Joshua Douglas

Binary Classifier Model ROC Analysis

Q1. *Manually calculate the sensitivity and specificity of the model, using a predicted_prob threshold of greater than or equal to 0.5.*

A1. The sensitivity and specificity can be calculated from the confusion matrix, provided below along with an overview of the confusion matrix interpretation. The sensitivity is defined as the true positive rate (True Positive / (True Postive + False Negative)) and the specificity is defined as the true negative rate (True Negative / (True Negative + False Positive)). These ratios can be easily calculated with the provided matrix, yielding a **sensitivity value of 0.806 and a specificity value of 0.772.**

Binary Classification Model Confusion Matrix Results

	Predicted Class Zero	Predicted Class One
Known Class Zero	39334.0	11589.0
Known Class One	9509.0	39568.0

Confusion Matrix Interpretation

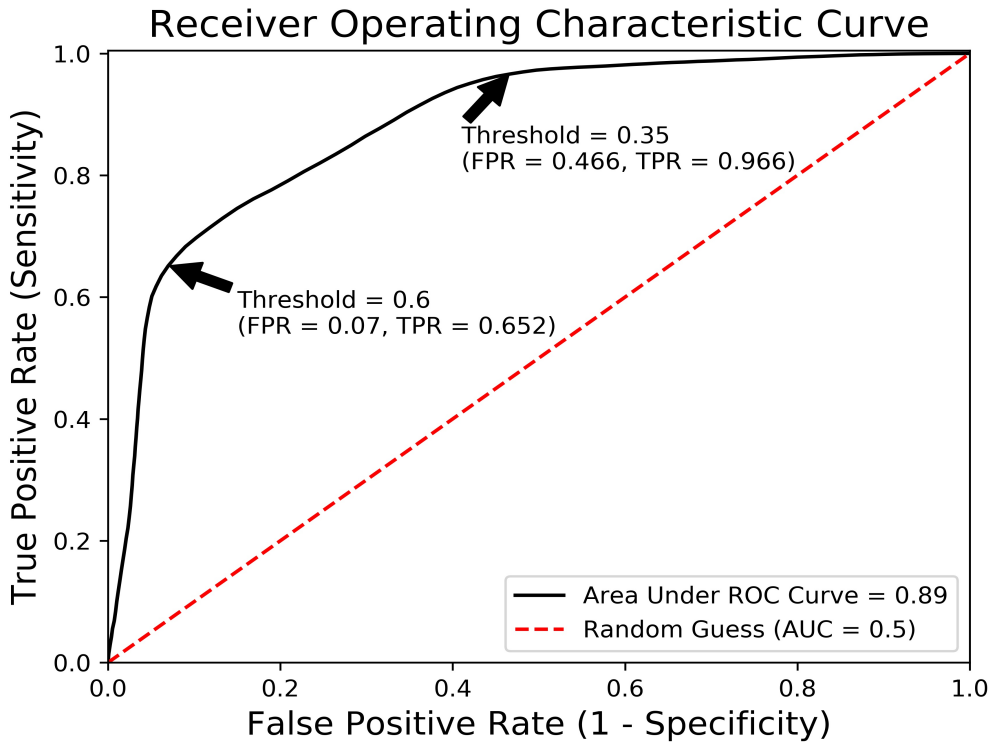
	Predicted Condition Negative	Predicted Condition Positive
Known Condition Negative	True Negative	False Positive
Known Condition Postiive	False Negative	True Positive

Q2. *Manually calculate the Area Under the Receiver Operating Characteristic Curve.*

A2. The Receiver Operating Characteristic (ROC) Curve is a plot of true positive rate (TPR) against the false positive rate (FPR), across the range of model threshold settings. In our case the ROC plot data can be obtained by computing the confusion matrix, and corresponding TPR/FPR values, for probability threshold values between zero and one. The definite integral, or true area under the curve, can be approximated using numerical methods such as the Riemann Sum or Trapezoid rule. These approximations approach the true value as the resolution of the calculation increases, and both methods were found to converge to similar values when the parameter sweep step-size was set to 0.01. Using the trapezoid rule **the Area Under the Curve is approximately 0.89**.

Q3. *Visualize the Receiver Operating Characterstic Curve.*

A3. The ROC curve is provided on the next page. The curve demonstrates that for all reasonable threshold values the proportion of correctly classified class 1 samples is greater than the proportion of the incorrectly classified class 0 samples (i.e. the model is better than random chance at all thresholds). With a AUC of 0.89 the model is a reasonably good classifier, depending on the context. The ROC curve can be used to set the appropriate threshold, depending on the desired ratio of FPR to TPR. A review of the data indicates that a threshold of 0.35 (FPR = 0.466, TPR = 0.966) is appropriate if it is a priority to correctly classify true positives and a threshold of 0.6 (FPR = 0.070, TPR = 0.652) is appropriate if it is a priority to minimize the number of incorrectly classified negative / class zero events.


Receiver Operating Characteristic Curve
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
Threshold = 0.35
(FPR = 0.466, TPR = 0.966)
Threshold = 0.6
(FPR = 0.07, TPR = 0.652)
Area Under ROC Curve = 0.89
Random Guess (AUC = 0.5)
0.0
0.2
0.4
0.6
0.8
1.0