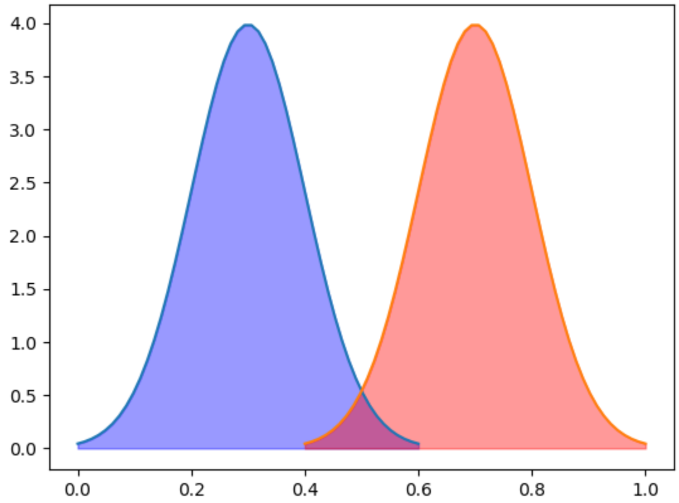
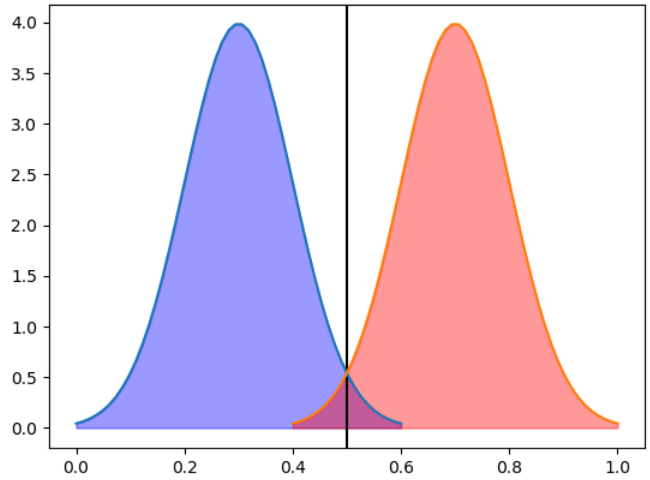
COMP 642 Assignment --- MODULE 6

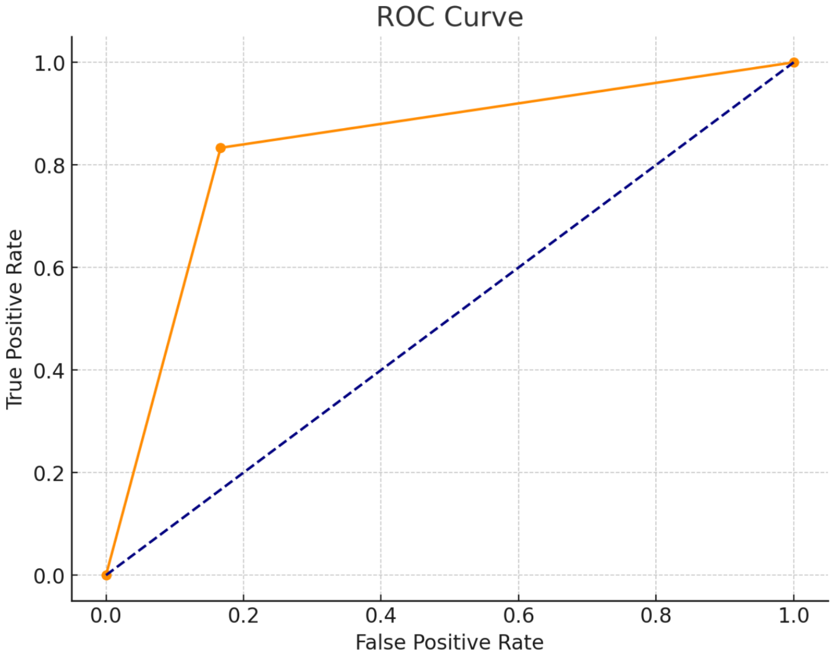
1. Suppose we have a dataset of credit card transaction information and for each transaction record we have a label indicating whether it is fraudulent. We build a Logistic Regression model that can classify whether the transaction is fraudulent or not.
   The distribution of the prediction is shown below, the x-axis represents the predicted probability, the y-axis represents the observations (like a histogram). Each blue and red pixel represents a transaction for which you want to predict the fraud status. There are 300 blue pixels transactions that were non-fraud, and 300 red pixels transactions that are fraud.



   a. The output of logistic regression will output a probability. In order to make the classification we need to choose a decision boundary. On the left-hand side of the decision boundary the prediction will be all negative (non-fraud), and on the right hand set the prediction will be all positive (fraud). If the line is drawn at 0.5, which is a common choice in binary classification, 250 transactions are classified as fraud and 250 are classified as non-fraud. The chart is shown below. Calculate the True Positive Rate, False Positive Rate and the position of this decision boundary on ROC curve.
      a. True Positive = classified fraud = 250
      b. False negative = total fraud – true positive = 300-250 = 50
      c. False Positive = total non-fraud – true positive = 300 – 250 = 50
      d. True negative = 250
      e. True Positive Rate = true positive / total fraud = 250/300 = 0.833
      f. False Positive Rate = false positive / total non-fraud = 50 / 300 = 0.167

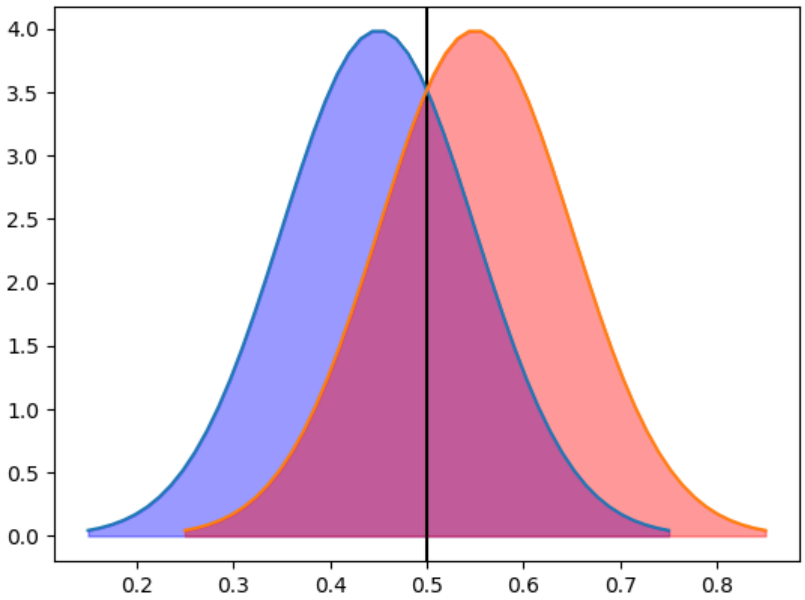


b. Draw the possible ROC curve for this model.

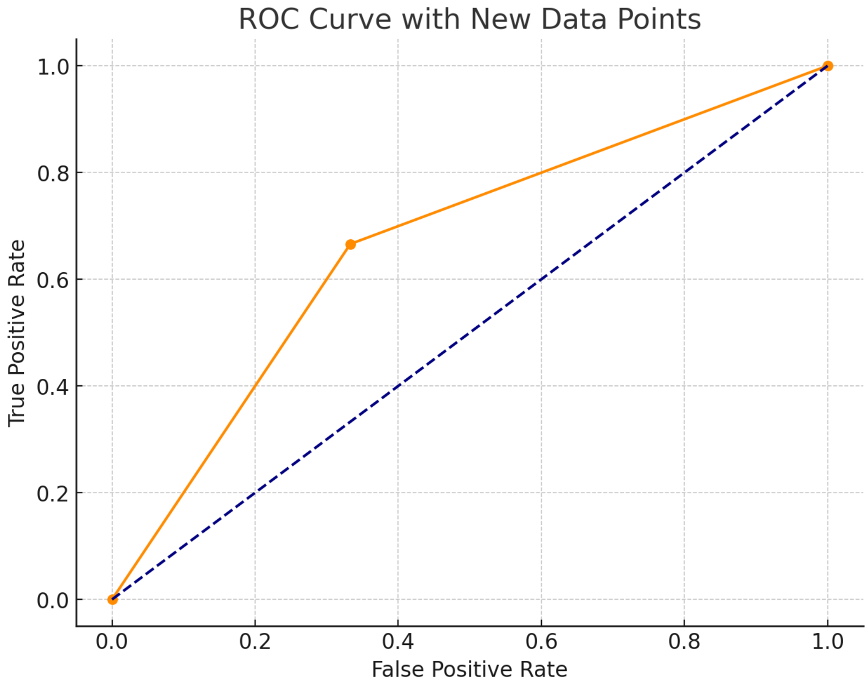


c. If we change the hyper parameters of the logistic regression model, the result prediction probability is shown below. The decision boundary is still 0.5 but 200 data are classified as fraud and 200 are classified as non-fraud. Calculate the same items as in part a.

   a. True Positive = classified fraud = 200
   b. False negative = total fraud – true positive = 300-200 = 100
   c. False Positive = total non-fraud – true positive = 300 – 200 = 100
   d. True negative = 200
   e. True Positive Rate = true positive / total fraud = 200/300 = 0.666
   f. False Positive Rate = false positive / total non-fraud = 100 / 300 = 0.333



d. Draw the possible ROC curve for this model.



e. Which model do you think is better and why?

The first ROC curve would be considered the better option. The first curve that we looked at had a higher true positive rate vs the ROC curve in question d. Another way to find a better ROC curve, if the answer is not as straight forward and many data points exist, is to find the area under the curve. We can see that there is greater area under the curve for our first ROC vs our second. In some instances, this is not as apparent and finding the area would help to solidify the best option.

f. Is making the decision boundary equal to 0.5 a good choice for this binary classification problem? Explain why or why not.
It is a good choice. In a logistic regression model, the results are given as probabilities with the decision boundary at 0.5 and is representative of the sigmoid function. This means it is a good interpretation of the performance of our model. However, this is a fraud detection problem, and the decision boundary may change if the risks for false negatives are more costly to the business/person.

2. The line below shows the degree of complexity of a machine learning model. Add to it annotations of: (more/less) bias, variance. (try to not look at the lecture notes), complexity.

More bias, less variance   Less bias, More variance
< -- >
less flexible (lower dimensionality)   more flexible (higher dimensionality)

Bias assumes that the model would possibly underfit if there is less flexibility, lower dimensionality and therefore greater assumptions being made about the model. Whereas variance has the opposite effect making less assumptions, over complicating the model with high dimensionality and causing the model to overfit.

.

3. Execute the hw_6.ipynb.

Submit a .doc or .pdf with your written answers. Submit your Python notebook. Submit a PDF of your Python notebook.