Accuracy

Non-Technical:

In a classification report, if the accuracy is reported as 95%, what does that mean in simple terms?

How would you explain to a non-technical person why the accuracy alone might not always reflect the model's true performance?

If the accuracy in the classification report is 98% in a dataset with 95% of one class and 5% of another, what do you think is happening behind the scenes?

How can accuracy be affected when dealing with imbalanced datasets, and how might that mislead you in evaluating the model?

What would it mean for the model's accuracy if it predicts 100% of the majority class correctly but fails to predict any of the minority class?

Technical:

How is accuracy calculated from the classification report, and why is it generally straightforward to compute?

In the context of the classification report, what does accuracy not capture about the model’s performance on individual classes?

Can you explain how accuracy would be impacted if you had a multi-class problem with one class being overrepresented in the dataset?

How would you interpret accuracy when using it in the classification report for a multi-class classification problem?

If the classification report shows a high accuracy but low precision or recall for certain classes, what would that indicate about the model?

Recall

Non-Technical:

In a classification report, what does a recall of 80% for a class mean in simpler terms?

If the recall is reported as low in the classification report, what does it tell us about the model's ability to identify positive cases of that class?

Why is recall often a more important metric than precision in certain scenarios, like medical diagnoses?

If recall for a class is high, what does that tell you about how many of the actual positive cases are correctly identified by the model?

If you have a recall of 100% in a classification report, what does that imply about the number of false negatives?

Technical:

How is recall computed from the classification report, and what role do false negatives play in this calculation?

If a model shows a high recall but low precision, what does that indicate in terms of the model's prediction behavior?

How would you interpret the recall value in a classification report for an imbalanced dataset, where the positive class is underrepresented?

If the recall for a class in the classification report is 0.75, what does that mean in terms of false negatives and correct positive predictions?

How can you improve recall without significantly affecting other metrics, and what strategies would you use based on the classification report?

Precision

Non-Technical:

In simple terms, what does the precision value of 90% mean in the classification report for a given class?

If precision is low for a class, what does that tell us about the model's tendency to make false positive errors for that class?

Why would we care about having a high precision when evaluating a classification model for fraud detection?

If precision is high but recall is low in a classification report, what does that tell us about the model's performance?

How can you interpret a precision of 100% in the classification report for a specific class?

Technical:

How is precision calculated from the classification report, and what role do false positives play in this metric?

If you observe that precision is high but recall is low, what might that suggest about the model's decision-making process, and how would you address this imbalance?

How would you analyze precision in a classification report for an imbalanced dataset where one class is underrepresented?

How does precision behave when the class distribution changes in a multi-class classification scenario?

If the precision for a class is 0.85, what does that tell you about the false positives for that class?

F1 Score

Non-Technical:

If the F1 score is 0.9 in the classification report, what does that mean about the balance between precision and recall?

Why is the F1 score considered a better metric in certain cases than accuracy, particularly when dealing with imbalanced classes?

What would it mean for the model if the F1 score is close to 0 in the classification report for a specific class?

How would you explain to a non-technical person the importance of the F1 score when evaluating a classification model's performance?

Why might the F1 score be a better measure of model performance in situations where both false positives and false negatives matter equally?

Technical:

How is the F1 score computed from the classification report, and why is it the harmonic mean of precision and recall?

What does it mean when the F1 score is significantly lower than both precision and recall for a class in the classification report?

How does the F1 score reflect trade-offs between precision and recall, and why is this important when evaluating model performance?

How would you improve the F1 score for a model with a low value, and what actions would you take based on the classification report?

How would you interpret a situation where the F1 score is high, but recall is low and precision is high for a class in the classification report?

Weighted Average

Non-Technical:

In the classification report, what does the "weighted average" tell us about the model's performance across all classes?

How does the weighted average differ from other averages in terms of handling imbalanced data?

If a model's weighted average of precision is low, what might that tell you about the model's performance across all classes?

Why is the weighted average a useful metric when we have a classification problem with classes of different sizes?

If the weighted average precision is significantly different from the macro average precision, what might that indicate about the distribution of classes?

Technical:

How is the weighted average calculated in the classification report, and how does it account for class imbalance?

In the context of a classification report, why is it important to compute the weighted average for precision, recall, and F1 score?

How does the weighted average differ from the macro average in multi-class classification tasks, and what are the implications of this difference?

How would you compute the weighted average F1 score, and what role does the number of instances in each class play in this calculation?

How do class frequencies influence the weighted average, and why might this metric be misleading in cases of extreme class imbalance?

Macro Average

Non-Technical:

In the classification report, how would you explain the macro average in simple terms?

If the macro average recall is low, what does that suggest about the model's performance across all classes?

How does the macro average treat each class in the classification report, and why might it be useful in comparing model performance?

What's the main difference between the macro average and the weighted average when it comes to evaluating performance in multi-class classification problems?

Why might the macro average be more useful than accuracy in some multi-class problems?

Technical:

How is the macro average calculated in the classification report, and what role does the performance of each class play in this calculation?

How does the macro average handle imbalanced datasets, and what are the trade-offs compared to the weighted average?

Why might the macro average be less sensitive to the performance on the majority class compared to the weighted average?

How would you interpret the macro average precision, recall, and F1 score in a multi-class classification task with imbalanced classes?

Can you explain how the macro average gives equal importance to each class, regardless of the class distribution, and how that affects the model evaluation?