1. **Logistic Regression in Python**

Which of these methods is used for fitting a logistic regression model using **statsmodels**?

- OLS()
- GLM()
- RFE()
- LogisiticRegression()

GLM()

✓ **Correct**

Feedback:

Correct! The GLM() method is used to fit a logistic regression model using statsmodels.

2. **Confusion Matrix**

Given the following confusion matrix, calculate the accuracy of the model.

Actual/Predicted	No	Yes
No	1000	50
Yes	250	1200

- 96%
- 88%
- 83.33%
- 60%

88%

✓ **Correct**

Feedback:

Correct!

Recall that the formula for accuracy is given as -

Accuracy=Correctly Predicted Labels/Total Number of Labels

Here, the number of correctly predicted labels is = 1000 + 1200 = 2200.

And the total number of labels is = 1000 + 250 + 50 + 1200 = 2500

Hence, you have -

Accuracy=2200/2500=0.88=88%

Can you find the answer now?

3. Diabetic based on Threshold

Suppose you are building a logistic regression model to determine whether a person has diabetes or not. Following are the values of predicted probabilities of 10 patients.

Patient	Probability(Diabetes)
A	0.82
B	0.37
C	0.04
D	0.41
E	0.55
F	0.62
G	0.20
H	0.91
I	0.74
J	0.33

Assuming you arbitrarily chose a cut-off of 0.4, wherein if the probability is greater than 0.4, you'd conclude that the patient has diabetes and if it is less than or equal to 0.4, you'd conclude that the patient doesn't have diabetes, how many of these patients would be classified as diabetic based on the table above?

- 4
- 5
- 6
- 7

6

✓ Correct

Feedback:

Yes! The cut-off is given to be 0.4. Hence, for a patient to be classified as diabetic, Probability(Diabetes) needs to be greater than 0.4. As you can see in the table above, there are 6 patients who have Probability(Diabetes) > 0.4. These are:

A: 0.82, D: 0.41, E: 0.55, F: 0.62, H: 0.91, I: 0.74

4. Log Odds

Suppose you are working for a media services company like Netflix. They're launching a new show called 'Sacred Games' and you are building a logistic regression model which will predict whether a person will like it or not based on whether consumers have liked/disliked some previous shows. You have the data of five of the previous shows and you're just using the dummy variables for these five shows to build the model. If the variable is 1, it means that the consumer liked the show and if the variable is zero, it means that the consumer didn't like the show. The following table shows the values of the coefficients for these five shows that you got after building the logistic regression model.

Variable Name	**Coefficient Value**
TrueDetective_Liked	0.47
ModernFamily_Liked	-0.45
Mindhunter_Liked	0.39
Friends_Liked	-0.23
Narcos_Liked	0.55

Now, you have the data of three consumers Reetesh, Kshitij, and Shruti for these 5 shows indicating whether or not they liked these shows. This is shown in the table below:

Consumer	TrueDetective_Liked	ModernFamily_Liked	Mindhunter_Liked	Friends_Liked	Narcos_Liked
Reetesh	1	0	0	0	1
Kshitij	1	1	1	0	1
Shruti	0	1	0	1	1

Based on this data, which one of these three consumers is most likely to like to new show 'Sacred Games'?

- Reetesh
- Kshitij
- Shruti

Reetesh

✓ **Correct**

Feedback:

Correct!

To find the person who is most likely to like the show, you can use log odds. Recall the log odds is given by:

$\ln(P1-P)=\beta0+\beta1X1+\beta2X2+\beta3X3+...+\beta nXn$

Here, there are five variables for which the coefficients are given. Hence, the log odds become:

$\ln(P1-P)=0.47X1-0.45X2+0.39X3-0.23X4+0.55X5$

As you can see, we have ignored the $\beta0$ since it will be the same for all the three consumers. Now, using the values of the 5 variables given, you get -

$(\text{Log Odds})\text{Reetesh}=(0.47\times1)-(0.45\times0)+(0.39\times0)-(0.23\times0)+(0.55\times1)=1.02$

$(\text{Log Odds})\text{Kshitij}=(0.47\times1)-(0.45\times1)+(0.39\times1)-(0.23\times0)+(0.55\times1)=0.96$

$(\text{Log Odds})\text{Shruti}=(0.47\times0)-(0.45\times1)+(0.39\times0)-(0.23\times1)+(0.55\times1)=-0.13$

As you can clearly see, the log odds of Reetesh is the highest, hence, the odds of Reetesh liking the show is the highest and hence, he is most likely to like the new show, Sacred Games.

5. **Calculating Sensitivity**

Suppose you got the following confusion matrix for a model by using a cutoff of 0.5.

Actual/Predicted	Not Churn	Churn
Not Churn	1200	400
Churn	350	1050

Calculate the sensitivity for the model above. Now suppose for the same model, you changed the cutoff from 0.5 to 0.4 such that your number of true positives increased from 1050 to 1190. What will the be the change in sensitivity?

Note: Report the answer in terms of new_value - old_value, i.e. if the sensitivity was, say, 0.6 earlier and then changed to 0.8, report it as (0.8 - 0.6), i.e. 0.2.

- 0.05
- -0.05
- 0.1
- -0.1

0.1

✓ **Correct**

Feedback:

Correct!

Recall that the formula for sensitivity is given by -

Sensitivity=TP/TP+FN

Here, TP = 1050 and FN = 350

Hence, initially, the sensitivity was -

Sensitivity=1050 / (1050+350)=0.75

Now, when you changed the threshold, the number of true positives changes from 1050 to 1190. Now, since the number of actual positives will be the same always, i.e. (1050 + 350 = 1400) from the original confusion matrix, you can now calculate the new sensitivity as -

Sensitivity=1190 / 1400=0.85

Hence, the change in sensitivity = 0.85 - 0.75 = 0.1

6. **Calculating Precision**

Consider the confusion matrix you had in the last question.

Actual/Predicted	Not Churn	Churn
Not Churn	1200	400
Churn	350	1050

Calculate the values of precision and recall for the model and determine which of the two is higher.

- Precision
- Recall
- Both are the same

Ans: Recall

✓ **Correct**

Feedback:

Recall that precision and recall are given by -

Precision=TP / TP+FP

Recall=TP / TP+FN

Here, TP = 1050; FP = 400; TN = 350

Hence, you get -

Precision=1050 / 1050+400≈72.41%

Recall=1050/ 1050+350=75%

As you can see, of the two, recall is higher.

7. **True Positive Rate**

Fill in the blanks.

The True Positive Rate (TPR) metric is exactly the same as _____.

- Sensitivity
- Specificity
- Precision
- False Positive Rate

Sensitivity

✓ **Correct**

Feedback:

Correct! Recall the formula for TPR is given as:

TPR=True Positives / Total Number of Actual Positives

Which can be rewritten as -

TPR=TP/TP+FN

And this is exactly the same as sensitivity as you might remember.

8. **Threshold**

Suppose someone built a logistic regression model to predict whether a person has a heart disease or not. All you have from their model is the following table which contains data of 10 patients.

Patient ID	Heart Disease	Predicted Probability for Heart Disease	Predicted Label
1001	0	0.34	0
1002	1	0.58	1
1003	1	0.79	1
1004	0	0.68	1
1005	0	0.21	0
1006	0	0.04	0
1007	1	0.48	0
1008	1	0.64	1

1009	0	0.61	1
1010	1	0.86	1

Now, you wanted to find out the cutoff based on which the classes were predicted, but you can't. But can you identify which of the following cutoffs would be a valid cutoff for the model above based on the 10 data points given in the table? (More than one option may be correct.)

- 0.45
- 0.50
- 0.55
- 0.60

0.50

✓ Correct

Feedback:

Yes!

See the table carefully. For patient 1007, the predicted probability is 0.48 and the predicted class is 0. This means that the cutoff has to be greater than 0.48. Also, for patient 1002, the predicted probability is 0.58 and the predicted class is 1. This means that the cutoff has to be lesser than 0.58.

Therefore, the cutoff can lie between 0.48-0.58 and hence, 0.50 are 0.55 can be valid cutoffs for the model above.

0.55

✓ Correct

Feedback:

Yes!

See the table carefully. For patient 1007, the predicted probability is 0.48 and the predicted class is 0. This means that the cutoff has to be greater than 0.48. Also, for patient 1002, the predicted probability is 0.58 and the predicted class is 1. This means that the cutoff has to be lesser than 0.58.

Therefore, the cutoff can lie between 0.48-0.58 and hence, 0.50 are 0.55 can be valid cutoffs for the model above.

9. **Evaluation Metrics**

Consider the same model given in the last question.

Patient ID	Heart Disease	Predicted Probability for Heart Disease	Predicted Label
1001	0	0.34	0
1002	1	0.58	1
1003	1	0.79	1
1004	0	0.68	1
1005	0	0.21	0
1006	0	0.04	0
1007	1	0.48	0
1008	1	0.64	1
1009	0	0.61	1
1010	1	0.86	1

Calculate the values of Accuracy, Sensitivity, Specificity, and Precision. Which of these four metrics is the highest for the model?

- Accuracy
- Sensitivity
- Specificity
- Precision

Sensitivity

✓ **Correct**

Feedback:

From the table given above, you can easily find out that -

TN = 3

FP = 2

FN = 1

TP = 4

Hence, your confusion matrix will look like:

Actual/Predicted	No Heart Disease	Heart Disease
No Heart Disease	3	2
Heart Disease	1	4

Hence, you get -

Accuracy=3+43+2+1+4=70%

Sensitivity=44+1=80%

Specificity=33+2=60%

Precision=44+2≈67%

As you can clearly see, sensitivity has the highest value.