Principles of Data Mining (420)
Spring 2022, Homework 03

Mohammed Mehboob
(mm2260@rit.edu)

March 3, 2022

ROC Curves for all attributes, except Hair and Tail Lengths:

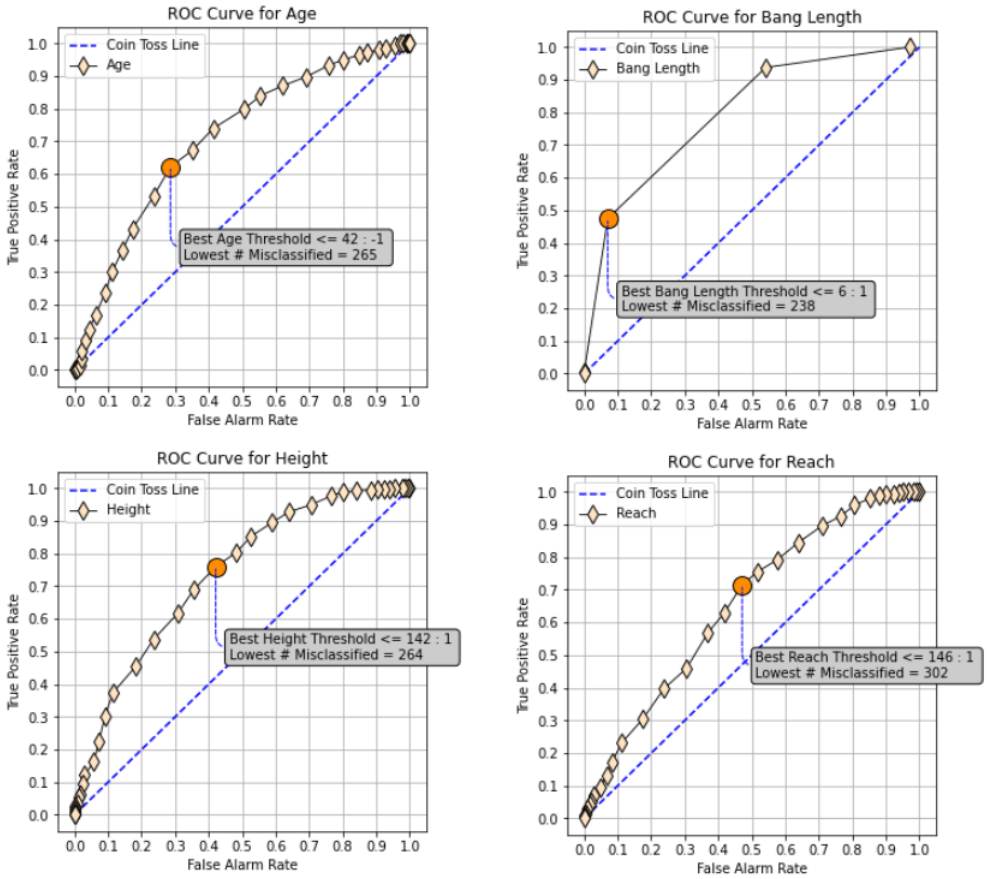

Figure 1: ROC Curves for Age, Height, Tail Length, Bang Length and Reach

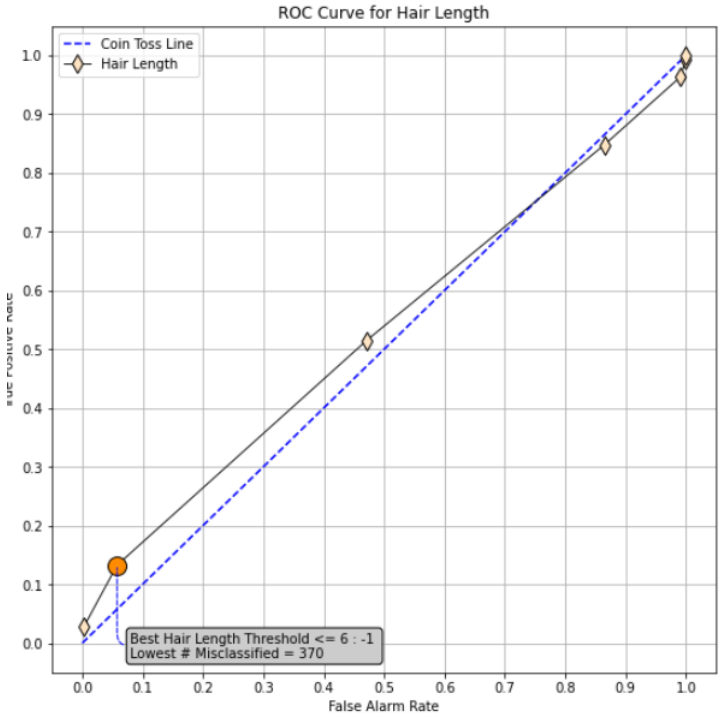

Figure 2: ROC Curve for the attribute Hair Length

The ROC curve for the Hair Length attribute was quite interesting. I initially thought that I'd made a mistake, since the curve dipped below the coin toss line. However, after a spending some more time trying to figure out if I did something wrong, I have concluded that it's probably the correct curve (or mostly, at least).

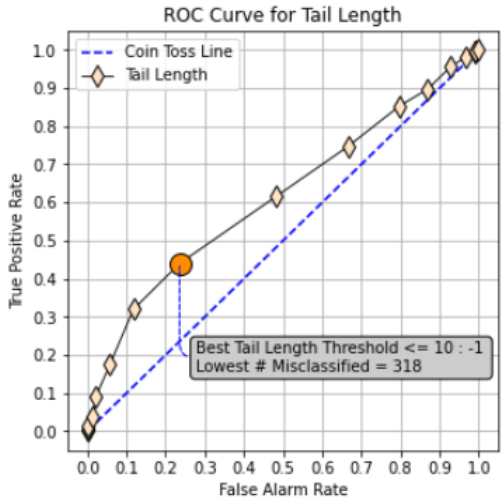

Figure 3: ROC Curve for the attribute Tail Length

The ROC curves for each attribute plotted on the same graph, for purposes of comparison with each other:

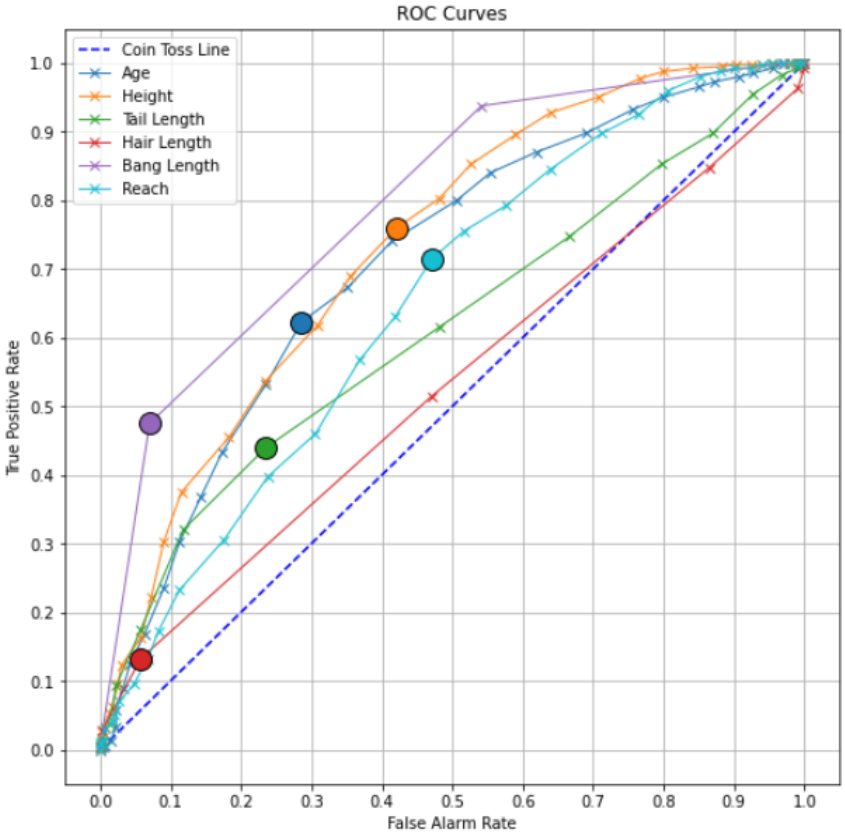

Figure 4: ROC Curves

On using the threshold-finding algorithm from previous homeworks, which finds the optimal threshold such that the total number of misclassifications is minimized gives us the information that the **"Bang Length" attribute would be the optimal classifier (judging just by the number of misclassifications**.

```
if( Bang Length <= 6 ):
      class = Bhutan
else:
      class = Assam
```

However, now that we have to depart from our notion of optimality being purely based on our "badness" or rate of misclassification, a good way of finding the best threshold based on our desired false alarm rate or true positive rate would be through examining the ROC curves.

I proceeded to add a vertical line on the graph of ROC curves denoting the False Alarm Rate being 0.09. Now that allows me to have a sense of what I'm looking for: as I have a reference to go off of, we are no longer looking at things

in two dimensions, but only one. Instead of comparing the entire curves, we can just compare each attribute by comparing the points where the ROC curves of each attribute intersect with the vertical line denoting the false alarm rate being 9%.

Once we do so, we can clearly see that the "Bang Length" attribute has the best true positive rate for our given false alarm rate, and it also happens to do fairly well with the total number of misclassifications. **So therefore, if we want a classifier with 9% false positive rate, the best one out of all the classifiers we have is the classifier by "Bang Length".**

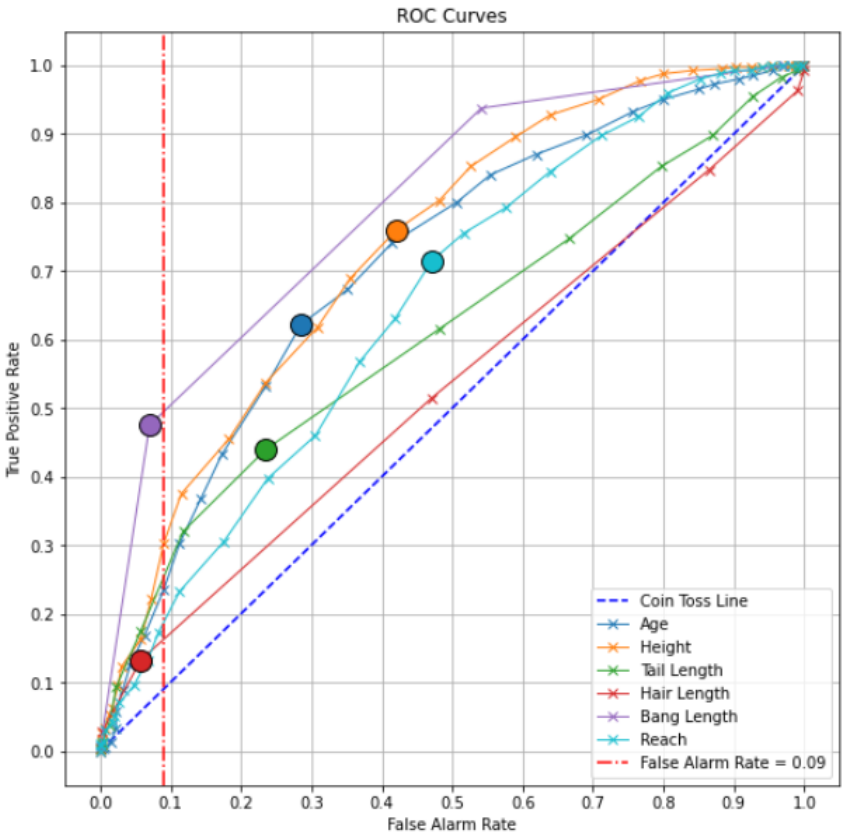

Figure 5: ROC Curves (9% False Alarm Rate Highlighted)

We'd want to use a very low false alarm rates in situations where it is likely that the observers will grow tired of all the false alarms and develop alarm deafness. Something like this has the potential to be especially devastating in a medical environment, perhaps. Wherein a patient might be in need of immediate assistance, however their caretakers have alarm deafness.

Similar to the previous part, where I compared each of the attributes by projecting the 2D ROC curve to a 1D respresentation by taking only their intersections with the vertical line, we can do something similar with the True Positive rate and horizontal lines.

Amongst all the curves which intersect with our horizontal line, the attribute "Bang Length" seems to offer better false alarm rates. Rather than figuring out which point on the ROC curve graph can I pick which is most towards the top-left corner, I am seeing myself first consider what true positive rate I would want to have.

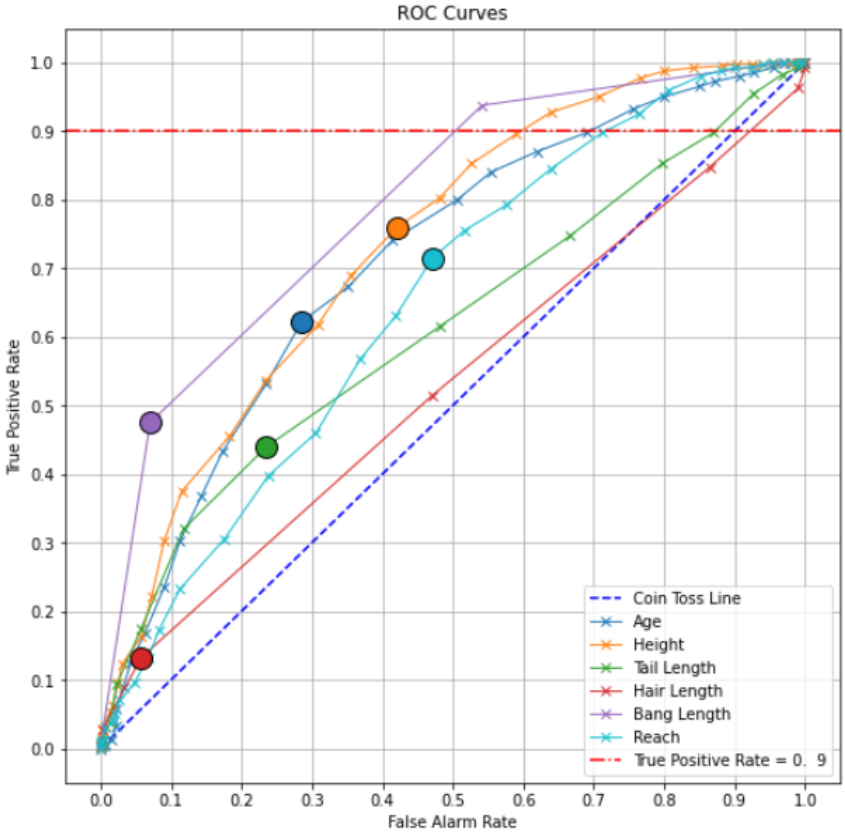

Figure 6: ROC Curves (90% True Positive Rate Highlighted)

Now, we would want a system with a very high true positive rate in situations such as: if someone felt sick, or are about to travel, they will most likely need to take a COVID text. The cheaper RAPID tests need to provide very accurate resuslts wherein we don't really mind taking the compromise that we may also get more false alarms, since they can always just get a RTPCR test to absolutely confirm of their infection status. So if someone tests positive on a RAPID test, it is very likely that they will actually have it.

My concluding thoughts are that - I would not have guessed for Bang Length to be the most optimal one-rule, for all three cases: where we wanted the minimal number of misclassifications, very low false alarm rates, and very high true positive rates. I think it had to do with the fact that there were only so many values that the attribute could take on. Which in retrospect, seems arbitrary, but at the time when I was looking at the data and saw that the Bang Length attribute only went from 0 to 10. I think I had a pre-existing bias that favored values that appeared to vary more; somehow I had formed a mental connection between training models and the amount of training data we have and the way that the values seem to vary more, perhaps. But this exercise helped me rethink my ideas.

I recall two specific concepts: namely, GIGO (Garbage in - Garbage Out) and Occam's Razor. That it doesn't matter if we use complex methods to understand the tiny changes datapoints can have if the data itself isn't good to begin with; and that even if the mathematical modeling we're doing seems simple, that does not necessarily imply that it would not be as useful as a super complex model.

On the other hand, questions (D) and (E), which asked for which one-rule I would use for a False Alarm Rate of 9% and True Positive Rate of 90% respectively helped me develop a better intuition of why exactly the "top right" of the ROC curve may not always be what you want. From what I understand now, as a data-scientist, we should ask ourselves the question of whether it's more important for us to have a very low false alarm rate and take the hit on the true positive rate, or have a hight true positive rate and a lower false alarm rate. Then, if we take the horizontal/vertical line made by the benchmark we've set (TPR = 90%, FAR=9%), we can then look at all the potential models we have and pick the one that seems to yield us the best true statistics.