BIG DATA ANALYTICS (CSCI -720)

Homework-03 **Name: Rajkumar Lenin Pillai**

The code which computes the cost value for each threshold and generates all the plots is present in file q1.py

Question-a.)

Solution:

If two different speed thresholds have the same lowest misclassification rate then the tie can be broken by using the lowest among the two speed thresholds. The best threshold found till now must be updated to new threshold only if the new threshold which is being considered is strictly less than the previous best threshold. As we are considering to maximize the public safety so having a lower threshold would result in people driving in much lesser speed as compared to the threshold with higher speed.

Question-b.)

Solution:

If two different speed thresholds have the same lowest misclassification rate then the tie can be broken by using the highest among the two speed thresholds. The best threshold found till now must be updated to new threshold with higher speed as we are trying to maximize the trust that public have in the police officers. So drivers who are not intentionally speeding will not be pulled over if we have a higher threshold.

Question-c.)

Solution:

Threshold value computed as the best threshold was 62.5 (mph).

Question-d.)

Solution:

Guess was that by changing the cost function such that the cost function will be equal to (2 x the number false alarms + 1 x number of missed speeders) the threshold will increase.

By changing the cost function such that the cost function will be equal to (2 x the number false alarms + 1 x number of missed speeders) increased the threshold value by 0.5 (mph) and the new threshold value was 63.0 (mph) .

Guess was that by changing the cost function such that the cost function will be equal to (1 x the number false alarms + 2 x number of missed speeders) the threshold will decrease.

By changing the cost function such that the cost function will be equal to (1 x the number false alarms + 2 x number of missed speeders) decreased the threshold value and the new threshold value was 57.5 (mph)

Changing the cost function to the temporary cost function made sense since using the first version of changing the cost function such that the cost function will be equal to (2 x the number false alarms + 1 x number of missed speeders) the threshold increased which means the algorithm has increased the threshold because the no of false positives have increased so it means that this temporary cost function can be used when we want to increase the trust that public have in police officers so that police officers will not pull up the drivers who are not intentionally trying to speed.

Similarly for the other temporary cost function which is (1 x the number false alarms + 2 x number of missed speeders) , the threshold decreased which means algorithm has decreased the threshold because the no of false negatives have increased so it means that this temporary cost function can be used when many reckless drivers who were trying to speed will be missed if threshold was high.

Question-e.)

Solution:

The first temporary cost function can be decomposed as follows:

temporary cost function =(1 x the number false alarms) + (R x number of missed speeders)

In the above equation the first term "(1 x the number false alarms)" is the objective function and "(R x number of missed speeders)" is the regularization term where R is any positive constant.

The second temporary cost function can be decomposed as follows:

temporary cost function =(R x the number false alarms) + (1 x number of missed speeders)

In the above equation the first term "(1 x number of missed speeders)" is the objective function and "(R x the number false alarms)" is the regularization term where R is any positive constant.

The regularization penalizes the threshold value being computed. For example in the first temporary cost function increasing the value of R will decrease the threshold and increasing the value of R in second temporary cost function will increase the threshold. The motive behind computing the threshold must be known in advance in order to choose the temporary cost function whichever is applicable. If public safety is important choose the first temporary cost function or if the public trust on police officers is important choose the second temporary cost function.

Question-f.)

Solution:

The no of aggressive drivers let through for the given data set using the first cost function is 102.

Question-g.)

Solution:

The no non-reckless drivers that would be pulled over for the given data set using the first cost function is 11.

Question-h.)

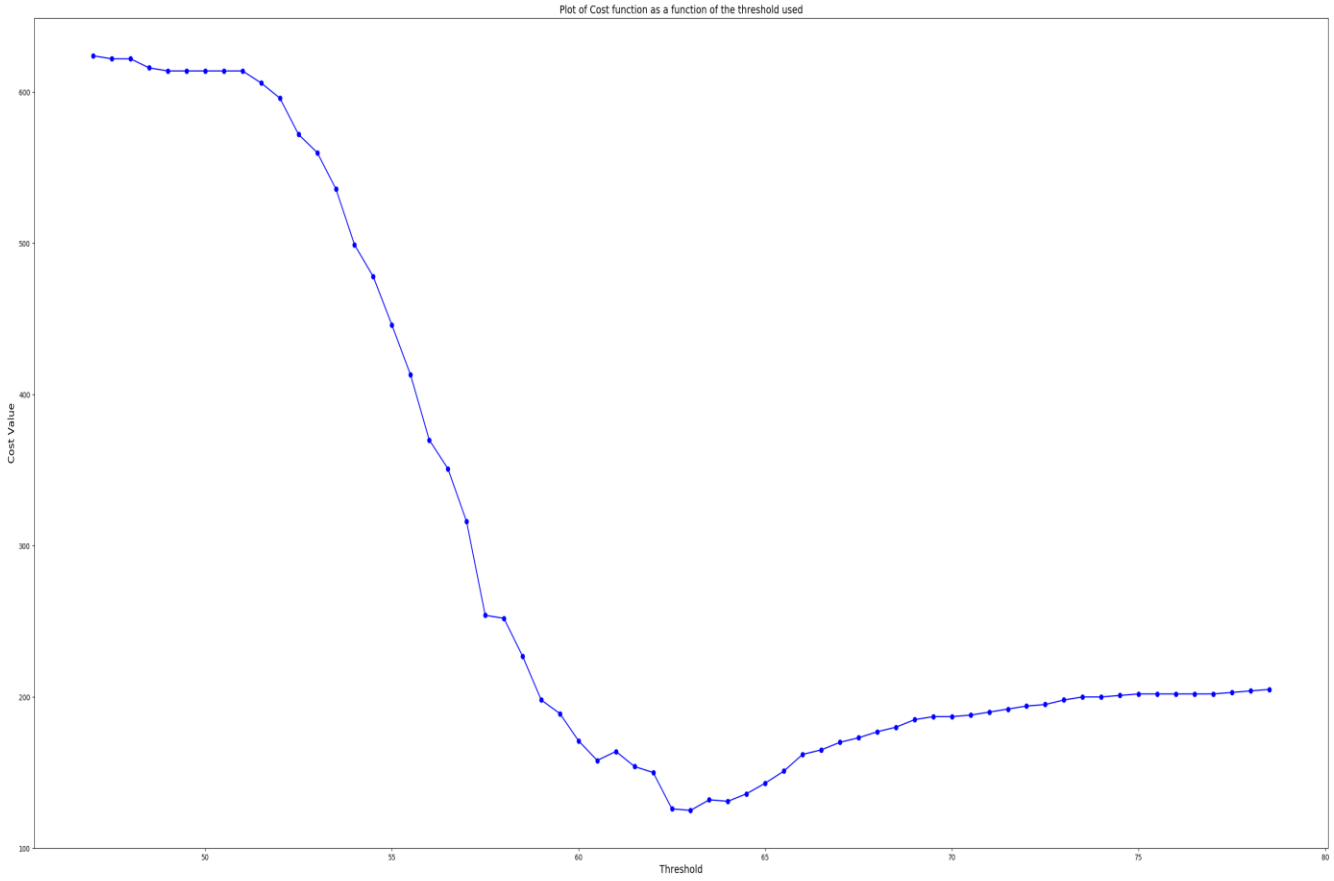
Solution:

The threshold for the given data set using otsu's method is 61.0 (mph) and the no of aggressive drivers let through for the given data set using the first cost function is 125 and the no of non-reckless drivers that would be pulled over for the given data set using the first cost function is 40.

Question-i.)

Solution:

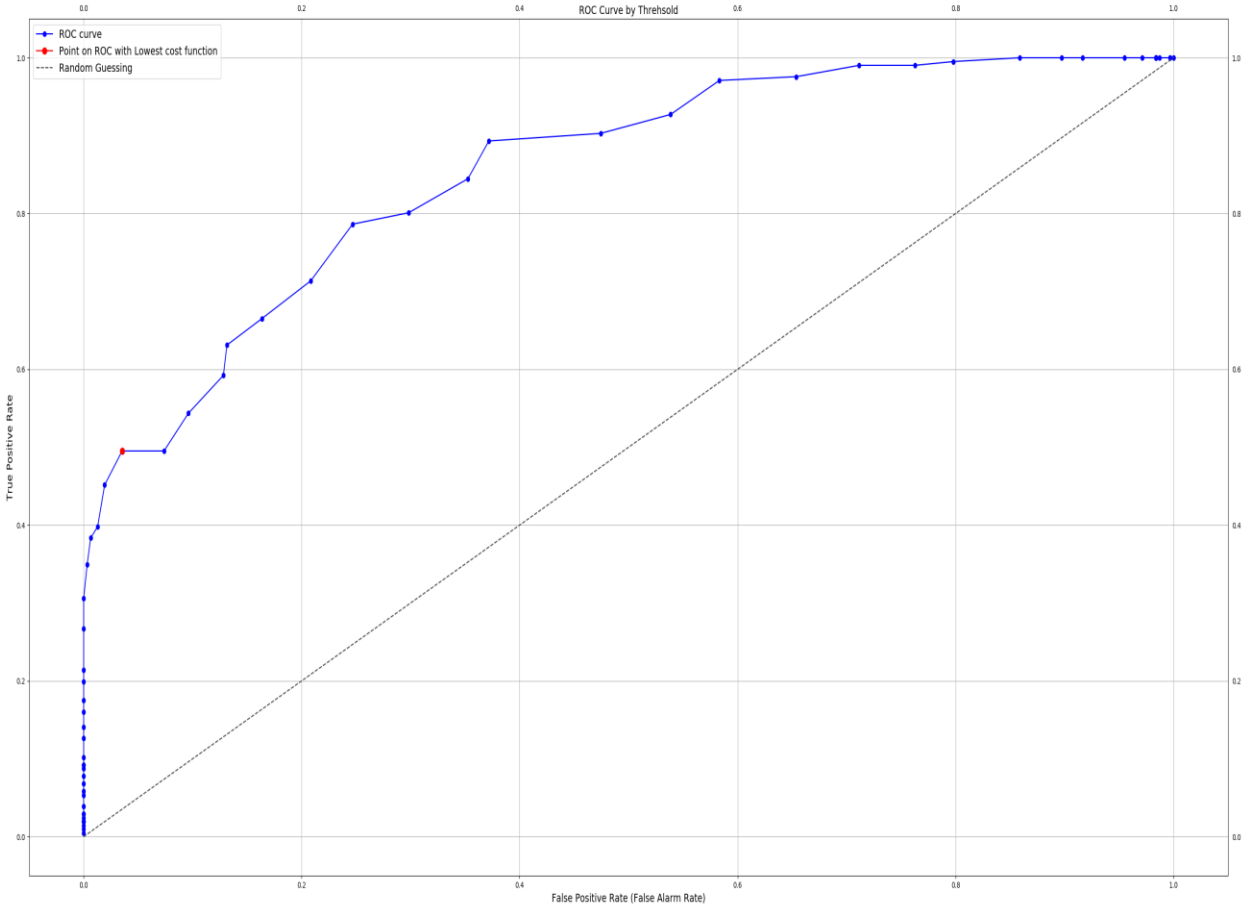
Plot of cost function as a function of the threshold used:-



Question-j.)

Solution:

ROC curve



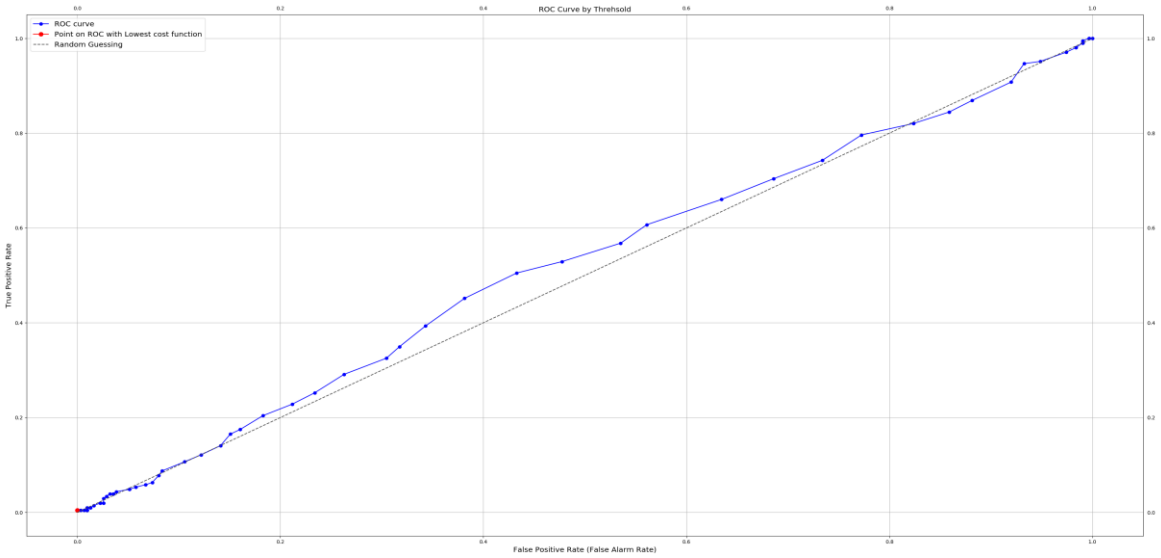
Question-k.)

Solution:

One dimensional classifier can be used with multidimensional data by combining all the variables in a multi-dimensional data into single variable by taking a cartesian or dot product and the possible values of this variable will be combination of all possible values of each feature and then use a one-dimensional classifier on this transformed data.

The challenging part was decomposing the temporary cost function and understanding the regularization term being used.

Initially some mistakes were made like initially the threshold computed by the algorithm was 78.5 mph which cannot be correct answer but later on it was found that after rounding the speed to 0.5 mph and sorting those values the corresponding column of aggressiveness of drivers also needs to be sorted in a likewise manner. Without sorting the aggressiveness values the ROC curve obtained was as follows:



Output of program without sorting the aggressiveness values:

```
Best threshold 78.5
NO of aggressive_let_through:   1
NO of non-reckless_pulled_over:   0

Process finished with exit code 0
```