ML Concepts - Classification

What is Thresholding?

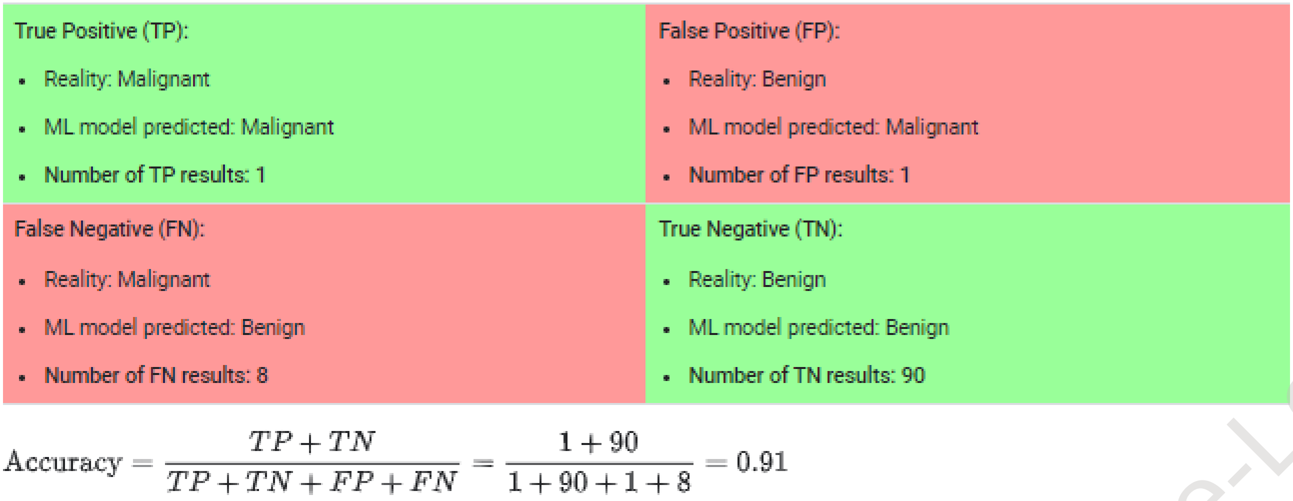
- Logistic regression returns a probability value between 0 and 1.
- You can use the returned probability as-is, for example, to estimate the likelihood of a user clicking on an ad.
- Alternatively, you can convert the probability to a binary value, such as classifying an email as spam or not spam.
- **Interpreting Probability Values**:
  - A high probability (e.g., 0.9995) suggests strong confidence in the prediction (e.g., very likely to be spam).
  - A low probability (e.g., 0.0003) indicates high confidence in the opposite prediction (e.g., very likely not spam).
- **Classification Threshold**:
  - To map a logistic regression value to a binary category, you need to set a classification threshold (decision threshold).
  - Values above the threshold indicate one class, while values below indicate the other.
- **Threshold Tuning**:
  - It's important to note that the choice of classification threshold is problem-dependent.
  - The threshold is a value you must tune based on your specific use case.
  - Changing the classification threshold can significantly affect the model's predictions.
- **Evaluation Metrics**:
  - Different metrics can be used to evaluate a classification model's predictions, such as accuracy, precision, recall, and F1-score.
- **Tuning Threshold vs. Hyperparameters**:
  - Tuning the threshold for logistic regression is a different process from tuning hyperparameters like learning rate.
  - It involves assessing the cost and impact of different types of classification errors.

In summary, choosing the right classification threshold is a crucial step in making accurate predictions with logistic regression. It's a problem-specific decision that requires considering the consequences of different types of classification errors.

True Vs False. Positive Vs Negative

- **True Positive (TP)**: The model correctly predicts the positive class.
- **True Negative (TN)**: The model correctly predicts the negative class.
- **False Positive (FP)**: The model incorrectly predicts the positive class.
- **False Negative (FN)**: The model incorrectly predicts the negative class.

Accuracy



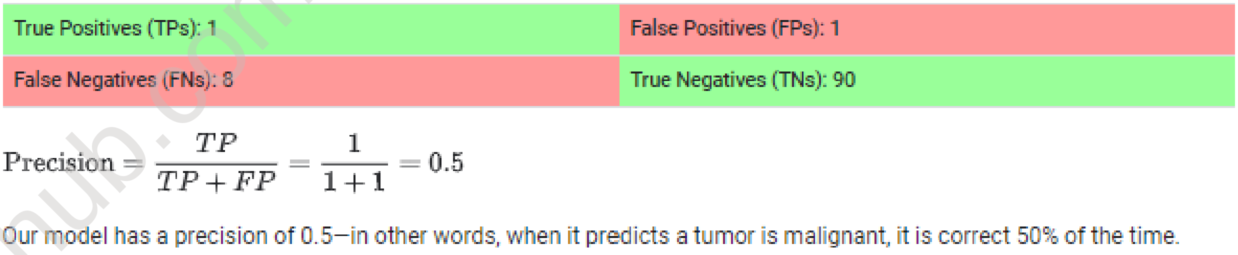
True Positive (TP): • Reality: Malignant • ML model predicted: Malignant • Number of TP results: 1	False Positive (FP): • Reality: Benign • ML model predicted: Malignant • Number of FP results: 1
False Negative (FN): • Reality: Malignant • ML model predicted: Benign • Number of FN results: 8	True Negative (TN): • Reality: Benign • ML model predicted: Benign • Number of TN results: 90

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1+90}{1+90+1+8} = 0.91$$

- Accuracy in binary classification is calculated using the formula: Accuracy = (TP + TN) / (TP + TN + FP + FN).
- In the example provided, there were 1 True Positive (TP), 1 False Positive (FP), 8 False Negatives (FN), and 90 True Negatives (TN).
- The calculated accuracy was 0.91 or 91%.
- However, a deeper analysis shows that out of 100 tumor examples, 9 were malignant and 91 were benign.
- The model correctly identified 90 out of 91 benign tumors but only 1 out of 9 malignant tumors.
- This highlights a significant issue, as 8 out of 9 malignancies were not detected.
- While the accuracy seems high, a model that always predicts benign would achieve the same accuracy.
- Accuracy alone is not sufficient for evaluating class-imbalanced datasets.
- The next section introduces precision and recall as better metrics for such scenarios.

Precision & Recall

- What proportion of positive identifications was actually correct?

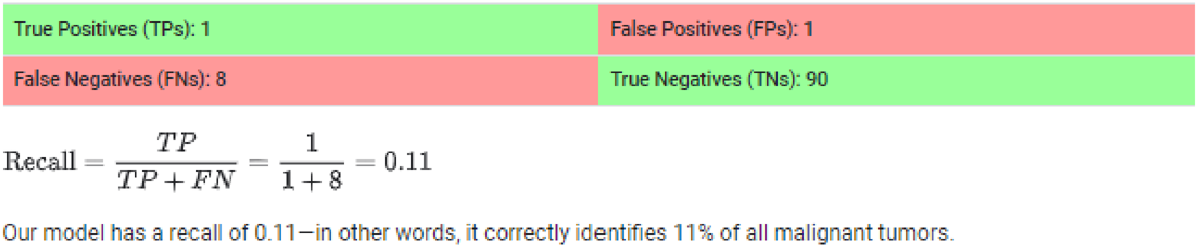


True Positives (TPs): 1	False Positives (FPs): 1
False Negatives (FNs): 8	True Negatives (TNs): 90

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{1}{1+1} = 0.5$$

Our model has a precision of 0.5—in other words, when it predicts a tumor is malignant, it is correct 50% of the time.

- What proportion of actual positives was identified correctly?



True Positives (TPs): 1	False Positives (FPs): 1
False Negatives (FNs): 8	True Negatives (TNs): 90

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{1}{1+8} = 0.11$$

Our model has a recall of 0.11—in other words, it correctly identifies 11% of all malignant tumors.

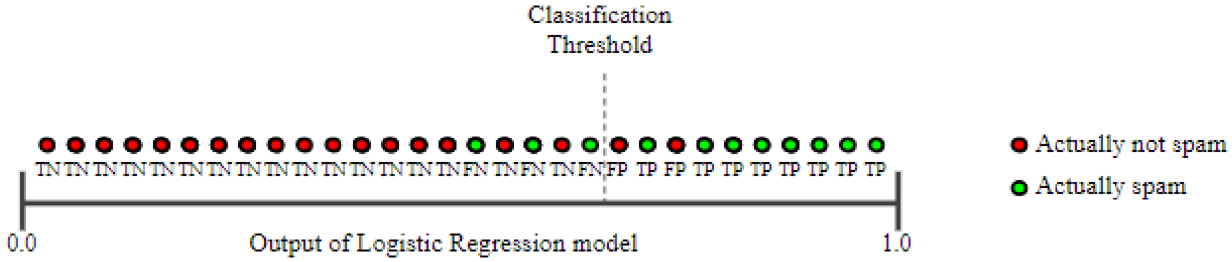

Classification
Threshold
TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN FN TN FN TN FN FP TP FP TP TP TP TP TP TP TP
0.0
Output of Logistic Regression model
1.0
● Actually not spam
● Actually spam

Figure 1. Classifying email messages as spam or not spam.

Let's calculate precision and recall based on the results shown in Figure 1:

True Positives (TP): 8	False Positives (FP): 2
False Negatives (FN): 3	True Negatives (TN): 17

Precision measures the percentage of **emails flagged as spam** that were correctly classified—that is, the percentage of dots to the right of the threshold line that are green in Figure 1:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{8}{8+2} = 0.8$$

Recall measures the percentage of **actual spam emails** that were correctly classified—that is, the percentage of green dots that are to the right of the threshold line in Figure 1:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{8}{8+3} = 0.73$$

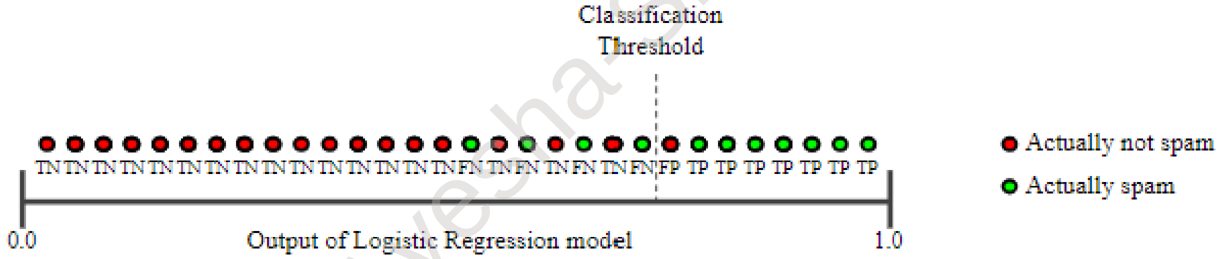

Classification
Threshold
TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN FN TN FN TN FN TN FN FP TP TP TP TP TP TP TP
0.0
Output of Logistic Regression model
1.0
● Actually not spam
● Actually spam

Figure 2. Increasing classification threshold.

The number of false positives decreases, but false negatives increase. As a result, precision increases, while recall decreases:

True Positives (TP): 7	False Positives (FP): 1
False Negatives (FN): 4	True Negatives (TN): 18

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{7}{7+1} = 0.88$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{7}{7+4} = 0.64$$

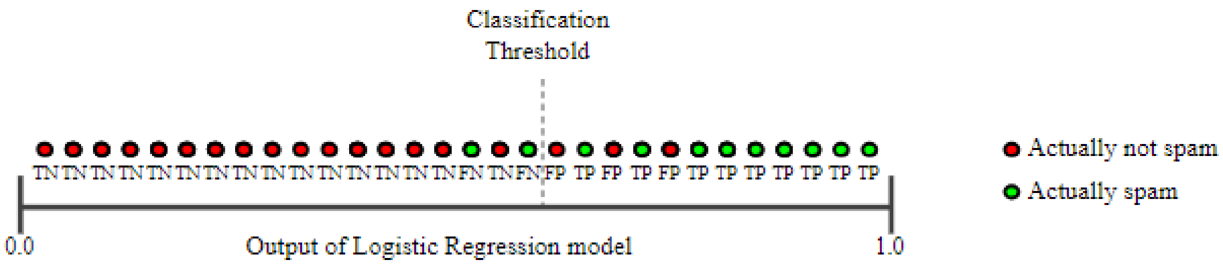

Classification Threshold
TN TN TN TN TN TN TN TN TN TN TN TN TN TN TN FN TN FN FP TP FP TP FP TP TP TP TP TP TP TP
0.0
Output of Logistic Regression model
1.0
Actually not spam
Actually spam

Figure 3. Decreasing classification threshold.

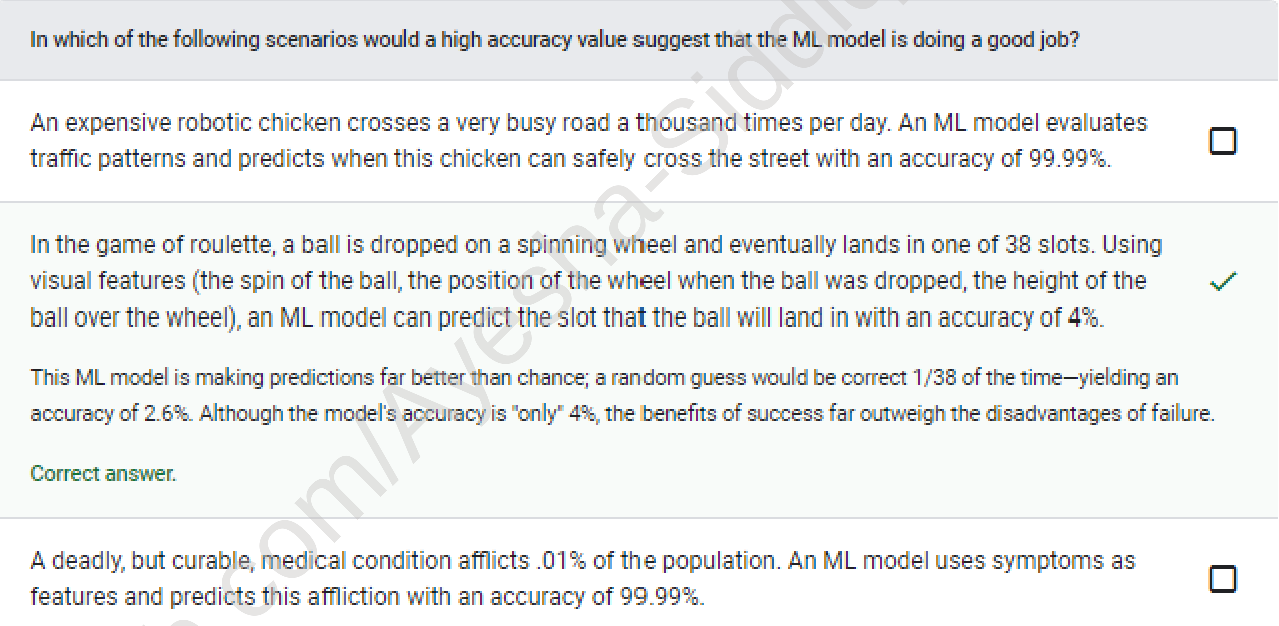
False positives increase, and false negatives decrease. As a result, this time, precision decreases and recall increases:

True Positives (TP): 9	False Positives (FP): 3
False Negatives (FN): 2	True Negatives (TN): 16

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{9}{9+3} = 0.75$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{9}{9+2} = 0.82$$

Quiz:



In which of the following scenarios would a high accuracy value suggest that the ML model is doing a good job?

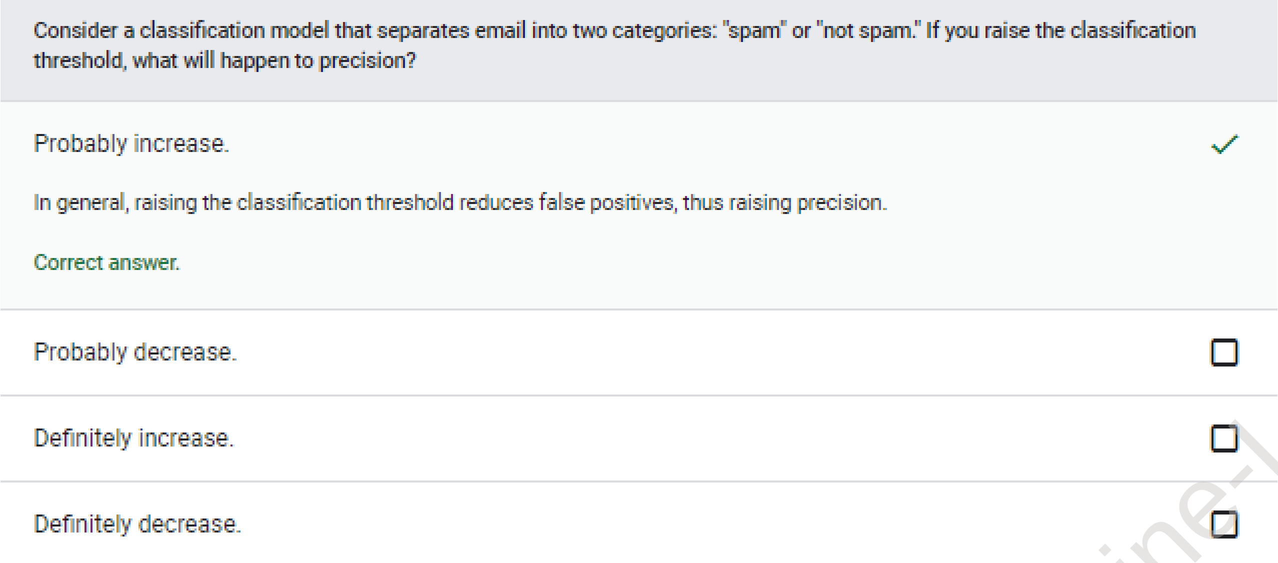
An expensive robotic chicken crosses a very busy road a thousand times per day. An ML model evaluates traffic patterns and predicts when this chicken can safely cross the street with an accuracy of 99.99%. □

In the game of roulette, a ball is dropped on a spinning wheel and eventually lands in one of 38 slots. Using visual features (the spin of the ball, the position of the wheel when the ball was dropped, the height of the ball over the wheel), an ML model can predict the slot that the ball will land in with an accuracy of 4%. ✓

This ML model is making predictions far better than chance; a random guess would be correct 1/38 of the time—yielding an accuracy of 2.6%. Although the model's accuracy is "only" 4%, the benefits of success far outweigh the disadvantages of failure.

Correct answer.

A deadly, but curable, medical condition afflicts .01% of the population. An ML model uses symptoms as features and predicts this affliction with an accuracy of 99.99%. □



Consider a classification model that separates email into two categories: "spam" or "not spam." If you raise the classification threshold, what will happen to precision?

Probably increase. ✓

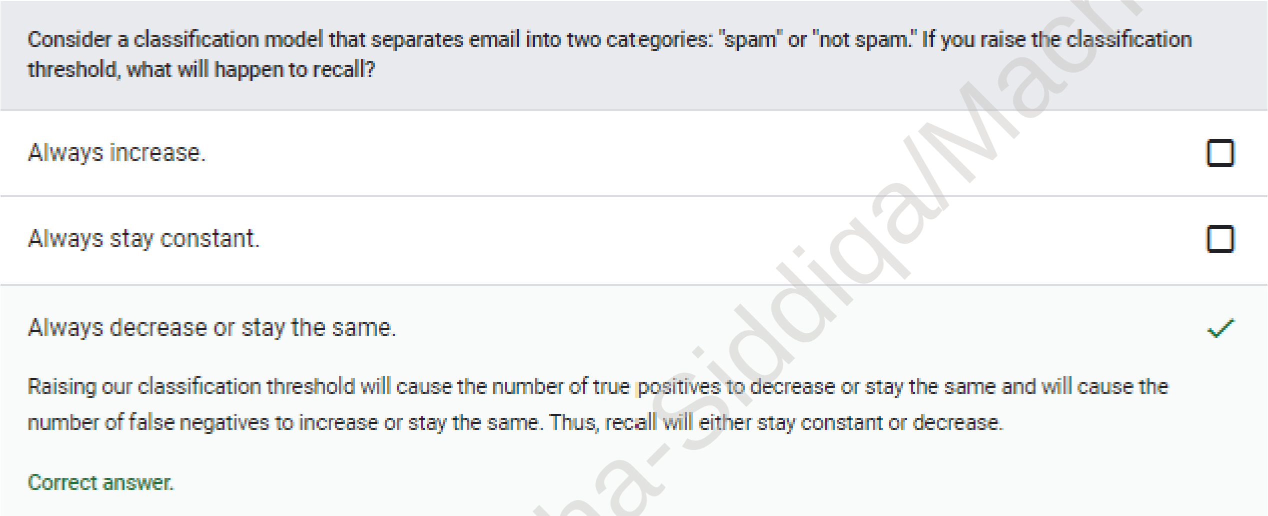
In general, raising the classification threshold reduces false positives, thus raising precision.

Correct answer.

Probably decrease. ☐

Definitely increase. ☐

Definitely decrease. ☐



Consider a classification model that separates email into two categories: "spam" or "not spam." If you raise the classification threshold, what will happen to recall?

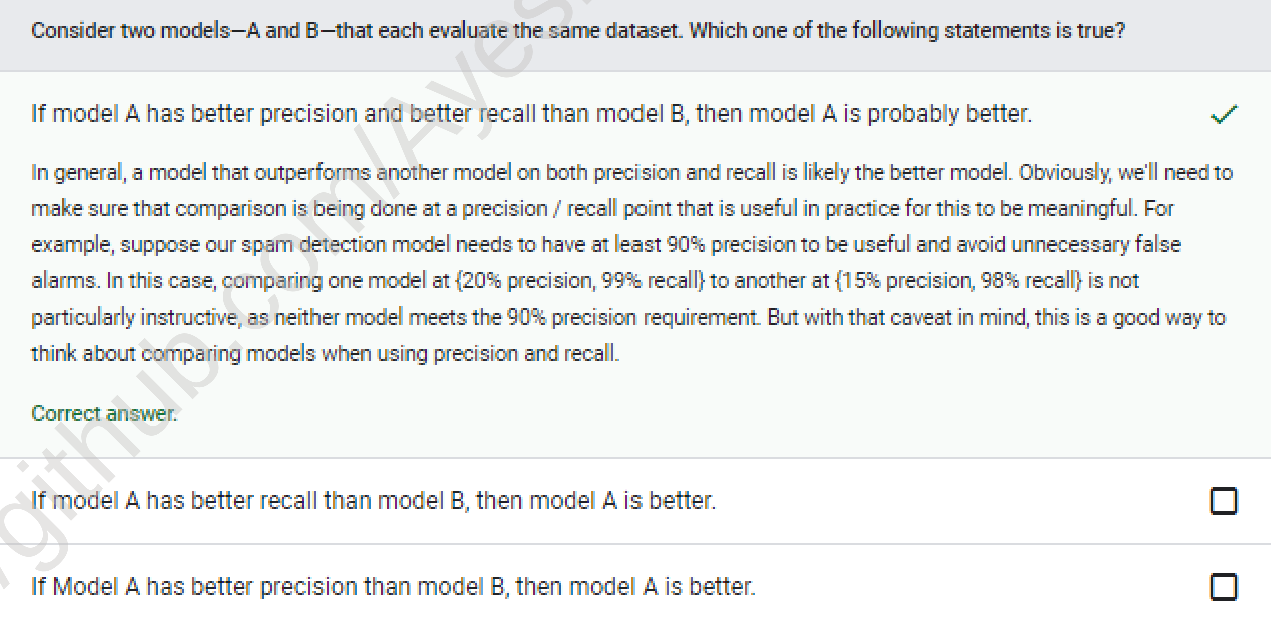
Always increase. ☐

Always stay constant. ☐

Always decrease or stay the same. ✓

Raising our classification threshold will cause the number of true positives to decrease or stay the same and will cause the number of false negatives to increase or stay the same. Thus, recall will either stay constant or decrease.

Correct answer.



Consider two models—A and B—that each evaluate the same dataset. Which one of the following statements is true?

If model A has better precision and better recall than model B, then model A is probably better. ✓

In general, a model that outperforms another model on both precision and recall is likely the better model. Obviously, we'll need to make sure that comparison is being done at a precision / recall point that is useful in practice for this to be meaningful. For example, suppose our spam detection model needs to have at least 90% precision to be useful and avoid unnecessary false alarms. In this case, comparing one model at {20% precision, 99% recall} to another at {15% precision, 98% recall} is not particularly instructive, as neither model meets the 90% precision requirement. But with that caveat in mind, this is a good way to think about comparing models when using precision and recall.

Correct answer.

If model A has better recall than model B, then model A is better. ☐

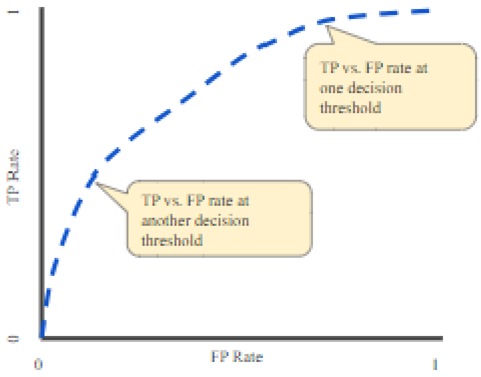
If Model A has better precision than model B, then model A is better. ☐

ROC curves

An ROC curve (Receiver Operating Characteristic curve) is a graph that illustrates the performance of a classification model at various classification thresholds.

- It plots True Positive Rate (sensitivity/recall) against False Positive Rate.

- True Positive Rate (TPR) is the ratio of correctly predicted positive observations to the actual positive observations.
- False Positive Rate (FPR) is the ratio of incorrectly predicted negative observations to the actual negative observations.
- The ROC curve shows how the TPR and FPR change as the classification threshold varies.



- AUC (Area under the ROC Curve) measures the entire area beneath the ROC curve from (0,0) to (1,1) and provides an aggregate performance metric across all possible classification thresholds.

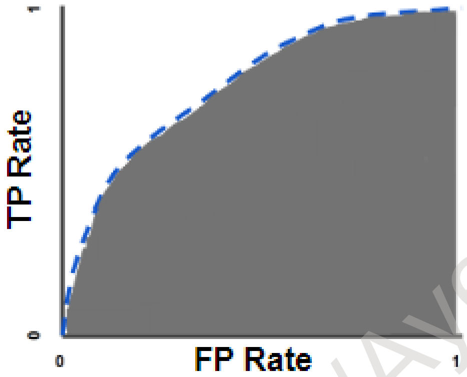


Figure 5. AUC (Area under the ROC Curve).

AUC ranges from 0 to 1, where 0 represents a model that is 100% wrong and 1 represents a model that is 100% correct.

- AUC is desirable because
  - It is scale-invariant: It measures how well predictions are ranked, rather than their absolute values.
  - Classification-threshold-invariant: It measures the quality of the model's predictions irrespective of what classification threshold is chosen.
- However, it's important to note that scale invariance and classification-threshold invariance may not always be desirable depending on the specific use case and requirements.

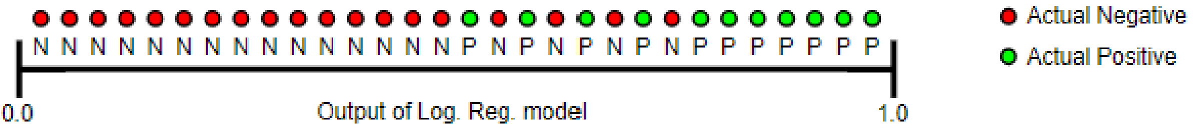
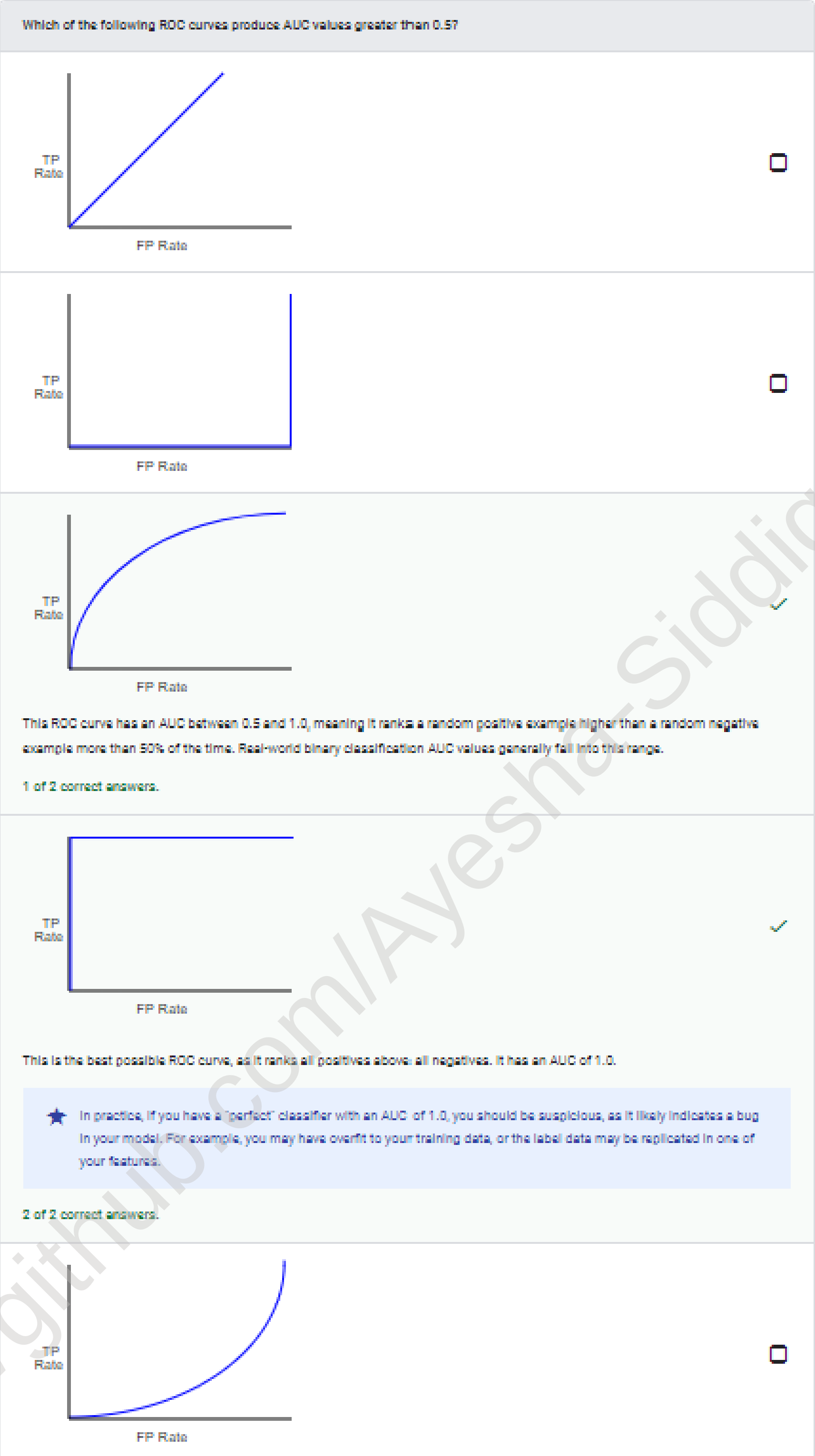


Figure 6. Predictions ranked in ascending order of logistic regression score.



Which of the following ROC curves produce AUC values greater than 0.5?

□

□

✓

This ROC curve has an AUC between 0.5 and 1.0, meaning it ranks a random positive example higher than a random negative example more than 50% of the time. Real-world binary classification AUC values generally fall into this range.

1 of 2 correct answers.

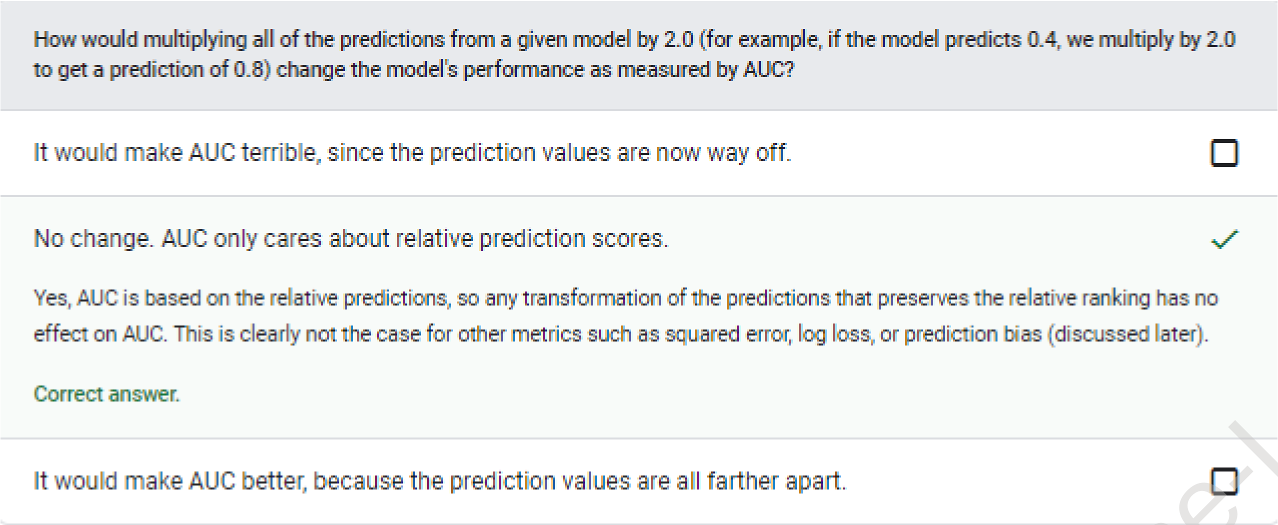
✓

This is the best possible ROC curve, as it ranks all positives above all negatives. It has an AUC of 1.0.

★ In practice, if you have a "perfect" classifier with an AUC of 1.0, you should be suspicious, as it likely indicates a bug in your model. For example, you may have overfit to your training data, or the label data may be replicated in one of your features.

2 of 2 correct answers.

□



How would multiplying all of the predictions from a given model by 2.0 (for example, if the model predicts 0.4, we multiply by 2.0 to get a prediction of 0.8) change the model's performance as measured by AUC?

It would make AUC terrible, since the prediction values are now way off. □

No change. AUC only cares about relative prediction scores. ✓

Yes, AUC is based on the relative predictions, so any transformation of the predictions that preserves the relative ranking has no effect on AUC. This is clearly not the case for other metrics such as squared error, log loss, or prediction bias (discussed later).

Correct answer.

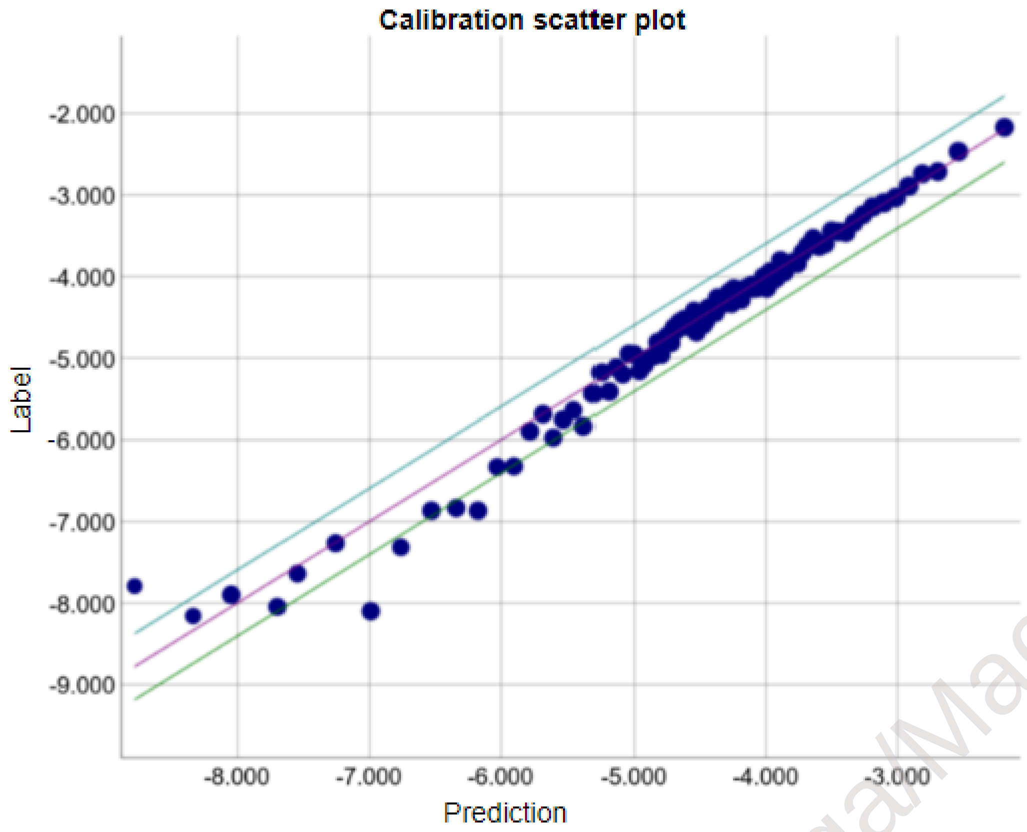
It would make AUC better, because the prediction values are all farther apart. □

*Prediction Bias

- Logistic regression predictions should be unbiased, meaning the average of predictions should be approximately equal to the average of observations.
- Prediction bias measures how far apart the average of predictions and the average of observations are.

$$\text{prediction bias} = \text{average of predictions} - \text{average of labels in data set}$$

- A significant nonzero prediction bias indicates a potential issue in the model.
- Possible causes of prediction bias include incomplete feature set, noisy data, buggy pipeline, biased training sample, and overly strong regularization.
- It is not recommended to correct prediction bias by adding a calibration layer, as it addresses the symptom rather than the underlying cause and can lead to maintenance challenges.
- A low prediction bias doesn't necessarily indicate a good model, as a poor model can also have low bias (e.g., a model that predicts the mean value for all examples).
- When examining prediction bias for logistic regression, it is important to group examples together in buckets for accurate comparison of predicted and observed values.
- Buckets can be formed by linearly breaking up the target predictions or forming quantiles.
- A calibration plot can be used to visualize prediction bias, with the x-axis representing the average predicted values and the y-axis representing the actual average values in the dataset for each bucket.
- Poor predictions in certain parts of the model may be due to inadequate representation in the training set, noisier subsets of the data, or excessive regularization.


Calibration scatter plot
Label
-2.000
-3.000
-4.000
-5.000
-6.000
-7.000
-8.000
-9.000
Prediction
-8.000
-7.000
-6.000
-5.000
-4.000
-3.000