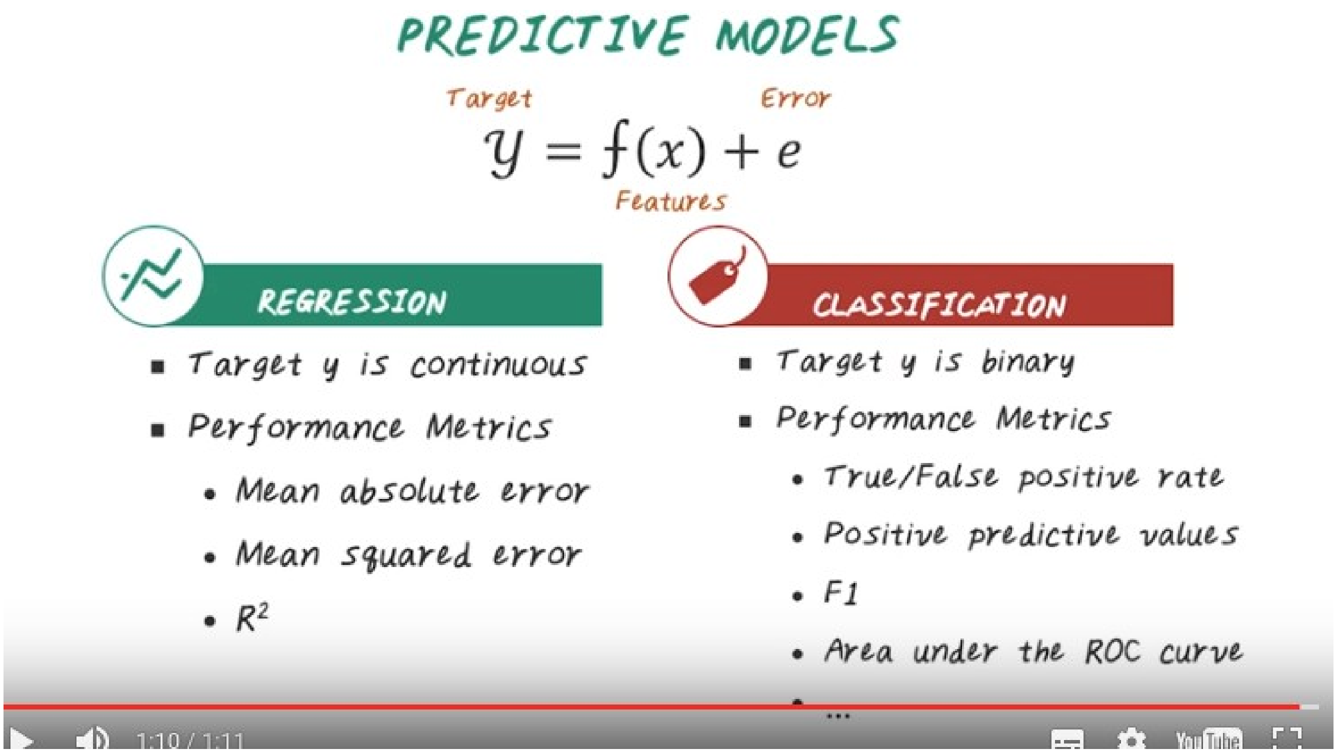
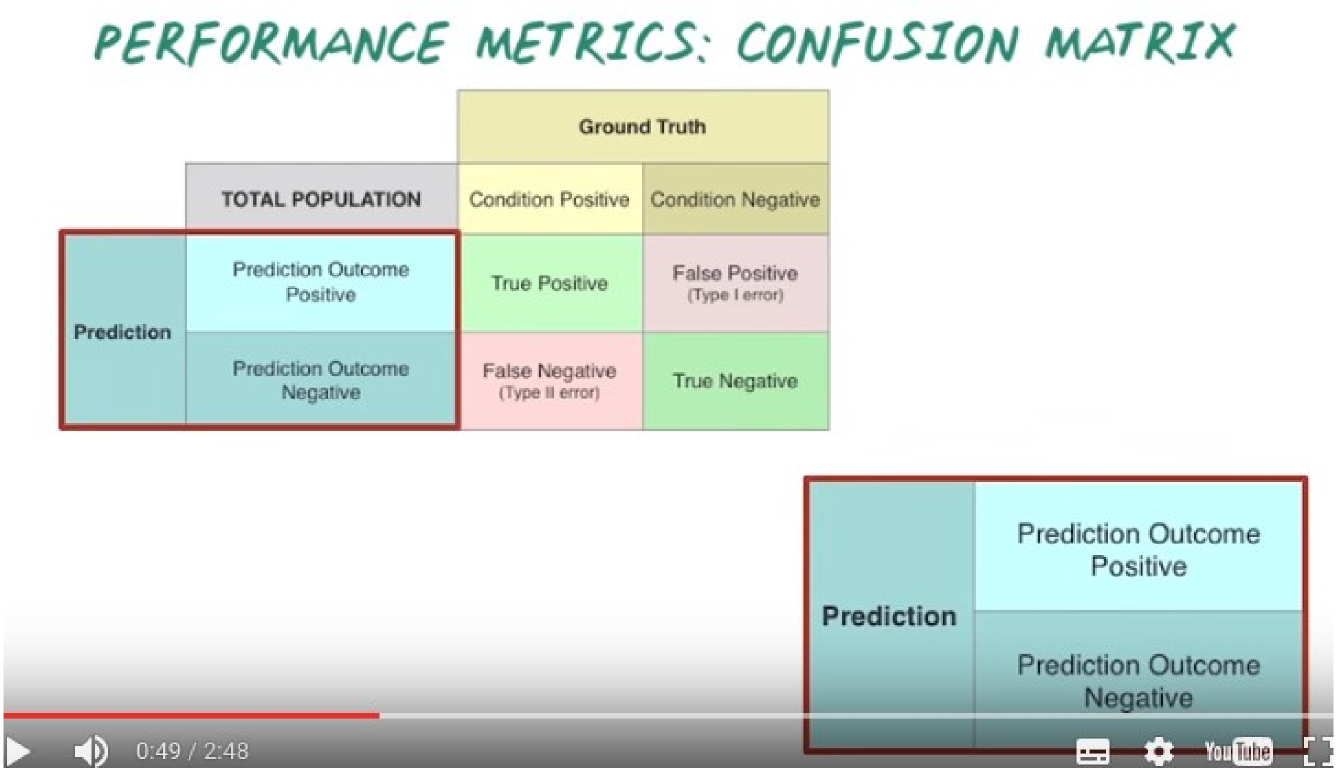
1. Last time, we talked about the overall process of predicted modeling. Now, we'll talk about how to evaluate the performance of predicted models (比如一個 decision tree 就是一個 model). When we are doing predicted modeling, one of our fundamental concerns is the quality of the models we developed and so we need metrics to evaluate the quality of the models. And there are many evaluation metrics, such as accuracy, sensitivity and specificity, which we'll talk about in this lesson. A big part of working with big data is developing multiple models and comparing them against one another. An evaluation metrics are the language we use to make these comparisons.

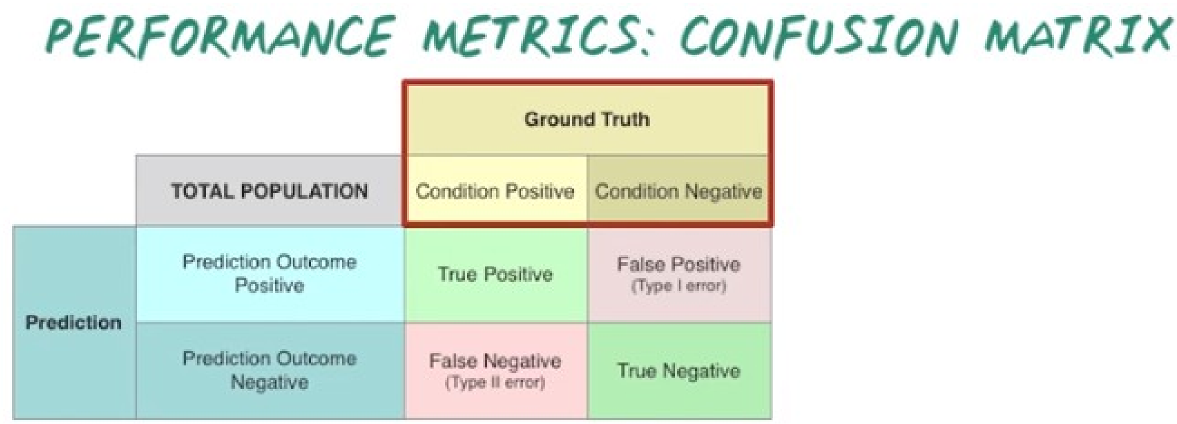


2. In the last lesson we talked about predictive modeling pipeline. It involved multiple computational steps. In this lesson, we'll focus on performance evaluation, and go into details, what are those different performance evaluation metrics. Predictive model is a function that map feature x to predict output target y. The most common predictive models are either classification problems or regression problem. For classifications the target variable y is either binary or categorical. There are many performance metrics associated with classification problems such as, true positive rate, false positive rate, positive predictive values, F1 score, area under the ROC curve, and many, many more. On the other hand if the target variable Y is continuous we also have a different performance metrics. Such as mean absolute error, mean squared error, and R squared. In this lesson, we'll go through all those different performance metrics and explain the definitions and how they're related to each other.



Condition Positive: 實際為 Positive 的,
Condition Negative: 實際為 Negative 的,
Prediction Outcome Positive: 預測為 Positive 的,
Prediction Outcome Negative: 實預測為 Negative 的.

3. Now let's start with performance metrics for a classification problem. In this lesson, we focus on describing metrics for evaluating binary classification problems, for example predicting whether patients would develop heart failure or not. These metrics can be generalized to the setting with multiple classes as well. But in this lesson, we'll focus on the binary setting. There are many evaluation metrics for classification model. Next we'll illustrate how they're all connected. To depending on the prediction and the Ground Truth value, we have four different possible outcomes. True Positive, False Positive, False Negative and True Negative. Next, we explain their definition and relationship to each other. So the output of a binary predictive model, can be either Positive or Negative. So here you can see a zoom in version of this. To differentiate with the Ground Truth value, we call this Prediction Outcome Positive, and Prediction Outcome Negative.



		Ground Truth	
	TOTAL POPULATION	Condition Positive	Condition Negative
Prediction	Prediction Outcome Positive	True Positive	False Positive (Type I error)
	Prediction Outcome Negative	False Negative (Type II error)	True Negative

Similarly, we have the Ground Truth value can be either Positive and Negative. Again, to differentiate with the prediction value, we call this Condition Positive and Condition Negative. If the Prediction Outcome is positive, and the Ground Truth's value is also positive, then we call this, True Positive. However if the Prediction Outcome is positive, but the Ground Truth condition is negative, then we call this False Positive, or Type I error. Conversely, if the Prediction Outcome is negative, and the Ground Truth Condition is positive, we call this False Negative, or Type II error. Finally, a True Negative occur, when Prediction Outcome and Ground Truth Condition are both Negative. In fact, this 2x2 matrix is called a Confusion Matrix, or a Contingency table. Now, we understand all the definitions of this basic matrix. Let's look at the relationship among this matrix. First, each row and column of this matrix sum up to the marginal. For example, True Positive plus False Positive equals Prediction Outcome Positive. And the True Positive plus the False Negative equals the Condition Positive. And the Total Population equals the Prediction Outcome Positive plus Prediction Outcome Negative. Or Condition Positive plus Condition Negative.

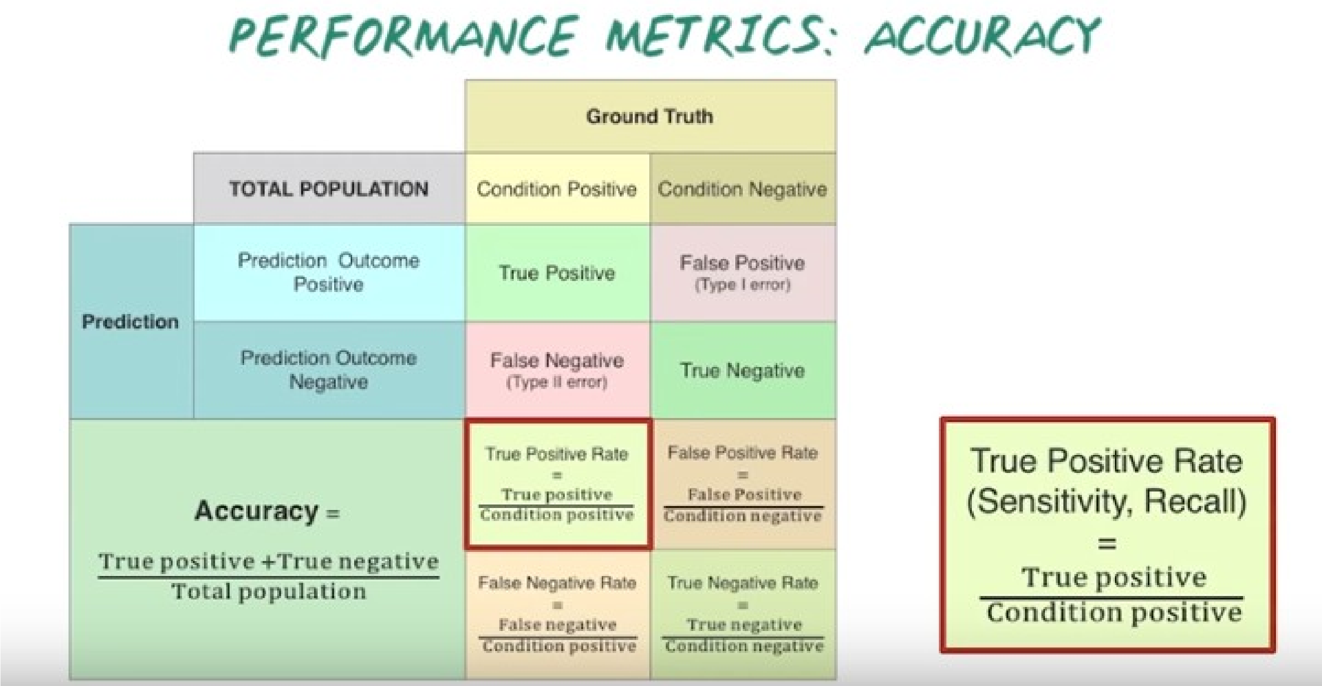
4. Now let's do a quiz to better understand the relationship of all this metric in this confusion matrix. Go ahead, fill in the true positive, true negative, condition positive, prediction outcome negative and the total population.

CONFUSION MATRIX QUIZ

		Ground Truth	
	TOTAL POPULATION 1000	Condition Positive 65	Condition Negative 935
Prediction	Prediction Outcome Positive 155	True Positive 55	False Positive 100
	Prediction Outcome Negative 845	False Negative 10	True Negative 835

Please fill in the missing numbers.

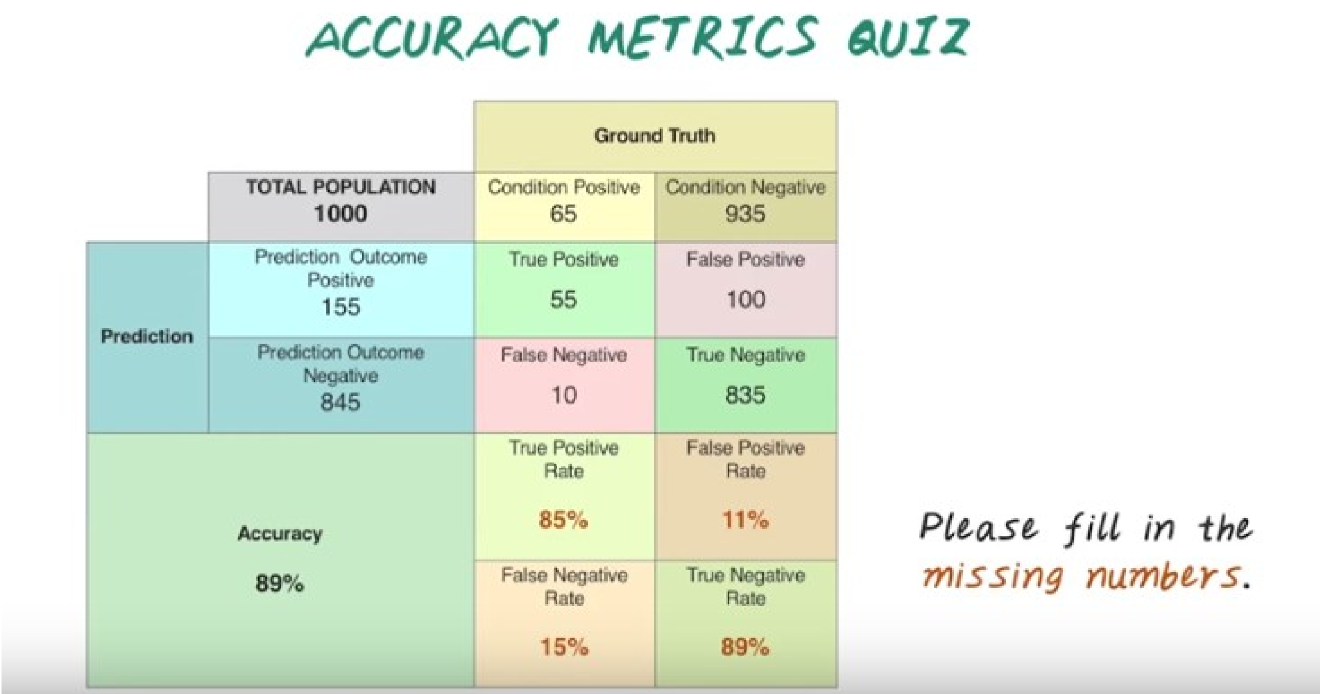
5. I hope this quiz is not too confusing for you. Here are the answers. The key insight for solving this quiz is leverage. The row sum and the column sums equals to the marginal. For example, true positive equals prediction outcome positive subtract false positive. For example, 155 minus 100 equal to 55. And the true negative equals the condition negative, subtract the false positive. 935 minus 100 equals to 835. Now you have all those four numbers and the condition positive can be calculated by taking the sum between true positive and false negative. 65 equals to 55 plus 10. And similarly prediction outcome negative equals the false negative plus true negative. 10 plus 835 equal to 845. And finally, the total population can be calculated easier by sum up, prediction outcome positive, and prediction outcome negative. Or you can sum up the condition positive with condition negative, so the total population equal to 1000.



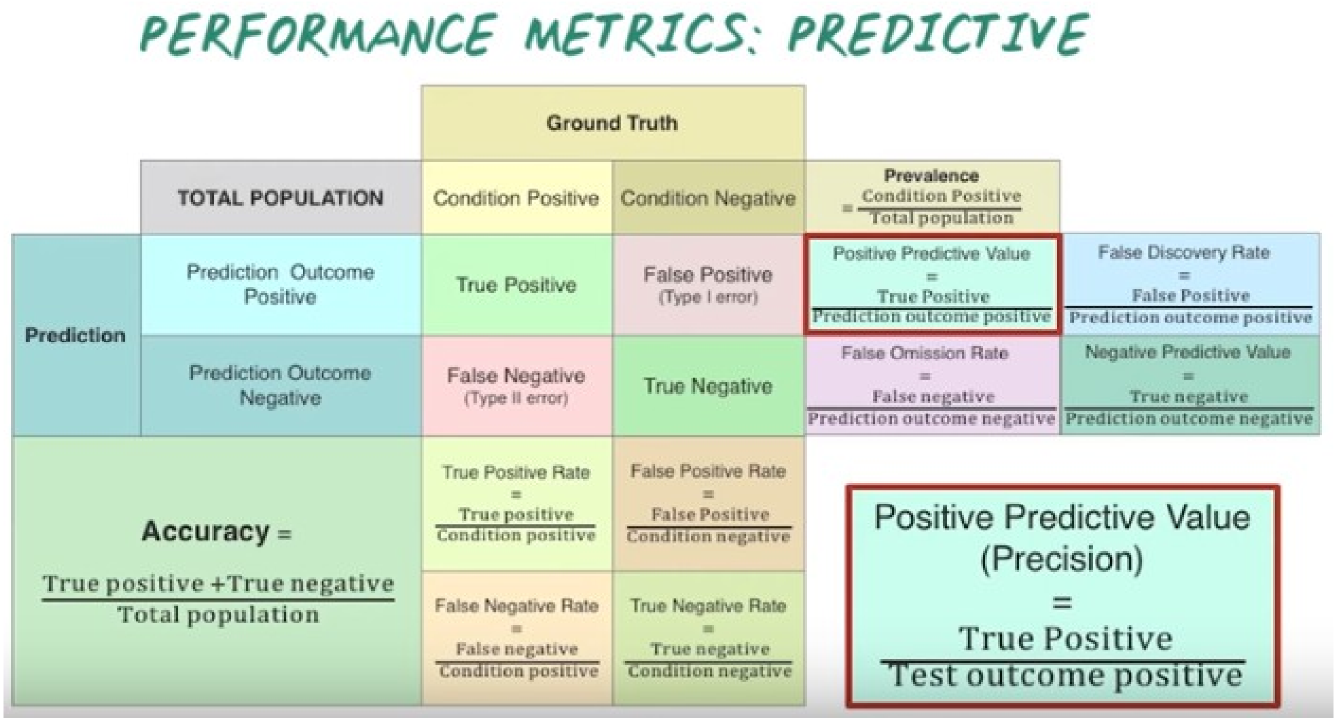
6. Knowing this two by two confusion matrix is just the beginning. There are many different metrics can be derived from those basic metrics. In particular, many metrics can be derived by taking ratios of different numbers from this basic metrics. First, let's introduce a set of metrics that are derived by normalizing some terms with the ground truth value, either the total population or the condition positive or the condition negative. So all these metrics are defined by some connative divided by those normalization terms. For example, accuracy is the ratio from the sum of true positive and true negative divided by the total population. Accuracy is the most basic metric that many people have intuitive understanding, but accuracy is not always the best measure when the class label are very imbalanced. If only 1% of the total population have heart failure, a trival models can be achieved 99% accuracy by simply predicting everyone would not have heart failure. So it is important to use the appropriate metric for measuring the performance of a predictive model. Another important metric is true positive rate or sensitivity or recall. So they have different names, because the same metric has been developed by different communities, and true positive rate equals the true positive divided by the condition positive in the ground truth. To explain the intuition, let's assume positive means heart failure and negative means without heart failure. The intuition of true positive rate is to measure among all people with heart failure, what percentage is correctly identified by the predictive model? A very related metric can be derived called false negative rate, which equals 1 minus true positive rate. The intuition of false negative rate is among all heart failure patients, what percentage of those patients is missed by the predicting model? False positive rate is the ratio between false positive and the condition negative. The intuition of false positive rate is to measure among all people without heart failure, what percentage of them is incorrectly predicted by the model as to have heart failure? Finally, true negative rate or specificity is the number of true negative divided by the ground truth condition negative, which also equals 1 minus false positive rate. The intuition behind true negative rate is to find among all

people without heart failure, what percentage of them is correctly classified as non-heart failure patients? Note that accuracy measures the combination performance of true positive and true negative, while true positive rate measures the performance only about true positive, and true negative rate measures the performance only about true negative. It is often possible to obtain high performance on a single metrics such as accuracy, while it's difficult to obtain high performance on all metrics. That's why we often test a predictive model against multiple metrics. First of all, different metrics are designed for different situations. It is important to choose the proper metric for your prediction study.

7. So now let's do another quiz on those new metric we just learned. Fill in the box based on the number already provided here. Let's calculate the true positive rate, false positive rate, false negative rate, and true negative rate. The detailed formula of all this metrics are provided in the instructor notes.

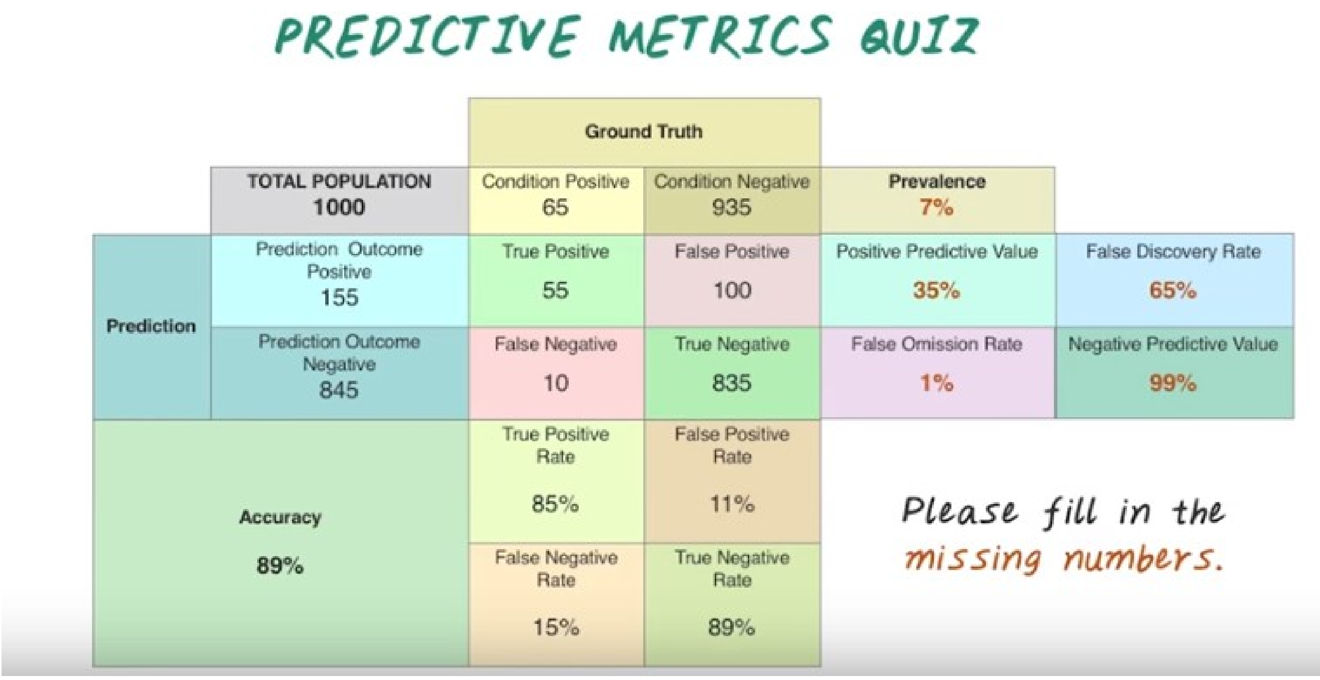


8. I hope your answers are accurate. [NOISE] The true positive rate can be calculated by taking the true positive divided by the condition positive (55/65) and that give us 85%, and the true negative rate can be calculated by taking the true negative divided by the condition negative. That give us 89%. Next, for false positive rate, we can calculate that by either taking the false positive divided by the condition negative, or just simply taking one minus the true negative rate, which is 11%. Finally the false negative rate can be calculated by taking the false negative divided by the condition positive, or simply one minus the true positive rate. That'd give us 15%.

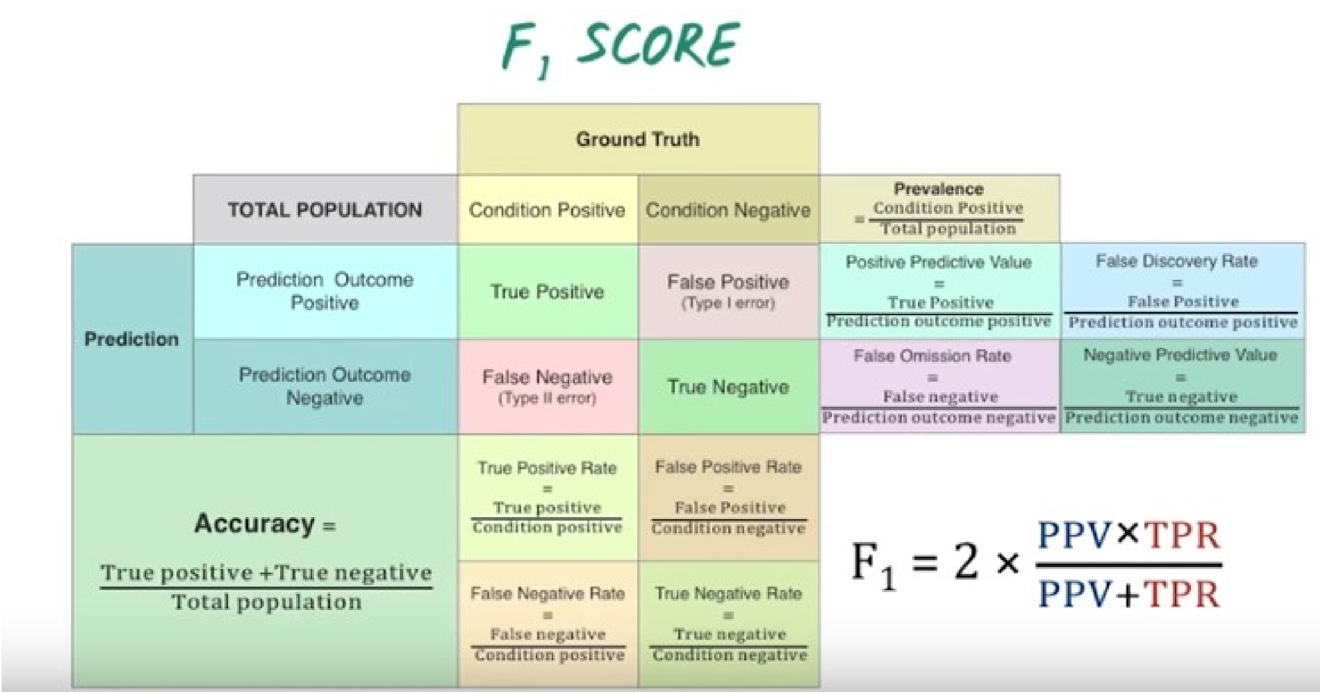


9. So far we have learned all the accuracy metrics which are normalization by the ground truth values. Next, we'll learn another set of metrics that are defined by some commodity divided by the prediction outcomes. Either the prediction outcome positive or prediction outcome negative. First prevalence. Prevalence is the ratio between condition positive and total population. And prevalence measures how likely the disease occurs in the total population. For example, for different disease condition, the prevalence can be very different. For heart failure. Among older population, the prevalence might be quite high such as 20%. For a rare type of cancer, the prevalence of that disease can be quite low maybe 0.001%. Now let's look at positive predictive value or precision. Positive predictor value is the true positive divided by predictive outcome positive. Positive predictive value measures among all patients that are predicted to have heart failure. What percentage of them would actually have heart failure? A relative metrics is called false discovery rate. Which measures the number of False Positives divided by the Prediction Outcome Positives, or 1 minus positive predicted value. The intuition behind False Discovery Rate is that, among all patients that are predicted to have heart failure, what percentage of those predictions is incorrect? Similarly we can have negative predictive value which defines as the true negative divide it by prediction outcome negative. The intuition behind negative predictive value is among all patient who are predicted to not have heart failure by the model. What percentage of the prediction is correct? And finally, we can also define false omission rate, which is the false negative divided by the prediction outcome negative. The intuition behind false omission rate is among all people who are predictive not to have heart failure by the model, what percentage of them will actually develop heart failure? That means the prediction is inaccurate there.

10. Now let's do another quiz. Please fill in the numbers for Prevalence, Positive Predictive Value, and False Discovery Rate.



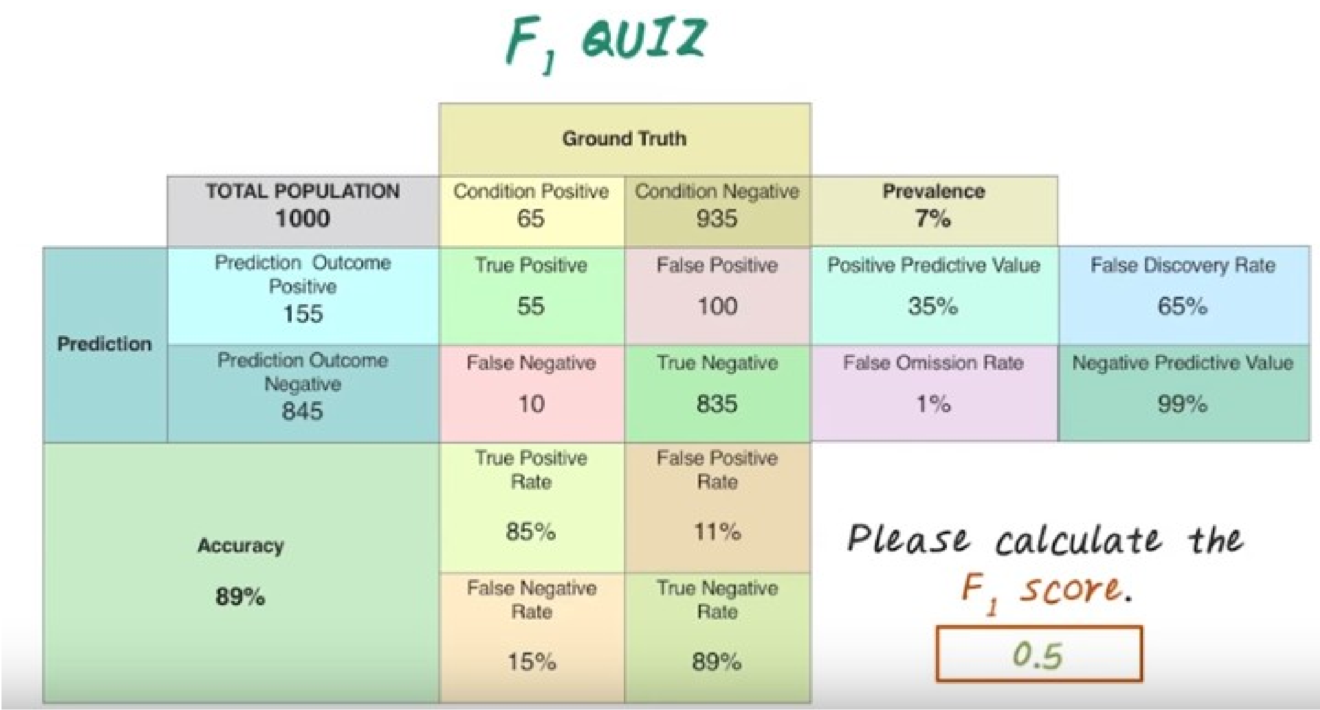
11. so here are the answers. Prevelance can be computed by taking the condition positive divided by the total population. So, it's close to seven percent, and the positive predictive value equals the true positive, divided by the prediction outcome positive. We have 35% here. And the false discovery rate can be calculated by taking false positive divided by prediction outcome positive, or simply, one minus positive predictive value, which is 65%. I hope you predicted correctly, all those numbers. [SOUND]



12. Another popular metric is called F1 score. It combines the true positive rate and the positive predictive value. And the F1 score is two times positive predict value times true positive rate divided by the sum of those two measures. In fact, this fancy formulation is really a harmonic mean of those

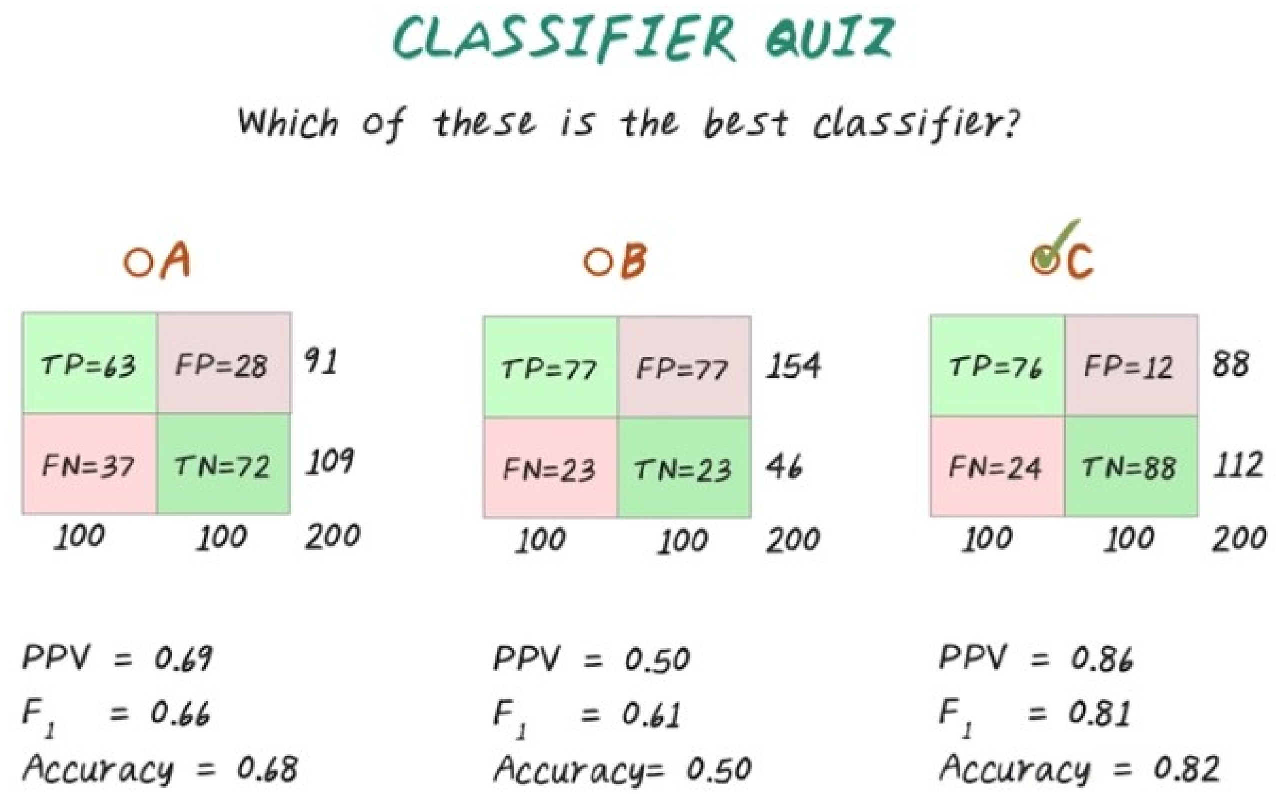
two measures, positive predictive value and the true positive rate.

13. Let's do another quiz about F1 score. Based on the information in this table, please calculate F1 score.



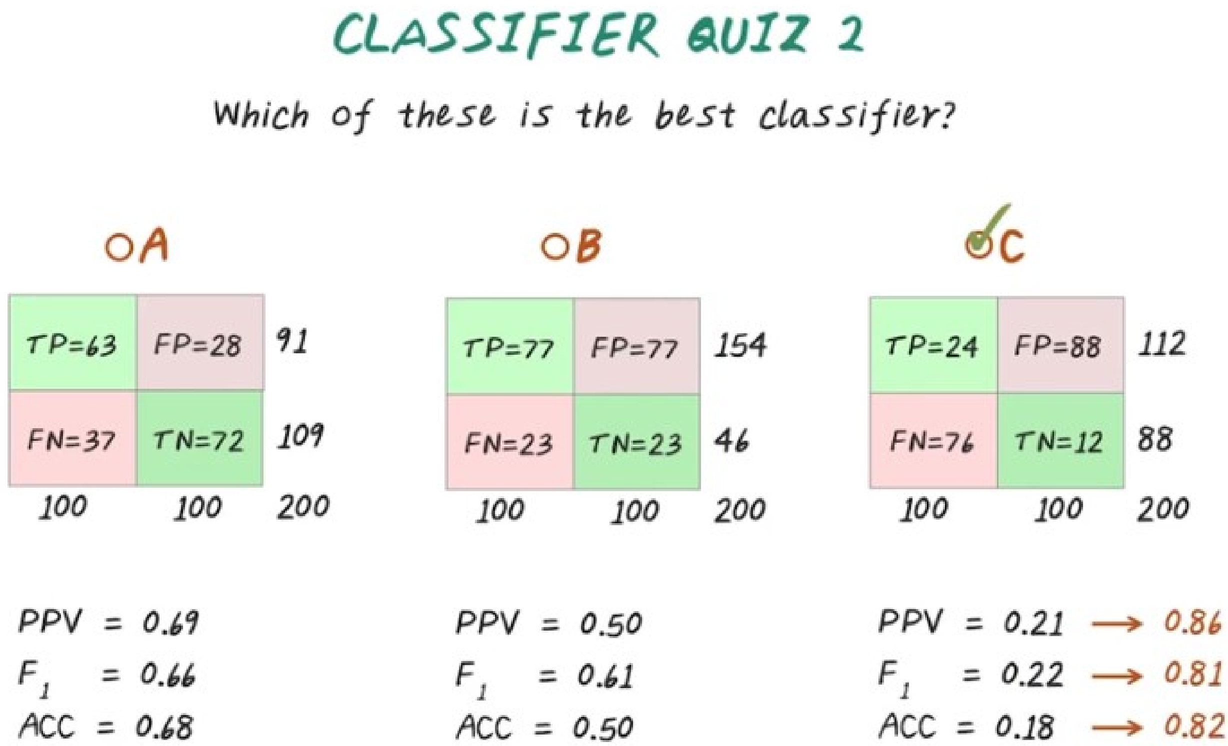
14. And the answer is close to 0.5. You can calculate the F1 score by taking the true positive rate times the positive predictive value, then divide it by the sum of those two measures. And scale that by 2, which will give you the number 0.5.

15. Now we have learned many different classification metrics. Let's use that to decide which one of this is the best classifier. Here are the different confusion matrix of these three classifiers and the corresponding performance metrics. This is the true positive rate, this is the false positive rate and this false negative rate and this is true negative rate, and the numbers on the side are the marginal. This are the ground truth's condition positive and the ground truth's condition negative. And this is predicted outcome positive and this predictive outcome negative. And this is the total population.

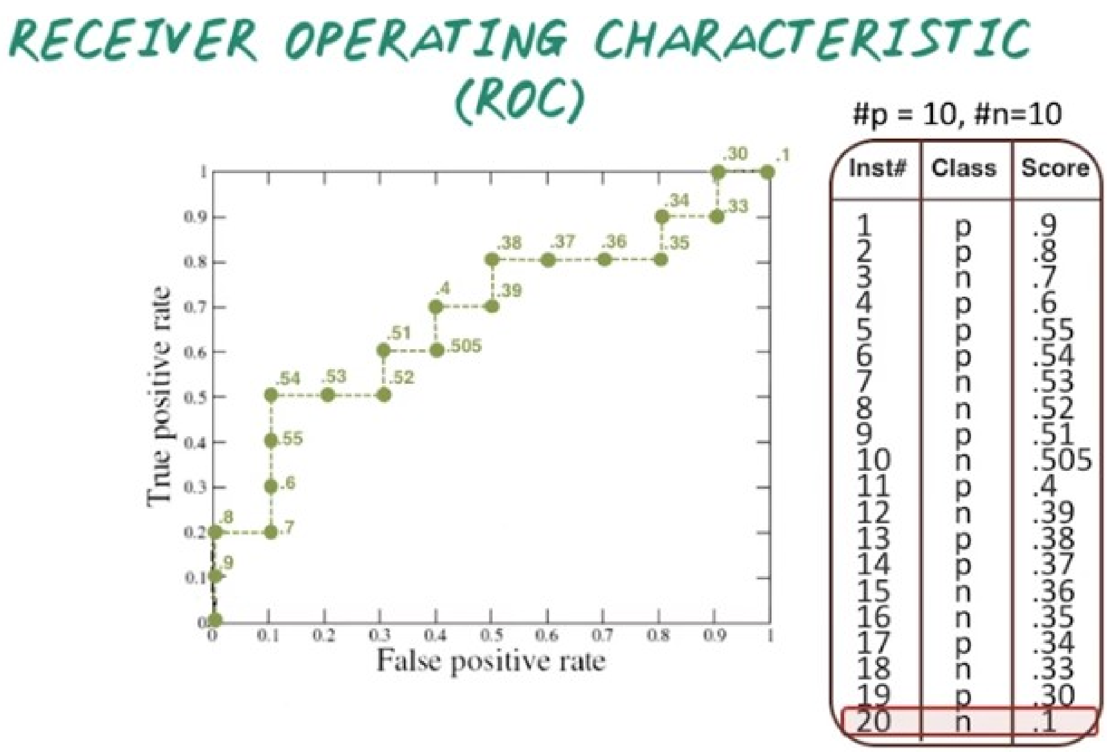


16. And the answer is C because it has higher performance matrix in all the three measures, positive predict values, F1 score, and accuracy.

17. Now we change the performance measures on classifier C. Can you tell me which one of this is the best classifier?



18. The answer is still C. This might be surprising to many of you. Although A seems to have a higher performance measure than B and C, C can be easily improved by reversing the prediction. We can change the positive to negative, and negative to positive. This way the performance measure will become highest, the same as before. So in this manner the C classifier will still perform the best.

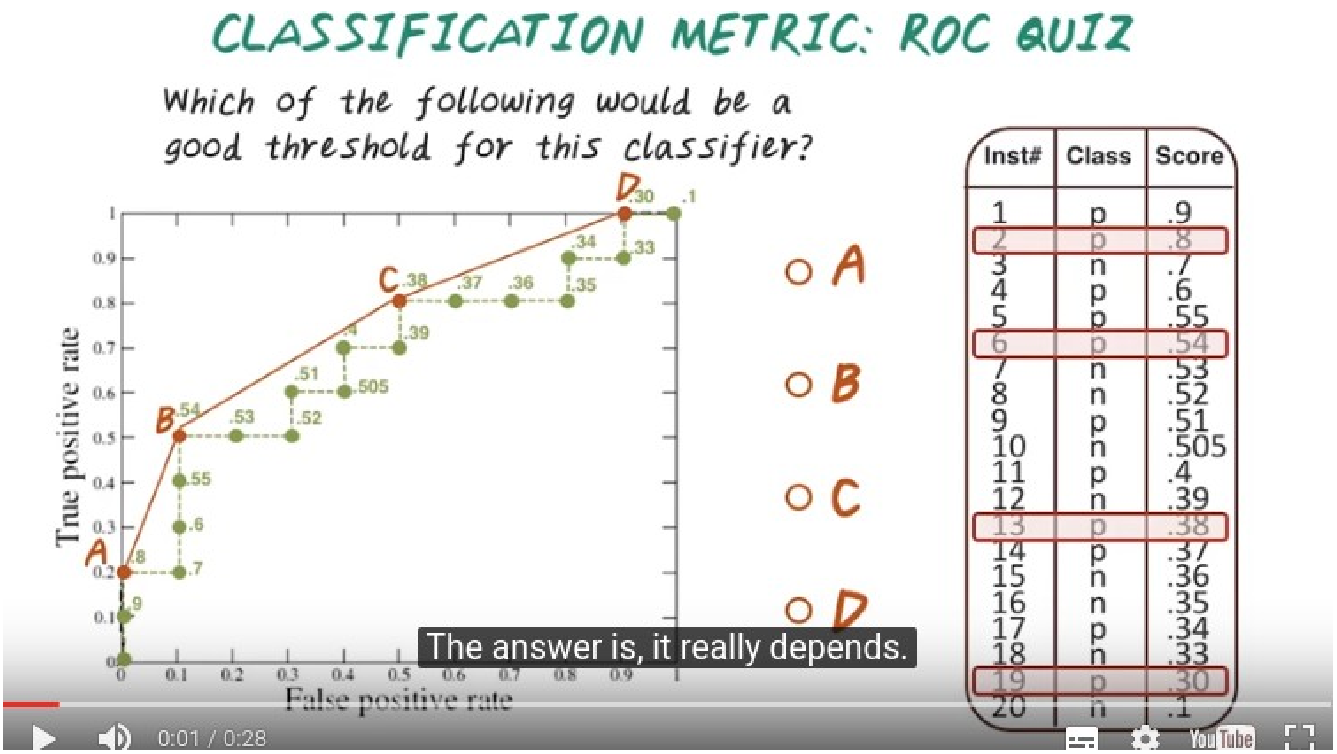


Class 那一列是該人真實的情況, 即實際上為 positive 還是 negative.

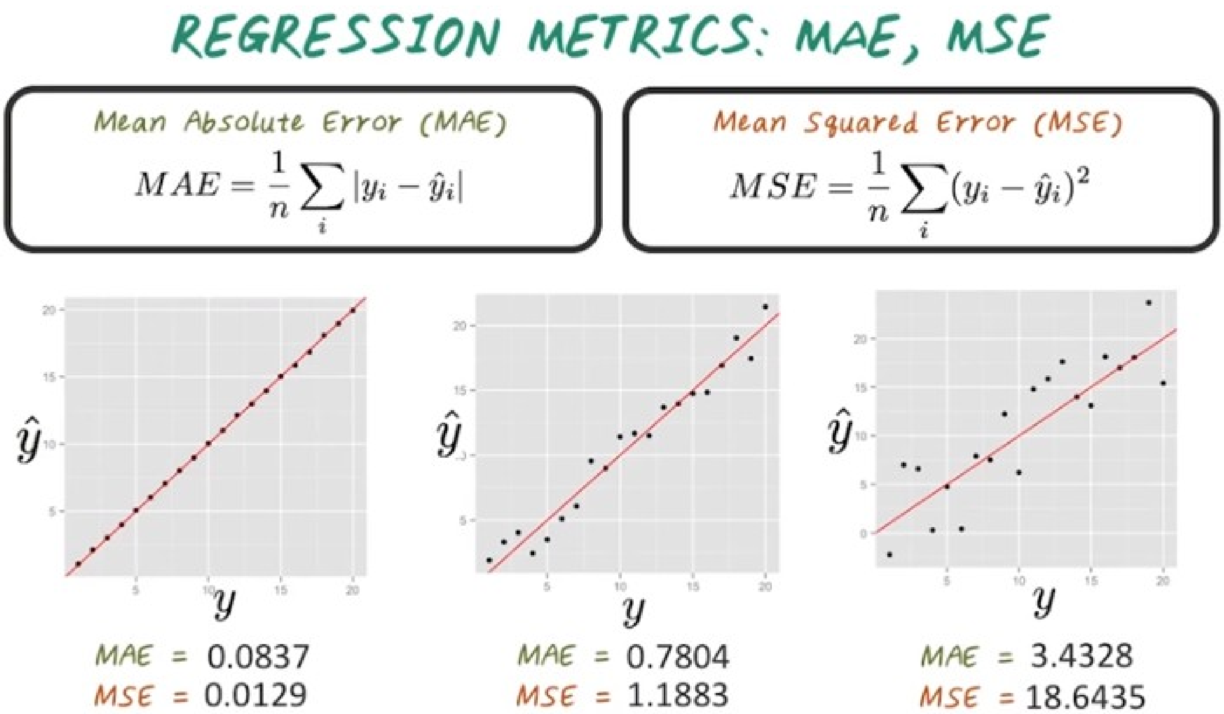
Score 那一列是 prediction score, 越接近 1, 我們就越預計它是 positive.

19. In general, predictive models output continuous prediction score. For example, in binary classification, it will be between 0 and 1. The value closer to 1 means prediction outcome positive, and the value closer to 0 means prediction outcome negative. In this case, we need a threshold as the prediction boundary. This threshold has a significant impact on all the performance metric we have discussed so far. And receiver operating characteristic, or ROC curve, provide a way to compare different classifiers as a prediction boundary is varied, and this curve is created by plotting the true positive rate against the false positive rate at various threshold values. So such a curve can be generated by going through all the data points in the descending order of their prediction score and using those prediction scores as threshold values. For example, we have 20 patients. Ten of them have positive outcomes indicated by letter p, and ten of them have a negative outcome indicated by letter n. We sort them based on the prediction score. Then we start using those prediction score as threshold points. We start from the point where the true positive rate and the false positive rate are both 0. Next, we pick the threshold value to be 0.9, which means the prediction score greater than or equal to 0.9 will be predicted as positive, and otherwise predicted as negative. In this case, the true positive rate will be 0.1, because only one out of the ten patients with positive outcome are predicted correctly (因為 10 個 p 中，我們目前只看了一個 p). The false positive rate in this case is 0, because no one has been mis-classified. Next, we change the threshold to 0.8. In this case, two positive patients are classified correctly out of ten. Therefore the true positive rate improved to 0.2. Still, the false positive rate remained 0 because no mis-classification has happened yet. Now let's change the threshold value to 0.7. Now we actually mis-classified this instance to be positive, but the actual outcome is negative. Therefore, the false positive rate become 0.1, since one out of ten negative examples are mis-classified as positive. The true positive rate remains the same. We can go through all those data points by changing the special values to complete this ROC curve. The ROC curve illustrates overall performance of a classifier when we're varying the threshold value. One important metric is the area under this ROC curve, or AUC. Since AUC doesn't depend on the choice of the threshold, it becomes the most popular performance metric for classification problems.

20. Now we understand ROC curve, and the concept of AUC. We can use that to choose the best classifier. However, to utilize that classifier to make a prediction, we still have to choose a threshold. Now which of the following would be a good threshold for this classifier? Is this A, 0.8, or B, 0.54, or C, 0.38, or D, 0.3?

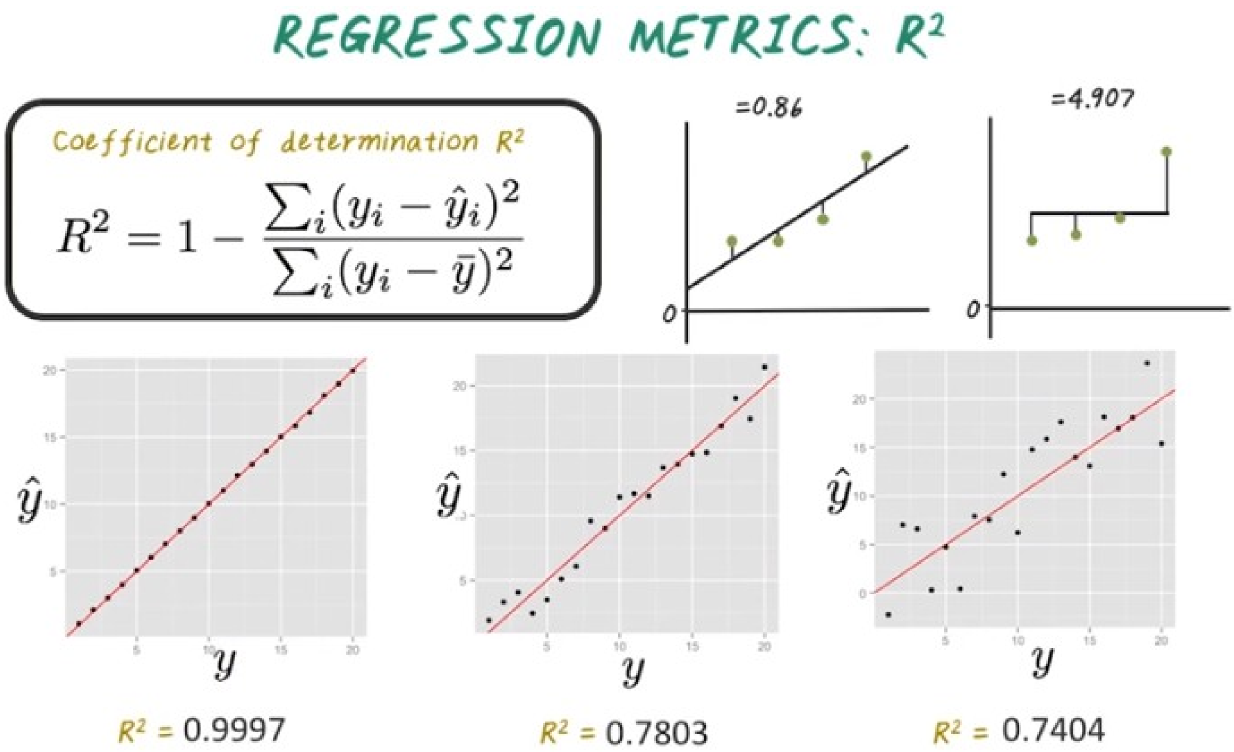


21. The answer is, it really depends. If you're really curious about false positive rate to be low, then A is a good choice. If you really want to have a high true positive rate, then C and D might be good choices. If you care true positive rate and the false positive rate equally, then B will be the good choice. So, the bottom line is, the optimal classification threshold may vary depending on your preferences.



yi 是實際值, yi^是 prediction.

22. So far we introduced many classification metrics. Next, we'll present some popular regression performance metrics. The most popular regression metrics are mean absolute error, MAE, or mean squared error, MSE. The mean absolute error, MAE, measures the average of the absolute errors, that is the difference between the prediction and the ground truth value, and the mean squared error MSE, on the other hand, measures the average of the squared error. MSE is easier to work with because the derivative of the square term is linear, but MSE will greatly affected by outliers because of the square term as well. On the other hand, MAE is more robust against the outliers, but it's harder to work with because this absolute value is not differentiable. Here are some visual illustrations. The X axis is the grand truth value and the Y axis is the prediction. As the amount of noise increases from left to right, both MAE and MSE increase. You can also notice that MSE increased a lot faster because the square of error term.



23. Both mean absolute error and mean squared error, a popular metrics. But they're not bonded in a fixed range, so it's not possible to compare across data sets. Next, we'll introduce another regression metrics called R squared which has a fixed maximum score of 1. So formally R squared or coefficient of determination is one minus the ratio between MSE and variance. For example, if we have a linear regression model looking like this, the mean square error can be equal to .86, while the variance equals 4.9 and r squared for this particular example is around 0.82, which is considered to be very good. In fact, R squared of 1 indicate the regression is perfectly fits data while r squared of 0 indicate the line does not fit the data at all. It is important to notice that by this definition, it's possible to have negative values of R squared. Which means the predictive model performed worse than a simple average over the original data. Again, the same visual illustration as we increase the noise level, we can see the R squared value also decreases.