Decision Tree Classifier (Advanced)

Measures of Classification

- Specificity
- Sensitivity
- Recall
- Precision
- Area Under Curve
- F1 Score

The Measures Themselves

1. **Specificity, Sensitivity, Precision and Recall**

   Sensitivity, Specificity, and Accuracy are the terms which are most commonly associated with a Binary classification test and they statistically measure the performance of the test.

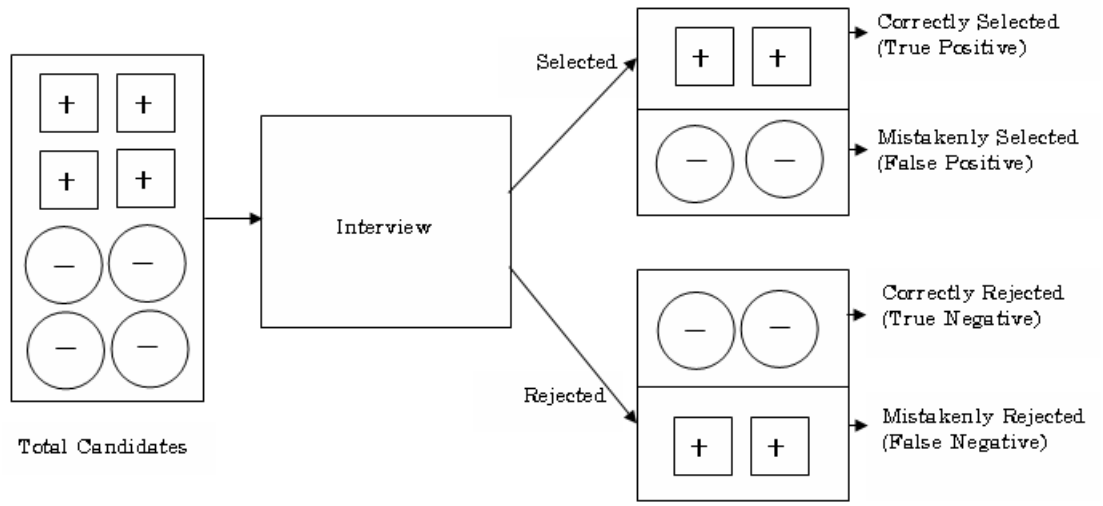
   Sensitivity, Specificity, and Accuracy are the terms which are most commonly associated with a Binary classification test and they statistically measure the performance of the test. In a binary classification, we divide a given data set into two categories on the basis of whether they have common properties or not by identifying their significance and in a binary classification test, as the name itself conveys, we deal with two datasets. Of these two categories, in general, Sensitivity indicates, how well the test predicts one category and Specificity measures how well the test predicts the other category. Whereas Accuracy is expected to measure how well the test predicts both categories.

A simple example to illustrate the idea

Imagine the scenario of a tough interview. The interviewer is supposed to be an experienced and sharp person so that we can blindly say that, all the candidates who are selected in the interview are exactly all the true excellent candidates who deserve the selection and no other candidates except those excellent ones are selected in the interview. I.e., an ideal or truly accurate interview will always give a positive result with excellent candidates only and a negative result with poor performers only.

This is not the case for all interviews. In practice this means that all excellent candidates may not be selected and this can be described by the term Sensitivity. Sensitivity measures the proportion of excellent candidates who were correctly identified to the total number of excellent candidates.. Similarly not all poor candidates may not be rejected and this can be described by the term Specificity. Specificity measures the proportion of poor candidates

who were correctly rejected to the total number of poor candidates. Whereas Accuracy measures the proportion of Excellent and poor candidates those are selected correctly to the total number of excellent and poor candidates.



In the above figure, the box 'Total Candidates' contains 4 excellent candidates represented by 4 +ve squares and 4 poor candidates represented by 4 –ve circles. Suppose there is an interview, which aims at selecting excellent candidates into an output box 'Selected' and to rejecting poor performers into another output box 'Rejected'.

In practice interviews are not perfect, so not all excellent applicants may be picked up by the interview. Suppose, in this case the interview gives a correct positive result in 2 out of the 4 who are excellent, therefore 2 are represented as True Positive in the 'Correctly Selected' portion of the output box 'Selected' in the figure. Also, the interview gives a correct negative result in 2 out of 4 who are not excellent; therefore 2 are represented as True Negative in the 'Correctly Rejected' section of the output box 'Rejected' in the figure.

At the same time, 2 candidates have a false positive result, even if they are not excellent. These are therefore represented as False Positive in the 'Mistakenly Selected' section of the output box 'Selected' in the figure, as the selection by the interviewer is incorrect. Similarly 2 candidates have a false negative result, even if they are excellent. These are therefore represented as False Negative in the 'Mistakenly Rejected' Section of the output box 'Rejected' in the figure, as the interviewer is mistaken. Here,

Correctly Selected **--->** True Positive (Excellent candidates correctly selected in box 'Selected') = 2 Mistakenly Selected **--->** False Positive (Poor candidates wrongly selected in box 'Selected') = 2 Correctly Rejected **--->** True Negative (Poor candidates correctly selected in box 'Rejected') = 2

Mistakenly Rejected **--->** False Negative (Excellent candidates wrongly selected in box 'Rejected') = 2

Sensitivity = True Positive / True Positive + False Negative
= Correctly Selected / Correctly Selected + Mistakenly Rejected
= Correctly Selected/ Total Excellent candidates who actually deserved Selection

i.e., 2/2+2= 2/4= 50% If the output box 'Excellent' bring together all the 4 excellent candidates and no excellent candidates in the output box 'Poor', then, the sensitivity will have its maximum value. i.e., 4/4+0 = 100%

Specificity = True Negative/ True Negative + False Positive
= Correctly Rejected / Correctly Rejected + Mistakenly Selected
= Correctly Rejected/ Total poor candidates who actually deserved Rejection

i.e., 2/2+2= 2/4= 50% If, the output box 'Poor' bring together all the 4 poor candidates and no poor candidates in the output box 'Excellent', then, the specificity will have its maximum value. I.e., 4/4+0 = 100%

Accuracy = (True Positive +True Negative)/ (True Positive+ False Positive + True Negative + False Negative)
= (Correctly Selected + Correctly Rejected)/ (Correctly Selected + Mistakenly Selected + Correctly Rejected
+ Mistakenly Rejected)
= (Correctly Selected + Correctly Rejected)/ (Total Excellent candidates who actually deserved Selection + Total poor candidates who actually deserved Rejection)

i.e., (2+2)/ (2+2+2+2) =4/8 =0.50

A comparative study is done, by considering some random outputs to identify the relationships between Accuracy, Sensitivity and Specificity, by exemplifying two datasets with, same number of data and different number of data. It is observed that, from the results, Accuracy alone is not a dependable factor, since Accuracy is derived from Sensitivity and Specificity. I.e. Accuracy= (Sensitivity + Specificity)/2. In the case where, the number of excellent candidates and poor performers are equal, if any one of the factors, Sensitivity or Specificity is high then Accuracy will bias towards that highest value. (I.e., if Sensitivity is high, Accuracy will bias towards Sensitivity, or, if Specificity if high, Accuracy will bias towards Specificity. If both are high, Accuracy will also high and if both are low, then Accuracy will be low.). But in the case where, the count of excellent candidates is low and the count of poor performers is high, Accuracy varies with Specificity without considering Sensitivity. Similarly, in the case where the number of excellent candidates is high and the

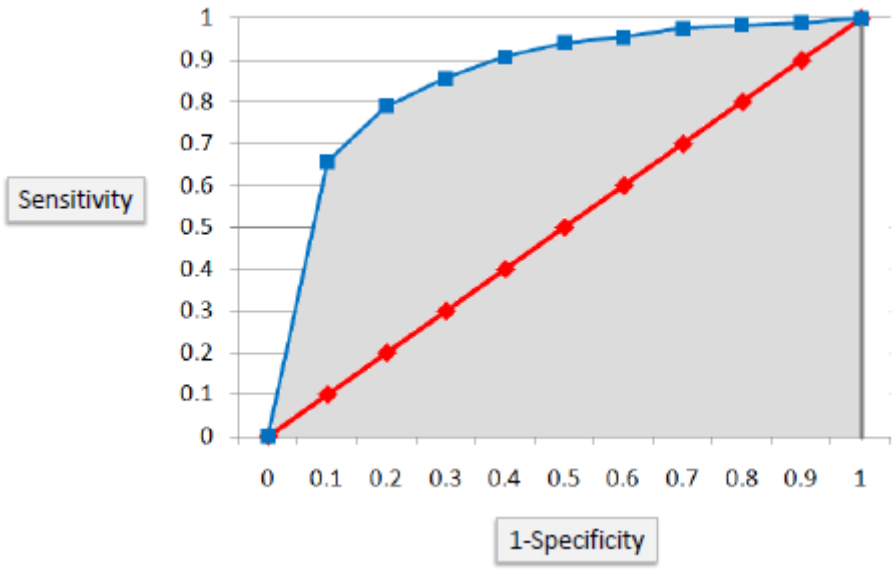
number of poor performers is low then, Accuracy tends to vary with Sensitivity without considering Specificity also.

To nullify the Accuracy variations regarding the imbalance in the count of data in the two datasets, we can apply data normalization. Through this, the variations of Accuracy with the predominance of Sensitivity or Specificity could brought to a normal state, i.e. by considering Sensitivity and Specificity together, rather than considering either of them.

Recall is another name for sensitivity.

2. Area Under Curve (AUC) Score

Area under ROC curve is often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1.



An area under the ROC curve of 0.8, for example, means that a randomly selected case from the group with the target equals 1 has a score larger than that for a randomly chosen case from the group with the target equals 0 in 80% of the time. When a classifier cannot distinguish between the two groups, the area will be equal to 0.5 (the ROC curve will coincide with the diagonal). When there is a perfect separation of the two groups, i.e., no overlapping of the distributions, the area under the ROC curve reaches to 1 (the ROC curve will reach the upper left corner of the plot).

3. F1 Score

In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results

divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

The traditional F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall:

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

The F1 score is also known as the Sørensen–Dice coefficient or Dice similarity coefficient (DSC).

The Dataset

The dataset is live Bitcoin trading prices from different markets around the world. This data is served by the Quandl API for python that gives all the attributes we need.

Quandl (/ˈkwɑːndəl/) is a platform for financial, economic, and alternative data that serves investment professionals. Quandl sources data from over 500 publishers. All Quandl's data are accessible via an API. API access is possible through packages for multiple programming languages including R, Python, Matlab, Maple and Stata.

The Code

All the codes can be found on the github repository here.

Additionally, the google colab notebook of the code can be found here.

Interpretation

As the various compared metrics clearly show here, even though the classes were split in a very intricate manner, the classifier did really well to categorise them properly with a standard test-train split of the database.