Machine Learning Assignment 3: Evaluation Metrics

Arush Sharma 16BCE1127

The dataset that I have used is Mushroom Dataset. I have used Logistic Regression for classifying.

For this assignment, I have used 5 evaluation metrics, namely

1. Accuracy
2. Confusion Matrix
3. Classification Threshold
4. ROC Curve
5. AUC Score

Evaluation Metric 1: Accuracy

Accuracy represents the number of labels correctly classified by the model. Accuracy is often the starting point for analyzing the quality of a predictive model, as well as an obvious criterion for prediction.

Better the accuracy, better is the model.

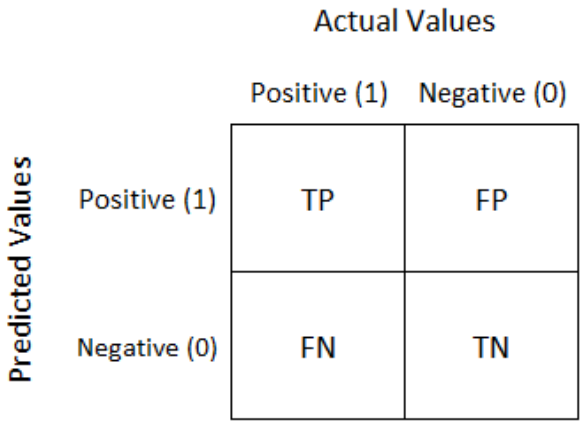
```
print(accuracy_score(y_predict,y_test))
```

```
0.7483998030526834
```

I am getting 74.839 % accuracy for my model which is neither good, neither bad.

Evaluation Metric 2: Confusion Matrix

A confusion matrix is a summary of prediction results on a classification problem.
The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.
The confusion matrix shows the ways in which your classification model is confused when it makes predictions.
It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.



For my mode, the Confusion Matrix is:

```
[[1490  427]
 [ 595 1550]]
```

For my model, the metrics of Confusion Matrix are:

True Positive: 1490

True Negative: 1550

False Positive: 427

False Negative: 595

Confusion Matrix can be used for calculation various evaluation metrics such as

Sensitivity, Specificity, False Positive Rate, Precision

1. **Sensitivity** (also called the true positive rate, the recall) measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition). **Better the sensitivity, better is the model.**

   For this model, I am getting 72.2 % sensitivity, that means the ratio of mushroom which are actually of class 1 and are also classified as of class 1 is 0.722.

2. **Specificity** (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).
   Better the specificity, better is the model.

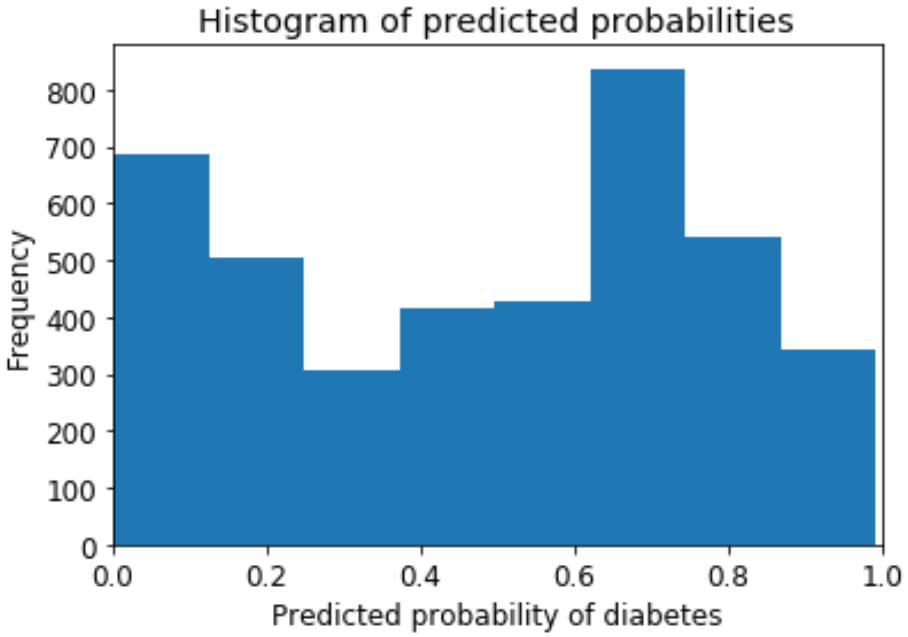
For this model, I am getting 77.7 % sensitivity, that means the ratio of mushroom which are actually of class 0 and are also classified as of class 0 is 0.777.

3. The **false positive rate** is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative events (regardless of classification). **Lower is the FPR, better is the model.**

   For this model, I am getting FPR as 22.2% which means the ratio of mushroom which are of class 0 but are classified as 1 to the actual number of mushrooms which are of class 0 is 0.222.

4. **Precision** is calculated as number of TP divided by TP + FP. This means of the mushrooms classified as of class 1, what proportion actually were of class 1? **Better the precision, better is the model.**
   For this model, I am getting precision of 78.4% the ratio of mushrooms actually of class 1 to the number of mushrooms classified as 1 is 0.784.

Evaluation Metric 3: Classification Threshold



From the histogram of predicted probabilities, we can see that more than 50% of the values lie to the right of 0.5 threshold, more would be predicted of class "0" in this case. Solution is to change the threshold to a higher value.

So, I have changed the threshold to 0.61(manually selected, by calculating sensitivity and specificity of new threshold).

Binarization:

In order to map a logistic regression value to a binary category, you must define a classification threshold (also called the decision threshold). A value above that threshold indicates "1"; a value below indicates "0". This is known as binarization. It is tempting to assume that the classification threshold should always be 0.5, but thresholds are problem-dependent, and are therefore values that you must tune.

The new confusion matrix for threshold 0.61 is

```
[[1876  209]
 [ 440 1537]]
```

- **Accuracy at new threshold**

  ```
  0.8402264894140817
  ```

  The accuracy increased from 0.74 to 0.84, a jump of 10%!

- **Sensitivity at new threshold**

  ```
  0.7774405665149215
  ```

  Increase of 5.5%

- **Specificity at new threshold**

  ```
  0.8997601918465228
  ```

  Increase of 12.2%

- **FPR at new threshold**

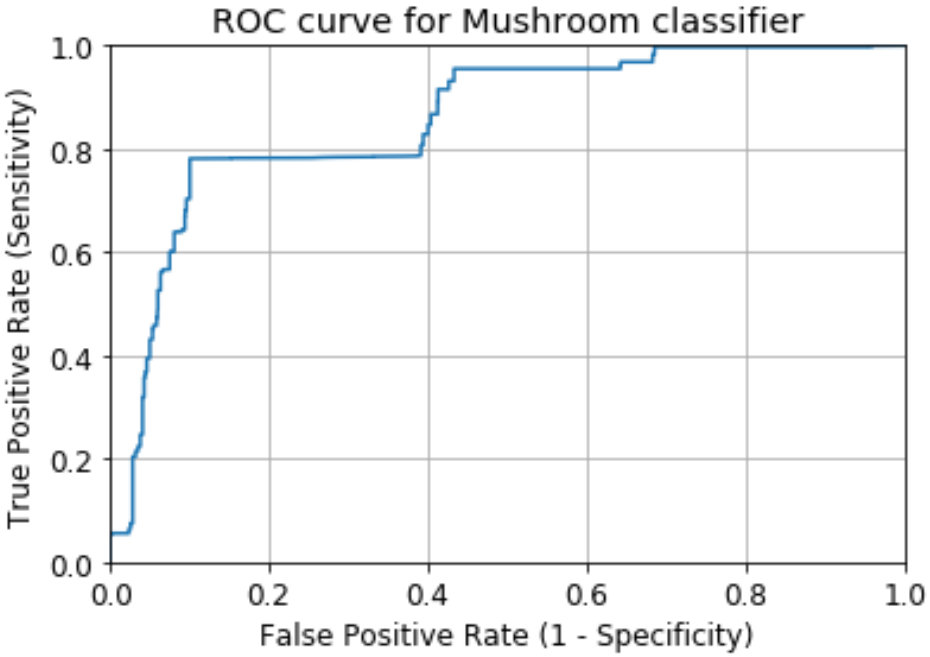
  ```
  0.10023980815347722
  ```

  Decreased by 12%

In short, the model has significantly improved across almost all metrics if the threshold is increased from 0.5(default) to 0.61(new).

Evaluation Metric 4: ROC Curve

In a **Receiver Operating Characteristic (ROC)** curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.



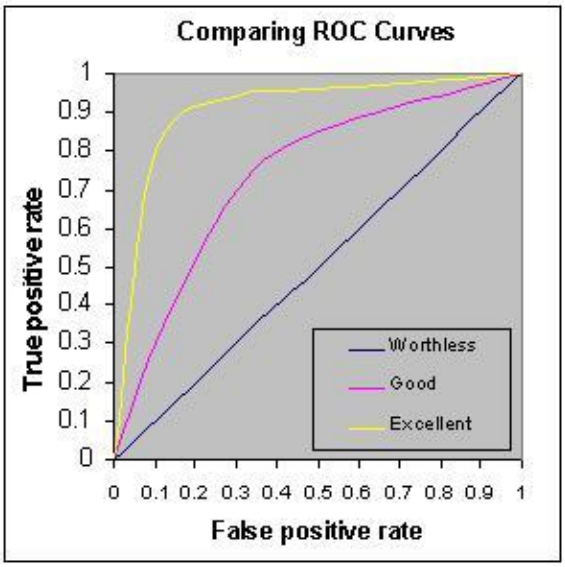
A ROC curve demonstrates several things:

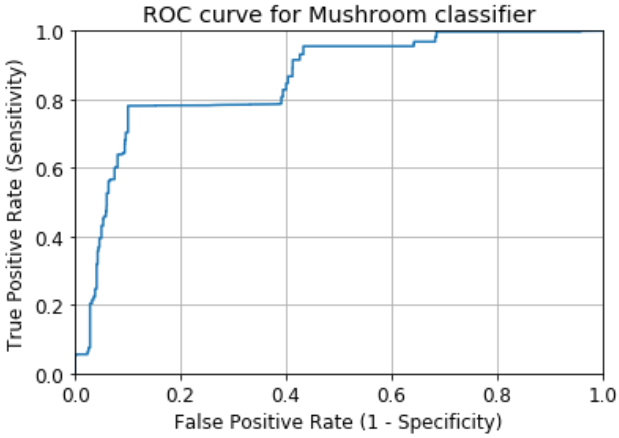
1. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
2. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
3. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
4. The slope of the tangent line at a cut point gives the likelihood ratio (LR) for that value of the test.

5. The area under the curve is a measure of text accuracy.

For our model, the ROC curve is closely following the left-hand border and then the top border of the ROC space. This means it is a good classifier.

Here I have taken the ROC Curve comparison,





We can see that our model is in between good to excellent one, based on ROC curve.

Evaluation Metric 5: AUC Score

AUC is the percentage of the ROC plot that is underneath the curve. AUC is useful as a single number summary of classifier performance

Higher value = better classifier

If you randomly chose one positive and one negative observation, AUC represents the likelihood that your classifier will assign a higher predicted probability to the positive observation

AUC is useful even when there is high class imbalance (unlike classification accuracy).

An area of 1 represents a perfect test; an area of .5 represents a worthless test. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

.90-1 = excellent (A)

.80-.90 = good (B)

.70-.80 = fair (C)

.60-.70 = poor (D)

.50-.60 = fail (F)

For my model, the AUC score is 0.851, according to the scoring given above, the model will come under good(B) category (.80-0.90).

SUMMARY

The assignment gave insights about how we can comprehensively measure the performance of a classifier. Most of people use only accuracy score for measuring the performance of a classifier. From this assignment, I learnt about using Confusion Matrix (Specificity, Sensitivity, Precision, Recall) to measure the performance, how changing threshold improves the model, what the ROC curve and AUC score tell about the classifier.