Optional

Subjective Questions - III

Q1. What is accuracy?

Accuracy is the number of correct predictions out of all predictions made.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Predictions}$$

Q2. Why is accuracy not a good measure for classification problems?

Accuracy is not a good measure for classification problems because it gives equal importance to both false positives and false negatives. However, this may not be the case in most business problems. For example, in the case of cancer prediction, declaring cancer as benign is more serious than wrongly informing the patient that he is suffering from cancer. Accuracy gives equal importance to both cases and cannot differentiate between them.

Q3. What is the importance of a baseline in a classification problem?

Most classification problems deal with imbalanced datasets. Examples include telecom churn, employee attrition, cancer prediction, fraud detection, online advertisement targeting, and so on. In all these problems, the number of the positive classes will be very low when compared to the negative classes. In some cases, it is common to have positive classes that are less than 1% of the total sample. In such cases, an accuracy of 99% may sound very good but, in reality, it may not be.

Here, the negatives are 99%, and hence, the baseline will remain the same. If the algorithms predict all the instances as negative, then also the accuracy will be 99%. In this case, all the positives will be predicted wrongly, which is very important for any business. Even though all the positives are predicted wrongly, an accuracy of 99% is achieved. So, the baseline is very important, and the algorithm needs to be evaluated relative to the baseline.

Q4. What are false positives and false negatives?

False positives are those cases in which the negatives are wrongly predicted as positives. For example, predicting that a customer will churn when, in fact, he is not churning.
False negatives are those cases in which the positives are wrongly predicted as negatives. For example, predicting that a customer will not churn when, in fact, he churns.

Q5. What are the true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), and false negative rate (FNR)?

TPR refers to the ratio of positives correctly predicted from all the true labels. In simple words, it is the frequency of correctly predicted true labels.

$$TPR = \frac{TP}{TP+FN}$$

TNR refers to the ratio of negatives correctly predicted from all the false labels. It is the frequency of correctly predicted false labels.

$$TNR = \frac{TN}{TN+FP}$$

FPR refers to the ratio of positives incorrectly predicted from all the true labels. It is the frequency of incorrectly predicted false labels.

$$FPR = \frac{FP}{TN+FP}$$

FNR refers to the ratio of negatives incorrectly predicted from all the false labels. It is the frequency of incorrectly predicted true labels.

$$FNR = \frac{FN}{TP+FN}$$

Q6. What are sensitivity and specificity?

Specificity is the same as true negative rate, or it is equal to 1 – false positive rate. It tells you out of all the actual '0' labels, how many were correctly predicted.

$$Specificity = \frac{TN}{TN+FP}$$

Sensitivity is the true positive rate. It tells you out of all the actual '1' labels, how many were correctly predicted.

$$Sensitivity = \frac{TP}{TP+FN}$$

Q7. What are precision and recall?

Precision is the proportion of true positives out of predicted positives. To put it in another way, it is the accuracy of the prediction. It is also known as the 'positive predictive value'.

$$Precision = \frac{TP}{TP+FP}$$

Recall is the same as the true positive rate (TPR) or the sensitivity.

$$Recall = \frac{TP}{TP+FN}$$

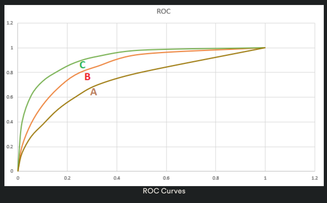
Q8. What is F-measure?

It is the harmonic mean of precision and recall. In some cases, there will be a trade-off between the precision and the recall. In such cases, the F-measure will drop. It will be high when both the precision and the recall are high. Depending on the business case at hand and the goal of data analytics, an appropriate metric should be selected.

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Q9. Explain the use of ROC curves and the AUC of an ROC Curve.

An ROC (Receiver Operating Characteristic) curve illustrates the performance of a binary classification model. It is basically a TPR versus FPR (true positive rate versus false positive rate) curve for all the threshold values ranging from 0 to 1. In an ROC curve, each point in the ROC space will be associated with a different confusion matrix. A diagonal line from the bottom-left to the top-right on the ROC graph represents random guessing. The Area Under the Curve (AUC) signifies how good the classifier model is. If the value for AUC is high (near 1), then the model is working satisfactorily, whereas if the value is low (around 0.5), then the model is not working properly and just guessing randomly. From the image below, curve C (green) is the best ROC curve among the three and curve A (brown) is the worst ROC curve among the three.



ROC Curves

Q10. How to choose a cutoff point in case of a logistic regression model?

The cutoff point depends on the business objective. Depending on the goals of your business, the cutoff point needs to be selected. For example, let's consider loan defaults. If the business objective is to reduce the loss, then the specificity needs to be high. If the aim is to increase the profits, then it is an entirely different matter. It may not be the case that profits will increase by avoiding giving loans to all predicted default cases. But it may be the case that the business has to disburse loans to default cases that are slightly less risky to increase the profits. In such a case, a different cutoff point, which maximises profit, will be required. In most of the instances, businesses will operate around many constraints. The cutoff point that satisfies the business objective will not be the same with and without limitations. The cutoff point needs to be selected considering all these points. If the business context doesn't matter much and you want to create a balanced model, then you use an ROC curve to see the tradeoff between sensitivity and specificity and accordingly choose an optimal cutoff point where both these values along with accuracy are decent.