Apollo Solutions Machine Learning Developer Test Interpretation questions

1- After analysing the distribution models, it can be concluded that the best balance between variance and bias is seen in distribution C because it captures the overall trend of the data without overfitting noise or assuming too much. In distribution A, the model is highly biased and has low variance, it is too simple like a straight line to curved data, it has high bias and fails to capture the true relationship, leading to underfitting. This results in poor generalization as the model will not be able to learn substantial relationships from data. In distribution B is of low bias and high variance, that adjusts to every point, it develops high variance, overfits the training data, and struggles to generalize to new data. This is overfitting, where the model performs well on training but poorly on testing. The model will not generalize to new data as it only memorizes the patterns that are likely to break down in other data sets.

Distribution C is a balance between these two. The model is general enough to capture the underlying pattern in the data without adding too much complexity that would result in overfitting.

2- The graph shown is a ROC curve and provides a visual representation of the trade-offs between true positive rate (TPR) and false positive rate (FPR) at various thresholds. It provides insights into how the model can balance tradeoffs between detecting positive instances and avoiding false positives at different thresholds. The AUC, or Area Under the Curve, is a single scalar value that ranges from 0 to 1 and provides a snapshot of the model's performance. Its only possible to calculate the AUC after generating the ROC curve because the AUC represents the area under the curve.

Between the two curves, the one which is closer to the top-left corner is a better fit. If the red curve is closer to the top-left corner compared to the green curve, it implies the model represented by the red curve is better. The reason being that it can achieve a greater True Positive Rate for a given False Positive Rate, i.e., it is better at correctly identifying positive cases while maintaining low false positives.

3-Looking at the model’s performance in the test evaluation, we can see that the training accuracy shown in Graph A and training loss in Graph B shows how well the model learned over time. If the accuracy increased and the loss decreased throughout the training process, that’s usually a good sign that the model was improving. But the real test of its performance comes from the confusion matrix, which tells us how well the model handled test samples for classes A, B, and C.

From the table, we can see how well the model handled the test samples, for instance in class A, the model got it right 50% of the time, but it misclassified 25% of the samples as B and another 25% as C. For class B, the model correctly identified 45% of the samples, but 15% were mistaken for A, and 35% ended up classified as C. Similarly, for class C, the accuracy was also 50%, with 10% misclassified as A and a significant 40% confused with B.

Overall, the model's performance in the test is not very good. While it accurately predicts about half of classes A and C and a few less for class B, the misclassifications especially between B and C are disconcerting. This means that the model has trouble differentiating these two classes effectively. During the testing procedure, it was clear that the model has a problem with correctly classifying some samples, particularly between classes B and C. This could mean that characteristics the model was trained on aren't strong enough to differentiate between these classes in a correct manner, leading to errors consistently.