5. (2.75 points)

The following confusion matrix shows the result from a claim fraud model with a discrimination threshold of 25%:

	Predicted	
Actual	Yes	No
Yes	72	162
No	63	1203

a. (0.5 point)

Identify a link function that can be used for a generalized linear model that has a binary target variable and briefly explain why this link function is appropriate.

b. (0.5 point)

Calculate the sensitivity and specificity from the above data.

c. (1.5 points)

Plot the receiver operating characteristic (ROC) curve with the discrimination threshold of 25%. Label each axis, the coordinates, and the discrimination thresholds of 100%, 25%, and 0% on the curve.

In addition, plot the ROC curve for each of the following two models:

i. A model with no predictive power
ii. A hypothetical "perfect" model

d. (0.25 point)

Briefly describe how the severity of claims will impact the selection of the model threshold.

CONTINUED ON NEXT PAGE

SAMPLE ANSWERS AND EXAMINER'S REPORT

QUESTION 5	
TOTAL POINT VALUE: 2.75	**LEARNING OBJECTIVE(S): A3, A4**
SAMPLE ANSWERS	
Part a: 0.5 point	

Sample 1

Logit link function $= \ln\left(\frac{\mu}{1-\mu}\right)$; 0≤μ≤1

Appropriate because the inverse of the logit function is the logistic function. Logistic function produces a variable between 0 and 1 which coupled with a discrimination threshold produces a binary 0 or 1 outcome.

Sample 2

The logit link function is appropriate because its inverse is the logistic which is $f(x) = \frac{1}{1+e^{\wedge}}$. This allows us to take an unbounded value of x and return a value between 0 and 1, which is what we want when estimating probabilities.

Sample 3

The logit link function is used for logistic regression in GLM. It is appropriate b/c it has the ability to map any number from "-∞ to ∞" to "0 to 1". We can then pick a threshold like 50%. Then if we get .39 which is below 50% we can assign a "No". If greater than 50%, then we can assign "Yes".

Sample 4

Logit function $g(x) = \ln\left(\frac{\mu}{1-\mu}\right)$

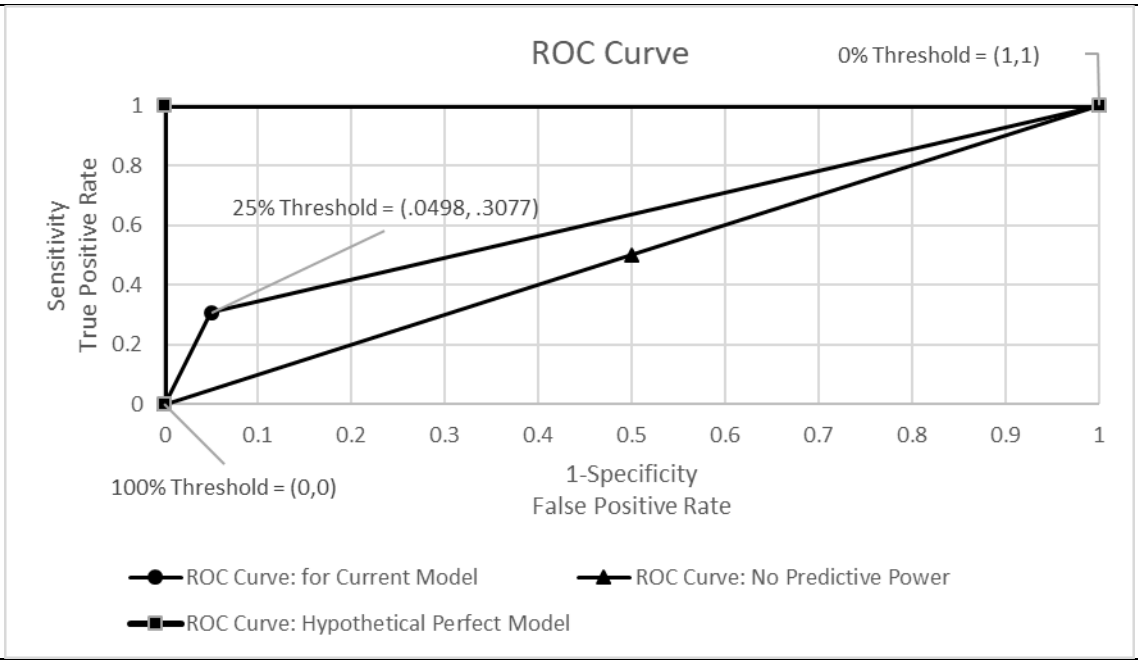
Appropriate b/c it maps to a range between 0&1, which is similar to a probability.

Part b: 0.5 point

$$Sensitivity = \frac{\text{True Positives}}{\text{Total Positive}} = \frac{72}{72+162} = 30.77\%$$

$$Specificity = \frac{\text{True Negatives}}{\text{Total Negatives}} = \frac{1203}{1203+63} = 95.02\%$$

Part c: 1.5 points



Part d: 0.25 point

Sample 1

The more severe the claims, the lower the discrimination threshold since the benefit of identifying fraudulent claims will outweigh the cost of investigating false negatives.

Sample 2

With high severity claims, we should lower the threshold so that more claims are predicted to be fraudulent and less fraudulent claims go undetected. Because the claims are more severe, the cost of investigating additional claims will be outweighed by the cost if those claims go undetected.

EXAMINER'S REPORT

Candidates were expected to know the appropriate link function to use for a generalized linear model (GLM) with a binary target variable, how to construct the receiver operating characteristic (ROC) curve and to understand the components within the ROC curve, and how the severity of claims impact the selection of the model threshold.

Part a

Candidates were expected to know which link function to use for a GLM that has a binary target variable.

Candidates did not receive full credit if they gave an incorrect link function or didn't explain why the link function would be appropriate for a binary target variable.

Common mistakes included:

- Mixing up the logit function and the logistic function
- Giving an incorrect link function such as:
    - Log
    - Binomial

- Negative Binomial
- Indicating an incorrect range of the linear predictor
- Saying the output of μ is either 0 or 1, without indicating that it is a range between 0 and 1.

Part b

Candidates were expected to know how to calculate the sensitivity and specificity from the confusion matrix.

A common mistake was using the incorrect denominator.

- For sensitivity, the most common incorrect denominator was adding the true positives and false positives
- For specificity, the most common incorrect denominator was adding the false negatives and true negatives.

Part c

Candidates were expected to plot the ROC curve with the 0%, 25%, and 100% discrimination thresholds, as well as the model with no predictive power and a hypothetical perfect model.

Common mistakes included:

- Mixing up the discrimination thresholds of 0% and 100%
- Incorrectly labeling the x-axis and y axis
  - Labeling the x-axis as sensitivity and the y-axis as 1 –specificity
  - Labeling the x-axis as specificity or false negative rate
  - Labeling the y-axis as true negative rate
- Incorrectly labeling or not labeling the discrimination threshold of 25%
- Not plotting the ROC curve for the current model, but instead just plotting the 25% threshold
- Incorrectly plotting the ROC curve or just plotting a point of (0, 1) for the hypothetical perfect model without drawing the ROC curve
- Incorrectly plotting the ROC for the model with no predictive power.

Part d

Candidates were expected to describe how the severity of claims will impact the selection of the model threshold.

Common mistakes included:

- Indicating that the severity of the model would not impact the selection
- Indicating that the threshold would increase if the severity increased
- Incorrectly describing how the sensitivity and false positive rate would be impacted by the severity
- Stating a preference to accept more false positive when severity was high, but did not explain how the threshold selection itself would be impacted by this preference.