Q & A

17 January 2020 04:45

Sigmoid Curve

This is the sigmoid curve equation: $y = P(Diabetes) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$. Here, let's say you take β_0 = -15 and β_1 = 0.065. Now, what will be the probability of diabetes for a patient with sugar level 220?

0.5 ✗ Incorrect

0.33 ✓ Correct

Feedback:

Here, the probability of diabetes for a person with sugar level x is given by $P(Diabetes) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$. Now, taking

$\beta_0 = -15$ and $\beta_1 = 0.065$, the probability of diabetes for a person with sugar level 220 will be given by $P(Diabetes) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

$$P = \frac{1}{1+e^{-(-15+0.065*220)}} = 0.33\,\beta_1$$

2.) For the sigmoid curve (β0 = -15 and β1 = 0.065), what will be the probability of diabetes for a patient with sugar level 240?
0.645

3.)
Now, let's say that for the ten points in our example, the labels are as follows:

Point no.	1	2	3	4	5	6	7	8	9	10
Diabetes	no	no	no	yes	no	yes	yes	yes	yes	yes

In this case, the likelihood would be equal to:
(1−P1)(1−P2)(1−P3)(1−P5)(P4)(P6)(P7)(P8)(P9)(P10)
Feedback : *Recall that likelihood is the product of (1−Pi) for all non-diabetic patients and (Pi) for all diabetic patients. Hence, the likelihood is given by (1−P1)(1−P2)(1−P3)(1−P5), (all non-diabetic patients) multiplied by (P4)(P6)(P7)(P8)(P9)(P10) (all diabetic patients).*

4.)

Log Odds

So, let's say that the equation for log odds is:

$$ln(\frac{P}{1-P}) = -13.5 + 0.06x$$

For x = 220, the log odds are equal to -0.3 and for x = 231.5, the log odds are equal to 0.39. For x = 243, the log odds are equal to:

1.08 ✓ Correct

Feedback:

For a given value of x, the log odds are equal to -13.5 + 0.06x. Putting in the value of x here, i.e. 243, you get that the log odds = 1.08. However, you can actually directly guess the answer, without any calculations. The last time you increased x by 11.5, i.e. from 220 to 231.5, the log odds increased by 0.69, i.e. from -0.3 to 0.39. Since the relationship between x and log odds is linear, when you increase x by 11.5 again to make it 243, the log odds will increase by 0.69 again to get to 1.08.

5.)

X	Odds
220	0.74
231.5	1.48
243	2.96

So, the odds for sugar level of 254.5 will be closest to which of the following values?

- 5
- 5.44
- 5.92 ✓ Correct

Feedback :

As you can see in the table, and as has been said by the professor, every time the value of x increases by 11.5, the value of the odds approximately doubles. Hence, if you increase x from 243 to 254.5, the odds will approximately double, from 2.96 to approx. 5.92.

6.) Standardising Variables
In a dataset with mean 50 and standard deviation 12, what will be the value of a variable with an initial value of 20 after you standardise it?
-2.5
Feedback :
The formula for standardising a value in a dataset is given by:
$(X-\mu)/\sigma$
Hence, you get:
$20-5012=-2.5$

7.)Standardising the train and test sets
As Rahim mentioned in the lecture, you use **'fit_transform'** on the train set but just **'transform'** on the test set. Recall you had learnt this in linear regression as well. Why do you think this is done?

The **'fit_transform'** command first fits the data to have a mean of 0 and a standard deviation of 1, i.e. it scales all the variables using:
$Xscaled=X-\mu\sigma$
Now, once this is done, all the variables are transformed using this formula. Now, when you go ahead to the test set, you want the variables to not learn anything new. You want to use the old centralisation that you had when you used fit on the train dataset. And this is why you don't apply 'fit' on the test data, just the 'transform'.
You can also refer to this StackOverflow answer.

8.)Which of the following command can be used to view the **correlation table** for the dataframe **telecom**?
telecom.corr()
Feedback :
Correct!
***telecom.corr()** will give you the correlation table for the dataframe telecom.*

9.) Checking Correlations
Take a look at the heatmap provided above. Which of the variables have the highest correlation between them?
MultipleLines_No and MultipleLines_Yes
Feedback :
The following are the correlation values between the four pair of variables given in the options:

1. *0.53*
2. *0.54*
3. *-0.82*
4. *-0.64*

As you can clearly see, the third pair, i.e. MultipleLines_No and MultipleLines_yes is the most correlated with a value of -0.82.

10.)Significant Variables
Which of the following variables are insignificant as of now based on the summary statistics above? (More than one option may be correct.)
Note: Use p-value to determine the insignificant variables.
PhoneService
Feedback :
Correct! For a variable to be insignificant, the p-value should be greater than 0.05. For this variable, the p-value is 0.228 which is clearly greater than 0.05.
Correct
TechSupport_Yes
Feedback :

Correct! For a variable to be insignificant, the p-value should be greater than 0.05. For this variable, the p-value is 0.888 which is clearly greater than 0.05.

11.) Negatively Correlated Variables
Which of the following variables are negatively correlated with the target variable based on the summary statistics given above? (More than one option may be correct.)
tenure
Feedback :
Correct! Recall to check whether a variable is positively or negatively correlated with the target variable, you simply need to see the sign on its coefficient. The coefficient for 'tenure' is -1.5172 which is indeed negative and hence, there is a negative correlation between the target variable and tenure.
Correct
MonthlyCharges
Feedback :
Correct! Recall to check whether a variable is positively or negatively correlated with the target variable, you simply need to see the sign on its coefficient. The coefficient for 'MonthlyCharges' is -2.1806 which is indeed negative and hence, there is a negative correlation between the target variable and MonthlyCharges.
Correct
TechSupport_Yes
Feedback :
Correct! Recall to check whether a variable is positively or negatively correlated with the target variable, you simply need to see the sign on its coefficient. The coefficient for 'TechSupport_Yes' is -0.0305 which is indeed negative and hence, there is a negative correlation between the target variable and TechSupport_Yes.

12.) p-values
After learning the coefficients of each variable, the model also produces a 'p-value' of each coefficient. Fill in the blanks so that the statement is correct:
"The null hypothesis is that the coefficient is __. If the p-value is small, you can say that the coefficient is significant and hence the null hypothesis ____."
zero, can be rejected
Feedback :
Yes! Recall that the null hypothesis for any beta was:
$\beta_i=0$
And if the p-value is small, you can say that the coefficient is significant, and hence, you can reject the null hypothesis that $\beta_i=0$

13.) Threshold Value
You saw that Rahim chose a cut-off of 0.5. What can be said about this threshold?
It was arbitrarily chosen by us, i.e. there's nothing special about 0.5. We could have chosen something else as well.
Feedback :
Correct! The threshold of 0.5 chosen as of now is completely arbitrary. You will learn how to choose an optimal threshold during model evaluation.

14.)Significance based on RFE
Based on the RFE output shown above, which of the variables is least significant?
gender_Male
Feedback :
*Correct! Recall that RFE assigns ranks to the different variables based on their significance. While 1 means that the variable should be selected, a rank > 1 tells you that the variable is insignificant. The ranking given to 'gender_Male' by RFE is **9** which is the highest and hence, it is the most insignificant variable present in the RFE output.*

15.)Churn based on Threshold
Suppose the following table shows the predicted values for the probabilities for 'Churn'. Assuming you chose an arbitrary cut-off of 0.5 wherein a probability of greater than 0.5 means the customer would churn and a probability of less than or equal 0.5 means the customer wouldn't churn, which of these customers do you think will churn? (More than one option may be correct.)

Customer	Probability(Churn)
A	0.45
B	0.67
C	0.98
D	0.49
E	0.03

B
Feedback :
The threshold mentioned in the question for churning is given to be 0.5 which means that the customers with a churn probability > 0.5 will churn and those with a churn probability of < 0.5. For customer B, the churn probability is 0.67 which is greater than 0.5 and hence, customer B will churn based on the decided threshold.
Correct
C
Feedback :
The threshold mentioned in the question for churning is given to be 0.5 which means that the

customers with a churn probability > 0.5 will churn and those with a churn probability of < 0.5. For customer C, the churn probability is 0.98 which is greater than 0.5 and hence, customer C will churn based on the decided threshold.

16.) Calculating Accuracy
From the confusion matrix you saw in the last question, compute the accuracy of the model.

Actual/Predicted	Not Churn	Churn
Not Churn	80	30
Churn	20	70

75%

Feedback :

Correct! The accuracy of a model is given by:

$$Accuracy = \frac{Correctly\ precited\ labels}{Total\ Number\ of\ Labels}$$

Here, the number of correctly predicted labels are present in the first row, first column and the last row, last column.

Hence, you get -

Correctly predicted labels = 80 + 70 = 150

And the total number of labels is simply the sum of all the numbers present in the confusion matrix. Therefore,

Total number of labels = 80 + 30 + 20 + 70 = 200

Hence, you get -

$$Accuracy = \frac{150}{200} = 75\%$$

17.) Confusion Matrix and Accuracy
Given the confusion matrix below, can you tell how many 'Churns' were correctly identified, i.e. if the person has actually churned, it is predicted as a churn?

Actual/Predicted	Not Churn	Churn
Not Churn	80	30
Churn	20	70

70
Feedback :
Yes! Look at the table carefully. The value in the last row and last column will give you this number. You can see that there are 70 people who had actually churned and were also predicted as churn.

18.) Confusion Matrix
Suppose you built a logistic regression model to predict whether a patient has lung cancer or not and you get the following confusion matrix as the output.

Actual/Predicted	No	Yes
No	400	100
Yes	50	150

How many of the patients were wrongly identified as a 'Yes'?
100
Feedback :
Look at the table carefully. The value in the first row and the second column will tell you this number. Hence, you get 100 patients which actually didn't have lung cancer but were identified as having lung cancer.

19.)Confusion Matrix
Take a look at the table again.

Actual/Predicted	No	Yes
No	400	100
Yes	50	150

How many of these patients were correctly labelled, i.e. if the patient had lung cancer it was actually predicted as a 'Yes' and if they didn't have lung cancer, it was actually predicted as a 'No'?

550

Feedback :

The sum of values of the numbers in the first row, first column and the last row, last column will give you the answer.

Actual/Predicted	No	Yes
No	400	100
Yes	50	150

From the table above, the value in the first row, first column is 400, and the value in the last row, last column is 150. Hence, you get the total correctly predicted labels as 400 + 150 = 550

20.) Accuracy Calculation

From the table you used for the last two questions, what will be the accuracy of the model?

Actual/Predicted	No	Yes
No	400	100
Yes	50	150

78.57%

Feedback :

The accuracy of a model is given by:

$$Accuracy = \frac{Correctly\ precited\ labels}{Total\ Number\ of\ Labels}$$

The number of correctly predicted labels as you found out from the last question is equal to 550. The total number of labels is (400 + 100 + 50 + 150) = 700. Hence, the accuracy becomes:

$$Accuracy = \frac{550}{700} \approx 78.57\%$$

21.)

Multivariate Logistic Regression (Variable Selection)

Based on the above information, what can you say about the log odds of these two customers?

PS: *Recall the log odds for univariate logistic regression was given as:*

$$ln(\frac{P}{1-P}) = \beta_0 + \beta_1 X$$

Hence, for multivariate logistic regression, it would simply become:

$$ln(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_n X_n$$

log odds (customer A) > log odds (customer B)

Feedback :

Recall the log odds are just the linear term present in the logistic regression equation. Hence, here we have 13 variables, so the log odds will be given by:

$$ln(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{13} X_{13}$$

Now, for the two customers, all beta and all x values are the same, except for X_2 (the variable for paperless billing), which is equal to 1 for customer A and 0 for customer B.

Hence, the value will exceed by the coefficient of 'PaperlessBilling' which is 0.3367.

*Basically, for customer A, this term would be = 0.3367 * 1*

*And for customer B, this term would be = 0.3367 * 0*

22.) Multivariate Logistic Regression (Variable Selection)
Now, what can you say about the odds of churn for these two customers?
For customer A, the odds of churning are higher than for customer B
Feedback :
Recall that in the last question, you were told that log odds for customer A are higher than those for customer B. So, the odds of churning for customer A are also higher than the odds of churning for customer B. This is because, as the number increases, its log increases and vice versa.

23.)Multivariate Logistic Regression - Log Odds
Multivariate Logistic Regression - Log Odds
Now, suppose two customers, customer C and customer D, are such that their behaviour is exactly the same, except for the fact that customer C has OnlineSecurity, while customer D does not. What can you say about the odds of churn for these two customers?
For customer C, the odds of churning are lower than for customer D
Feedback :
Recall that the log odds for customer C will differ from those for customer D, by a margin of βOnlineSecurity. Now since in this case, this coefficient isds negative (-0.3739), this means that the log odds of customer C will be 0.3739 less than that of customer D. Since the log odds of customer C are lower, naturally, the actual odds for C would also be lower.

24.)

False Positives

What is the number of False Positives for the model given below?

Actual/Predicted	Not Churn	Churn
Not Churn	400	100
Churn	50	150

400

100 ✓ Correct

Feedback :

Correct! The false positives are the values which were actually 'Not Churn' but have been predicted as Churn. Hence, from the matrix above, the answer would be 100.

You can also have a look at the labelled confusion matrix you just learnt about:

Actual/Predicted	*Not Churn*	*Churn*
Not Churn	*True Negatives*	*False Positives*
Churn	*False Negatives*	*True Positives*

25.)

Sensitivity

Sensitivity is defined as the fraction of the number of correctly predicted positives and the total number of actual positives, i.e.

$$Sensitivity = \frac{TP}{(TP+FN)}$$

What is the sensitivity of the following model?

Actual/Predicted	Not Churn	Churn
Not Churn	400	100
Churn	50	150

- 60%
- **75%** ✓

  Feedback:

  Sensitivity is given as:

  $$Sensitivity = \frac{TP}{(TP+FN)}$$

  Here, TP (True Positives) = 150

  and FN (False Negatives) = 50

  Hence, you get:

  $$Sensitivity = \frac{150}{(150+50)} = 75\%$$

26.)

Evaluation Metrics

Among the three metrics that you've learnt about, which one is the highest for the model below?

Actual/Predicted	Not Churn	Churn
Not Churn	400	100
Churn	50	150

- Accuracy
- Sensitivity
- **Specificity** ✓

  Feedback:

  The formula for the three metrics are given as:

  $$Accuracy = \frac{Correctly\ Predicted\ Labels}{Total\ Number\ of\ Labels}$$

  $$Sensitivity = \frac{Number\ of\ actual\ Yeses\ correctly\ predicted}{Total\ number\ of\ actual\ Yeses} = \frac{TP}{TP+FN}$$

  $$Specificity = \frac{Number\ of\ actual\ Nos\ correctly\ predicted}{Total\ number\ of\ actual\ Nos} = \frac{TN}{TN+FP}$$

  Hence, you get:

  $$Accuracy = \frac{400+150}{400+100+50+150} = 78.57\%$$

  $$Sensitivity = \frac{150}{150+50} = 75\%$$

  $$Specificity = \frac{400}{400+100} = 80\%$$

  As you can clearly see, Specificity (80%) is the highest among the three.

27.)

False Negatives

What is the number of False Negatives for the model given below?

Actual/Predicted	Not Churn	Churn
Not Churn	80	40
Churn	30	50

- 80
- 40
- **30** ✓ Correct

Feedback :

The false negatives are the values which were actually 'Churn' but have been predicted as 'Not Churn'. Recall the labelling for the confusion matrix:

Actual/Predicted	*Not Churn*	*Churn*
Not Churn	*True Negatives*	*False Positives*
Churn	*False Negatives*	*True Positives*

Hence, you can see from the matrix above that the element in the 2nd row, 1st column gives you the value of 'False Negatives'. From the model given in the question, you can see that this number is equal to 30.

28.)

Specificity

Specificity is defined as the fraction of the number of correctly predicted negatives and the total number of actual negatives, i.e.

$$Specificity = \frac{TN}{(TN+FP)}$$

What is the approximate specificity of the following model?

Actual/Predicted	Not Churn	Churn
Not Churn	80	40
Churn	30	50

- 60%
- **67%**

Feedback :

Specificity is given as:

$$Specificity = \frac{TN}{(TN+FP)}$$

Here, TN (True Negatives) = 80

and FP (False Positives) = 40

Hence, you get:

$$Specificity = \frac{80}{(80+40)} = 66.67\% \approx 67\%$$

29.)

Evaluation Metrics

Which among accuracy, sensitivity, and specificity is the highest for the model below?

Actual/Predicted	Not Churn	Churn
Not Churn	80	40
Churn	30	50

- ○ Accuracy
- ○ Sensitivity
- ◉ **Specificity**

Feedback:

The formula for the three metrics are given as:

$$Accuracy = \frac{Correctly\ Predicted\ Labels}{Total\ Number\ of\ Labels}$$

$$Sensitivity = \frac{Number\ of\ actual\ Yeses\ correctly\ predicted}{Total\ number\ of\ actual\ Yeses} = \frac{TP}{TP+FN}$$

$$Specificity = \frac{Number\ of\ actual\ Nos\ correctly\ predicted}{Total\ number\ of\ actual\ Nos} = \frac{TN}{TN+FP}$$

Hence, you get:

$$Accuracy = \frac{80+50}{80+40+30+50} = 65\%$$

$$Sensitivity = \frac{50}{30+50} = 62.5\%$$

$$Specificity = \frac{80}{80+40} \approx 67\%$$

As you can clearly see, Specificity (~67%) is the highest among the three.

30.) Other Metrics

In the code, you saw Rahim evaluate some other metrics as well. These were:

$$False\ Positive\ Rate = \frac{FP}{TN+FP}$$

$$Positive\ Predictive\ Value = \frac{TP}{TP+FP}$$

$$Negative\ Predictive\ Value = \frac{TN}{TN+FN}$$

- As you can see, the 'False Positive Rate' is basically (1 - Specificity). Check the formula and the values in the code to verify.
- The positive predictive value is **the number of positives correctly predicted by the total number of positives predicted**. This is also known as **'Precision'** which you'll learn more about soon.
- Similarly, the negative predictive value is **the number of negatives correctly predicted by the total number of negatives predicted**. There's no particular term for this as such.

Calculate the given three metrics for the model below and identify which one is the largest among them.

Actual/Predicted	Not Churn	Churn
Not Churn	80	40
Churn	30	50

- ◉ **Negative Predictive Value**

Feedback:

Correct! The values that you'll get are:

$$False\ Positive\ Rate = \frac{FP}{TN+FP} = \frac{40}{80+40} \approx 33\%$$

You could have also used the specificity value you calculated in the last question (~67%) and simply calculated this as 1-Specificity = 1 - 0.67 = 33%

$$Positive\ Predictive\ Value = \frac{TP}{TP+FP} = \frac{50}{50+40} = 55.55\% \approx 56\%$$

$$Negative\ Predictive\ Value = \frac{TN}{TN+FN} = \frac{80}{80+30} \approx 72.72\% \approx 73\%$$

As you can clearly see, the Negative Predictive Value is the highest of the three.

31.)TPR and FPR
Given the following confusion matrix, calculate the value of True Positive Rate (TPR) and False Positive Rate (FPR).

Actual/Predicted	Not Churn	Churn
Not Churn	300	200
Churn	100	400

TPR = 80%
FPR = 40%

Feedback :

Correct! Recall the formulas for TPR and FPR were:

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN}$$

Here,

TP = 400; FN = 100; FP = 200; TN = 300

Hence, you get -

$$TPR = \frac{400}{400+100} = 80\%$$

$$FPR = \frac{200}{200+300} = 40\%$$

32.)True Positive Rate
You have the following table showcasing the actual 'Churn' labels and the predicted probabilities for 5 customers.

Customer	Churn	Predicted Churn Probability
Thulasi	1	0.52
Aditi	0	0.56
Jaideep	1	0.78
Ashok	0	0.45
Amulya	0	0.22

Calculate the True Positive Rate and False Positive rate for the cutoffs of 0.4 and 0.5. Which of these cutoffs, will give you a better model?
Note: The good model is the one in which TPR is high and FPR is low.

Cutoff of 0.5
Feedback :
Now, at the cutoff of 0.4, you get the following values of predicted probabilities:

Customer	Churn	Predicted Churn Probability	Predicted Churn Label
Thulasi	1	0.52	1
Aditi	0	0.56	1
Jaideep	1	0.78	1
Ashok	0	0.45	1
Amulya	0	0.22	0

From the above table, you can easily calculate:

True Positives = 2

False Positives = 2

Also, from the original table, you have:

Actual Positives = 2

Actual Negatives = 3

Hence, you get:

$$TPR = \frac{True\ Positives}{Total\ Actual\ Positives} = \frac{2}{2} = 100\%$$

$$FPR = \frac{False\ Positives}{Total\ Actual\ Negatives} = \frac{2}{3} \approx 67\%$$

Performing similar steps for a cutoff of 0.5 will give you -

$$TPR = 100\%$$

$$FPR \approx 33\%$$

(Do calculate it yourself to verify)

As you can see, with both the cutoffs, the TPR is 100% but for the cutoff 0f 0.5 you have a lower value of FPR. So clearly, a cutoff of 0.5 gives you a better model.

Please note that 0.5 just gives the better model among 0.4 and 0.5. It might be possible that there is a cutoff point which gives an even better model.

33.)Changing the Threshold
You initially chose a threshold of 0.5 wherein a churn probability of greater than 0.5 would result in the customer being identified as 'Churn' and a churn probability of lesser than 0.5 would result in the customer being identified as 'Not Churn'.
Now, suppose you decreased the threshold to a value of 0.3. What will be its effect on the classification?
More customers would now be classified as 'Churn'.
Feedback :
Correct! Now since you have decreased the cutoff to 0.3, it would mean that:
Customers with churn probability > 0.3 will be identified as 'Churn'.
Customers with churn probability < 0.3 will be identified as 'Not Churn'.
Initially, the threshold was 0.5. Look at the customers in the 0.3-0.5 probability range. They were being identified as 'Not Churn' before, but now, are being identified as 'Churn'. Hence, naturally, the number of people being identified as 'Churn' will increase.

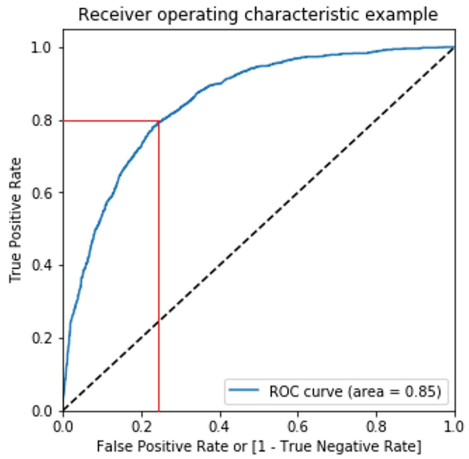
34.) TPR and FPR
Fill in the blanks:
When the value of TPR increases, the value of FPR _____.
increases
Feedback :
Correct! This can be clearly seen from the ROC curve as well. When the value of TPR (on the Y-axis) is increasing, the value of FPR (on the X-axis) also increases.

35.)Area Under the Curve
You have the following five AUCs (Area under the curve) for ROCs plotted for five different models. Which of these models is the best?

Model	AUC
A	0.54
B	0.82
C	0.79
D	0.66
E	0.56

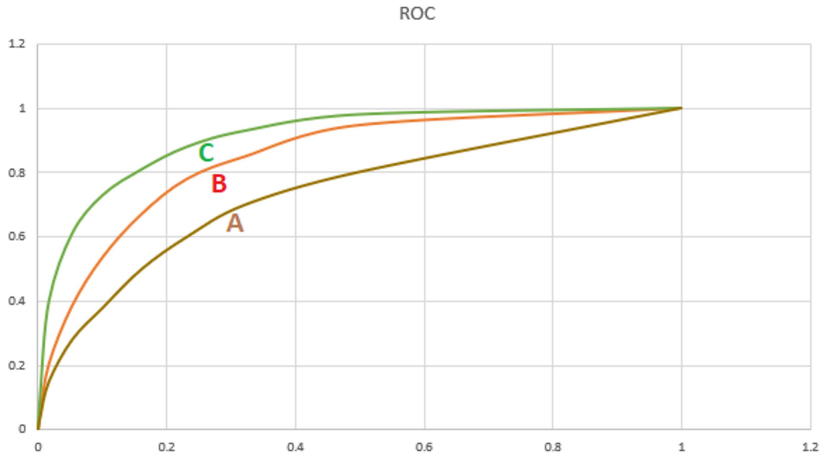
B
Feedback :
Correct!
Recall that when the ROC curve is more towards the top left corner of the graph, the model is deemed to be more accurate. Hence, a greater area under the curve would mean the model is more accurate. Among the five models given, B has the highest AUC and hence is the most accurate model. Also, note that the highest value of AUC can be 1.

36.)ROC Curve
Following is the ROC curve that you got.



As you can see, when the 'True Positive Rate' is 0.8, the 'False Positive Rate' is about 0.24. What will be the value of specificity, then?
0.76
Feedback :
Correct!
Recall that the False Postive Rate is nothing but (1 - True Negative Rate) and the True Negative Rate is simply the specificity. Hence,
False Positive Rate = 1 - Specificity
or, Specificity = 1 - False Postive Rate
Here, the False Positive Rate is 0.24. Therefore, Specificity = (1 - 0.24) = 0.76.

37.)ROC Curve
Which of the following ROC curve represents the best model?



C
Feedback :
Yes! Recall that the area under the curve tells you how good a model is. If the curve is more towards the top-left corner, area is more, and hence, the model is better. As you can see, of the three curves, curve 'C' is most towards the top-left corner and thus, has the highest area resulting in it being the best model.

38.)Choosing the Optimal Cut-off
Suppose you created a dataframe to find out the optimal cut-off point for a model you built. The dataframe looks like the following:

Threshold	Probability	Accuracy	Sensitivity	Specificity
0.0	0.0	0.21	1.00	0.00
0.1	0.1	0.39	0.96	0.22
0.2	0.2	0.56	0.88	0.49
0.3	0.3	0.59	0.81	0.53
0.4	0.4	0.62	0.78	0.63
0.5	0.5	0.74	0.73	0.74
0.6	0.6	0.81	0.64	0.79
0.7	0.7	0.78	0.42	0.83
0.8	0.8	0.63	0.21	0.92
0.9	0.9	0.56	0.03	0.98

Based on the table above, what will the approximate value of the optimal cut-off be?

0.5
Feedback :
Correct! The optimal cut-off point exists where the values of accuracy, sensitivity, and specificity are fairly decent and almost equal. At the cut-off of 0.5, the metric values are 0.74, 0.73, and 0.74 respectively. This is the optimal value of threshold that you can have.

39.)Choosing a model evaluation metric
As you learnt, there is usually a trade-off between various model evaluation metrics, and you cannot maximise all of them simultaneously. For e.g., if you increase sensitivity (% of correctly predicted churns), the specificity (% of correctly predicted non-churns) will reduce.

Let's say that you are building a telecom churn prediction model with the business objective that your company wants to implement an aggressive customer retention campaign to retain the 'high churn-risk' customers. This is because a competitor has launched extremely low-cost mobile plans, and you want to avoid churn as much as possible by incentivising the customers. Assume that budget is not a constraint.

Which of the following metrics should you choose the maximise?
Sensitivity
Feedback :
Yes, high sensitivity implies that your model will correctly identify almost all customers who are likely to churn. It will do that by over-estimating the churn likelihood, i.e. it will misclassify some non-churns as churns, but that is the trade-off you need to choose rather than the opposite case (in which case you may lose some low churn risk customers to the competition).

40.)Accuracy of the Model
Using the threshold of 0.3, what is the approximate accuracy of the model now?
77%
Feedback :
Correct!
Use the following code to calculate the accuracy:
metrics.accuracy_score(y_train_pred_final.Churn, y_train_pred_final.final_predicted)
You'll see that you get an accuracy of about 77.14%.

41.)Confusion Matrix
Get the confusion matrix after using the cut-off 0.3. What is the number of 'False Negatives' now?
2793
842
283
Feedback :
Correct! When you run the following code to get the confusion matrix,
confusion2 = metrics.confusion_matrix(y_train_pred_final.Churn, y_train_pred_final.final_predicted)

you'll get the following confusion matrix:

Actual/Predicted	Not Churn	Churn
Not Churn	2793	842
Churn	283	1004

Also, recall that the labels in the confusion matrix are:

Actual/Predicted	Not Churn	Churn
Not Churn	True Negatives	False Positives
Churn	False Negatives	True Positives

You can clearly see that the number of 'False Negatives' is now 283. Also, note that the number of 'False Negatives' has now dropped significantly and the number of 'True Positives' has increased. Thus, choosing a lower cut-off has definitely helped in capturing the 'Churns' better

42.) Sensitivity
In the last question you saw that in the confusion matrix, the Churns are being captured better now. Using the confusion matrix, can you tell what will the approximate sensitivity of the model now be?
78
Feedback :
Correct!
The sensitivity is given as:
Sensitivity=TPTP+FN
Hence, you get:
Sensitivity=10041004+283≈78.01%
It is recommended that you calculate this in your Jupyter notebook as well.

43.)

Calculating Precision

Calculate the precision value for the following model.

Actual/Predicted	Not Churn	Churn
Not Churn	400	100
Churn	50	150

- (●) **60%**

  Feedback :

  Correct! The formula for precision is given by:

  $$Precision = \frac{TP}{TP+FP}$$

  From the matrix given,

  TP = 150

  FP = 100

  Hence, you get,

  $$Precision = \frac{150}{150+100} = 60\%$$

44.)

There is a measure known as **F1-score** which essentially combines both precision and recall. It is the basically the harmonic mean of precision and recall and its formula is given by:

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

The F1-score is useful when you want to look at the performance of precision and recall together.

Calculate the F1-score for the model below:

Actual/Predicted	Not Churn	Churn
Not Churn	400	100
Churn	50	150

- ( ) 33%

- (●) **67%**

  Feedback :

  Correct!

  From the confusion matrix given,

  TP = 150

  FP = 100

  FN = 50

  Hence, you get -

  $$Precision = \frac{150}{100+150} = 0.6$$

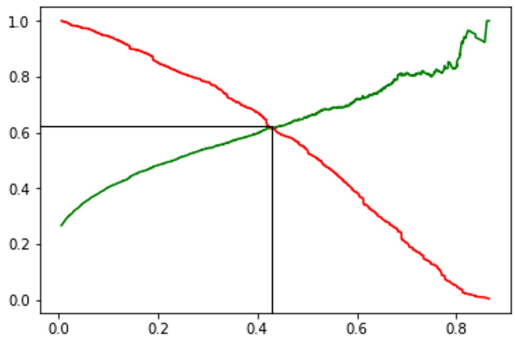
  $$Recall = \frac{150}{150+50} = 0.75$$

  So, the F1-score becomes -

  $$F = 2 \times \frac{0.6 \times 0.75}{0.6+0.75} \approx 66.67\% \approx 67\%$$

45.)Optimal Cut-off

When using the sensitivity-specificity tradeoff, you found out that the optimal cutoff point was 0.3. Now, when you plotted the precision-recall tradeoff, you got the following curve:



What is the optimal cutoff point according to the curve given above?
0.42
Feedback :
Yes! The optimal cutoff point is where the values of precision and recall will be equal. This is similar to what you saw in the sensitivity-specificity tradeoff curve as well. So, when precision and recall are both around 0.62, the two curves are intersecting. And at this place, if you extend the line to the X-axis as given, you can see that the threshold value is 0.42.

46.)

Calculating Accuracy

Recall that in the last segment you saw that the cutoff based on the precision-recall tradeoff curve was approximately 0.42. When you take this cut-off, you get the following confusion matrix on the test set.

Actual/Predicted	Not Churn	Churn
Not Churn	1294	234
Churn	223	359

What will the approximate value of accuracy be on the test set now?

- 60%
- 72%
- 75%
- **78%**

  Feedback :

  Correct!

  Recall that the accuracy of the model is given by:

  $$Accuracy = \frac{Correctly\ Predicted\ Labels}{Total\ Number\ of\ Labels}$$

  Hence, you get -

  $$Accuracy = \frac{1294+359}{1294+234+223+359} \approx 78.34\%$$

47.)

Calculating Recall

For the confusion matrix you saw in the last question, what will the approximate value of recall be?

Actual/Predicted	Not Churn	Churn
Not Churn	1294	234
Churn	223	359

62%

Feedback :

Correct!

Recall that the recall is given by -

$$Recall = \frac{TP}{TP+FN}$$

Here,

TP = 359

FN = 223

Hence, you get -

$$Recall = \frac{359}{359+223} \approx 61.68\% \approx 62\%$$

It is recommended that you calculate these values in the Jupyter Notebook provided as well.

48.) What information would you infer from the woe trend of tenure variable?
As tenure increases, the chances of churning decrease
Feedback :
If you calculate woe value of all the 10 buckets, you would notice, as tenure increases from 1 year to 72 years, woe values are also increasing continuously from -1.46 to 2.12. Thus it results in decreasing the chances of churning over the bucket.

49.)Coarse binning is not required for tenure variable as there is a clear monotonic trend in fine binning
Feedback : *If you create a plot out of woe value, you can clearly visualise a monotonic plot.*

50.) Woe Analysis
What does negative woe signify in 'contract' variable (refer sheet-3)?
% of churners (bad customers) are more than % of no-churners (good customers)
Feedback :
The woe is expressed by ln (percentage of non-churns in bucket/ percentage of churns in the bucket). If the woe is negative, it means the percentage of churns in that bucket is greater than the percentage of non-churns in that bucket.

51.) Woe Analysis
Compare the woe trends of both variables (tenure and contract).
Based on the woe trend, which variable when increased in value, will decrease the likelihood of churn?
Both
Feedback :*"Tenure", as well as "Contract" both, negatively impacts churns rate over the bucket. It means that if the tenure increases, churn rate decreases and vice-versa. Similarly for contract variable as well, the two-year contract has the low chance of churn than one year of the contract.*

52.)What is the total information value of both the variables?
Contract = 1.24 , Tenure = 0.83
Feedback :*Information Value for each bucket can be calculated as: IVbucket = WOEbucket * (% good - % bad) Total IV for Tenure = IVbucket-1(0-1) + IVbucket-2(2-5) + IVbucket-3(6-11)+IVbucket-4(12-19) + IVbucket-5(20-28) + IVbucket-1(29-39) + IVbucket-2(40-49) + IVbucket-3(50-59)+IVbucket-4(60-68) + IVbucket-5(69-72) Total IV for Tenure = 0.23+0.12+0.02+0.01 +0.00+0.01+0.02+0.05+0.12+0.26 = 0.83 Total IV for Contract = IVbucket-1(month-to-month) + IVbucket-2(One-year) + IVbucket-3(two-year) Total IV for Contract = 0.33+0.17+0.74 = 1.24*

53.) Contract variable has stronger predictive power than tenure
Feedback :*Predictive power can be measured based on information values, higher the information value, higher the predictive power. In this example as well, Contract variable shows IV of 1.24 and Tenure variable shows IV of 0.83*

54.) Missing value
NA bucket can be merged with -
None
Feedback :*The woe value for NA bucket is nearly -0.51, whereas woe values for other buckets are very different from the NA bucket. So basically, there is not bucket which shows same woe values as NA does.*

55.) Fibonacci Series
Description
Compute and display Fibonacci series upto n terms where n is a positive integer entered by the user.
You can go here to read about Fibonacci series.
Sample Input:
5
Sample Output:
0
1
1
2
3

```
n=int(input())
def fibonacci(n):
   x1 = 0
   x2 = 1
   print(x1)
   for i in range(1,n):
      print(x2)
      x3= x1+x2
      x1 = x2
      x2 = x3
fibonacci(n)
```

```
n=int(input())
secondLast=0
last=1
if n>0:
   print(secondLast)
if n>1:
   print(last)
for i in range(3,n+1):
   nextNumber=last+secondLast
   print(nextNumber)
   secondLast=last
   last=nextNumber
```

56.) Prime Numbers

Description
Determine whether a positive integer n is a prime number or not. Assume n>1.
Display “number entered is prime” if n is prime, otherwise display “number entered is not prime”.
Sample Input:
7
Sample Output:

```
n=int(input())
if n > 1:
  for i in range(2,n):
     if (n % i) == 0:
        print("number entered is not prime")
        break
  else:
     print("number entered is prime")
else:
  print("number entered is not prime")
```

```
n=int(input())
out=True
for i in range(2,n):
   if(n%i==0):
      out=False
      break
if out==True:
   print("number entered is prime")
else:
   print("number entered is not prime")
```

57.)Armstrong number
Description
Any number, say n is called an Armstrong number if it is equal to the sum of its digits, where each is raised to the power of number of digits in n.

For example:
$153=1^3+5^3+3^3$

Write Python code to determine whether an entered three digit number is an Armstrong number or not.
Assume that the number entered will strictly be a three digit number.
Print "True" if it is an Armstrong number and print "False" if it is not.
Sample Input:
153
Sample Output:
True
Execution time limit
Default.

```
n=int(input())
asum = 0
a1 = n
while a1 > 0:
  dig = a1 % 10
  asum += dig ** 3
  a1 //= 10
if n == asum:
  print("True")
else:
  print("False")
```

```
n=int(input())
digits=list(map(int,str(n)))
num=sum(list(map(lambda x:x**3,digits)))
print(num==n)
```

58.) Selecting dataframe columns
Description
Write a program to select all columns of a dataframe except the ones specified.
The input will contain a list of columns that you should skip.
You should print the first five rows of the dataframe as output where the columns are **alphabetically sorted.**

Sample Input:
['PassengerId', 'Pclass', 'Name', 'Sex','Embarked']
Sample Output:

```
   Age Cabin    Fare  Parch  SibSp   Ticket
0  34.5   NaN  7.8292      0      0   330911
1  47.0   NaN  7.0000      0      1   363272
2  62.0   NaN  9.6875      0      0   240276
3  27.0   NaN  8.6625      0      0   315154
4  22.0   NaN 12.2875      1      1  3101298
```

Execution time limit

```
import pandas as pd
import ast,sys
df=pd.read_csv("https://media-doselect.s3.amazonaws.com/generic/X0kvr3wEYXRzONE5W37xWWYYA/test.csv")
input_str = sys.stdin.read()
to_omit = ast.literal_eval(input_str)
df=df[df.columns[~df.columns.isin(to_omit)]]
print(df.loc[:, sorted(list(df.columns))].head())
```

59.) Two series

Description
Given two pandas series, find the position of elements in series2 in series1.
You can assume that all elements in series2 will be present in series1.
The input will contain two lines with series1 and series2 respectively.
The output should be a list of indexes indicating elements of series2 in series 1.
Note: In the output list, the indexes should be in ascending order.
Sample Input:
[1,2,3,4,5,6,7]
[1,3,7]
Sample Output:
[0,2,6]

Execution time limit

```
import ast,sys
```

```
import pandas as pd
input_str = sys.stdin.read()
input_list = ast.literal_eval(input_str)
series1=pd.Series(input_list[0])
series2=pd.Series(input_list[1])
position =[]
for i in list(series2):
   if i in list(series1):
      pos=(list(series1).index(i))
      position.append(pos)
out_list=list(position)
print(list(map(int,out_list)))#do not alter this step,
```

```
import ast,sys
import pandas as pd
input_str = sys.stdin.read()
input_list = ast.literal_eval(input_str)
series1=pd.Series(input_list[0])
series2=pd.Series(input_list[1])
out_list=[pd.Index(series1).get_loc(num) for num in series2]
print(list(map(int,out_list)))
```

60.) Cleaning columns

Description

For the given dataframe, you have to clean the "Installs" column and print its correlation with other numeric columns of the dataframe.(print df.corr())

You have to do the following:

1. Remove characters like ',' from the number of installs.
2. Delete rows where the Installs column has irrelevant strings like 'Free'
3. Convert the column to int type

You can access the dataframe using the following URL in your Jupyter notebook:

https://media-doselect.s3.amazonaws.com/generic/8NMooe4G0ENEe8z9q5ZvaZA7/googleplaystore.csv

Note: You should try this problem on your own Jupyter notebook before submitting. Do not clean any column other than "Installs".

Sample Output:

```
          Rating  Installs
Rating    1.000000  0.051355
Installs  0.051355  1.000000
```

Execution time limit

```
import pandas as pd
df=pd.read_csv("https://media-doselect.s3.amazonaws.com/generic/8NMooe4G0ENEe8z9q5ZvaZA7/googleplaystore.csv")
df.Installs=df.Installs.str.replace(',','')
df.Installs=df.Installs.str.replace('+','')
df=df[df.Installs!='Free']
df.Installs=df.Installs.astype(int)
print(df.corr())
```