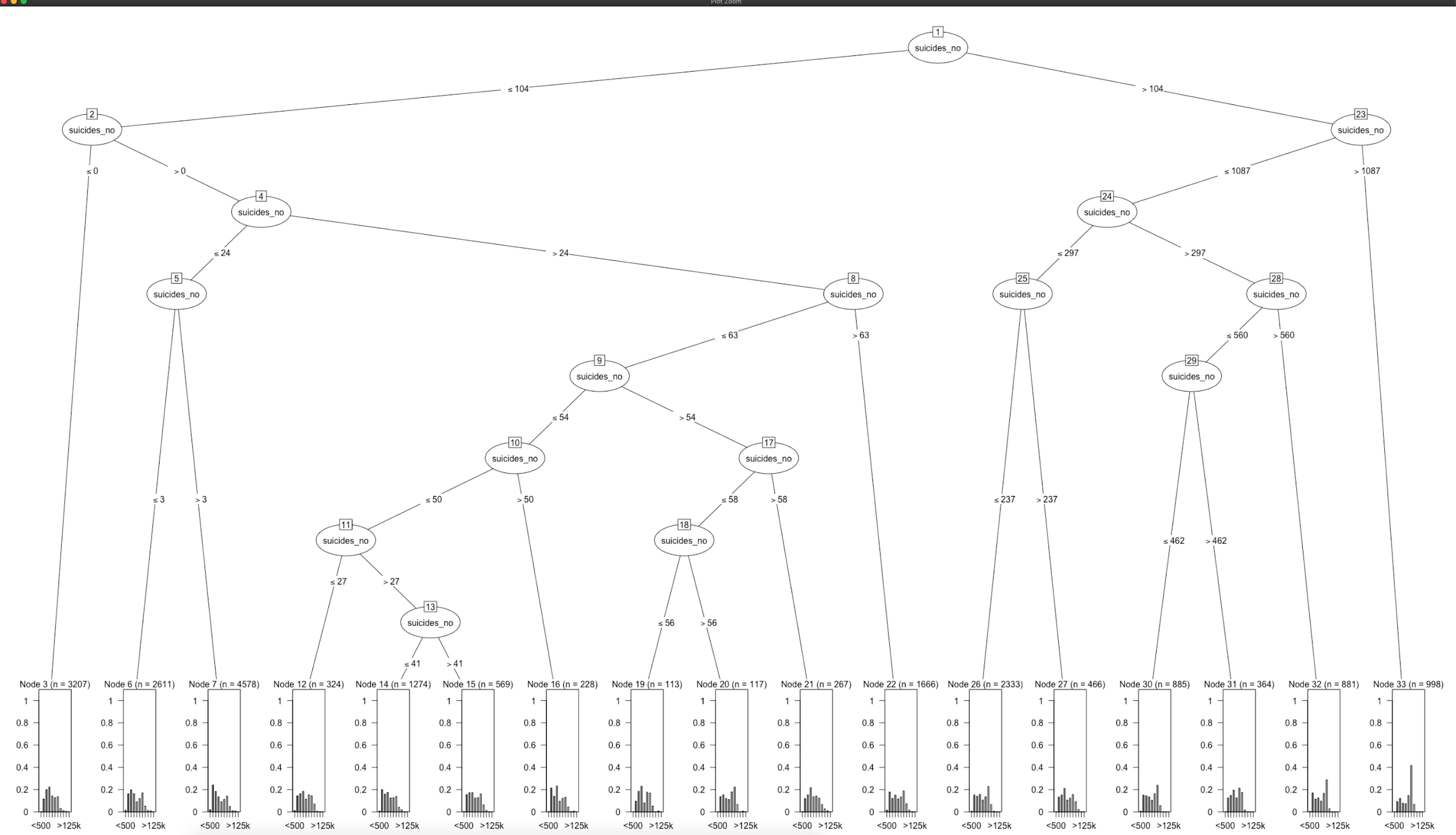
By:Michael Dichiara & Steven Warwick

Setup Phase

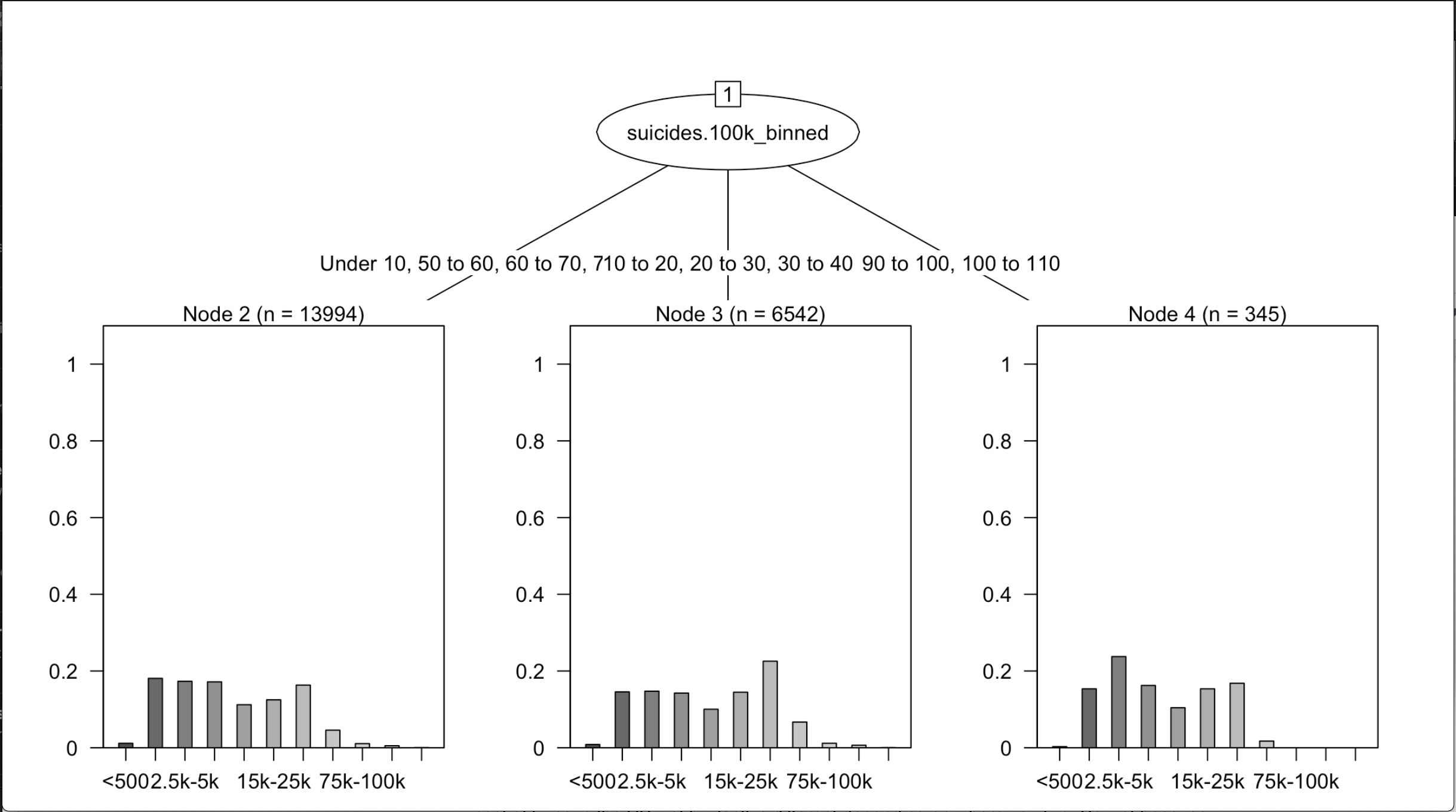
To begin the set up phase, we had to bin a new column based off of the gdp per capita in the suicide dataset. This column was used later in evaluating our baseline performance. The test set and train set had to be made for the two fold cross validation. We decided to perform cross validation through a sample z test on the sex column and the gdp\_per\_capita\_binned column to see if it was indeed balance. Every single value in both of those columns were balanced and was determined by the z-value result. Therefore, the data did not need to be balanced between the train and test datasets. Baseline performance was determined by evaluating which value of the gdp\_per\_capita\_binned had the most occurences of suicides. In our case it was the gdp per capita range 25k to 50k with a total amount of 1,652,154 suicides. We then divided this number by the entire amount of suicides in the dataset giving us 32.63%. The baseline model would assign all records of suicides to the gdp range 25k to 50k with an accuracy of 32.63%.

Modeling Phase

First we tried to use the CART model to model the suicide deaths. We ran into problems with it such as it not branching enough. We have two different variables which were GDP and suicides. When we would run it with CART we would only get a handful of branches which just did not feel right. We then attempted using the Random Forest model. The results we had for the plot graph were not helpful but the predicted table did return some interesting results like how the table would count every instance of a suicide then place the gdp per capia range it fell under below. It was interesting but was not overall helpful with what we were working on. Next we tried C5.0 which gave us very nice branching. As this gave us the kind of model we were looking for, we decided to use it. This tree was made from the column values gdp\_per\_capita\_binned and suicides\_no. Each node shows the number of suicide instances that were classified under the gdp per capita range they fell in. The range that had the most occurences in this case is 25k to 50k range similar to our baseline model. One minor problem we should point out is that R does not let you zoom in on a model. If you want to view it in R you will need to plug the computer into a wide screen tv. For convenience we have included a picture of it below.



We also had another C5 chart that turned out pretty well which is the image below. It is using the columns gdp\_per\_capita\_binned and suicides.100k\_binned. Each node represents the amount of occurences of the suicides.100k\_binned ranges based on the gdp range they were classified under. The gdp per capita range that had the most suicide ranges is harder to dictate from these bar graphs but based on the 2 largest n= valued nodes have 25k to 50k as the top or close to top number of occurences.



Evaluation Phase

The model does show that lower areas with lower GDP do have a higher suicide count. That being said I also believe that we do not have enough variables to make a blanket statement of areas with lower GDP have a significantly higher suicide rate. I would expect to see a much sharper drop toward the end of the GDP but we just don’t see that. I believe to accurately model this we would need to include info such as time and location. It may be that it is a GDP drop that would increase that amount instead of just how low it is. For example the cost of living in the US just by itself is different depending on where you live. So just because one country has a lower GDP than another does not mean that the quality of life is not any different.

For the cost analysis the confusion matrix came out kind of messy. The diagonal is true positive (so if it says Actual: <500 and Predicted: <500 that is the diagonal), the row is false positive and the column being false negative, there is no true negative. The cost of a suicide to a male is about $48,000 with a cost to the economey of somewhere between $1,300,000-$850,000 per person. For easy math let's put it at $1,100,000. The average cost of a therapist is $20 - $250 per session. So for easy math we will say that it is $200 per session and they go 50 weeks out of the year giving us $10,000. True positive occurs when we predicted it would happen so action was taken and so they did not commit suicide, so the price is -$1,090,000(negative being money not lost). False positives would be when we predicted it so they received therapy but they would not have anyway, so the price would come out to $10,000. A false negative would be if it was not predicted so they did not receive therapy but did commit suicide which would give us the number $1,100,000. True negative would be if it was not predicted and they did not commit suicide the price would be $0. The justification for these prices is <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5061092/> from table 1 and 2. The price for therapy was just a quick google search but the main point is it is much cheaper than that person dying. I can not really put this into a matrix nicely so I'm going to the cost of every part multiplied by the cost I gave each part summed together for each of the different brackets. <500 = $820,000. 500-2.5k = $1,214,460,000. 2.5k-5k = $684,270,000. 5k-10k = $1,029,780,000. 10k-15k = $7,400,000. 15k-25k = $101,480,000. 25k-50k = $1,342,050,000. 75k-100k = 740,000. 100k-125k=420,000. >125k= 20,000. From the numbers shown here it is clear that this would not save money for a country. Part of the problem is that the price for false negative and true positive are very close so you can not really afford any false negatives. At the same time, while I did put a price on someone's life in terms of for the economy, the price for keeping people happy could also have the economy boon. The best model we had was the C5.0 model where we used the columns gdp\_per\_capita\_binned and suicide\_nos because it showed the best visualization of what we were targeting for.

Contributions

Michael Dichiara performed the setup phase with the sample z test, the new gdp\_per\_capita\_binned column and determined the baseline performance. He also performed the C5.0 models and the random forest model. Also wrote half of this report and applied the clustering algorithm to our code to see if the results would be helpful.

Steven Warwick performed the CART model, half of the modeling phase report most of the evaluation phase report and the cost evaluation

John Yucetas did not contribute anything to this project. He was supposed to perform the random forest model but we did not hear from him. This was definitely a communication and a participation problem with this group member.