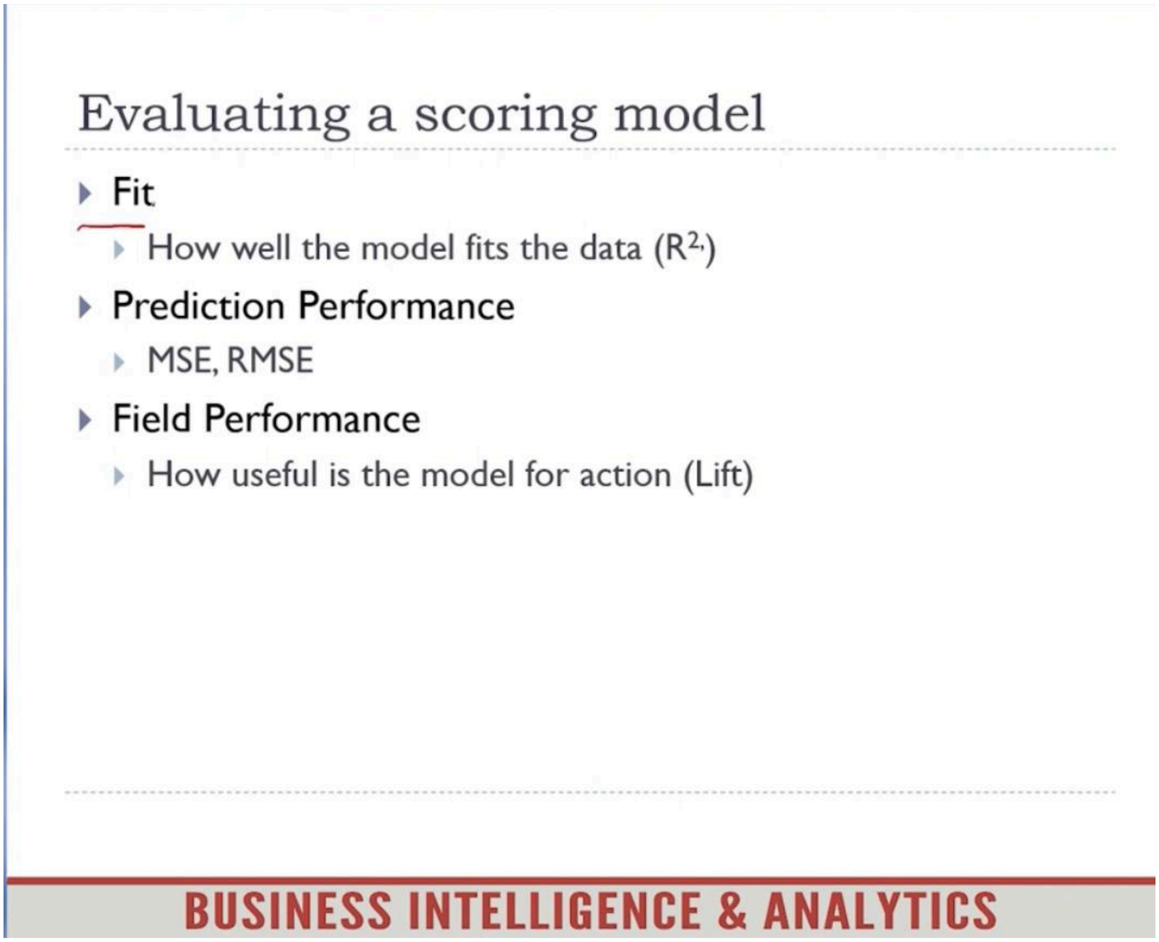
Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:06
Lecture:23

SCORING MODELS | Business Intelligence & Analytics

Now, we have to understand certain fundamental concepts related to classification, related to application of classification and also the performance of algorithms in practice in business applications. So, we have seen so far in our discussions when we discuss modeling, there is something known as fit. This is at the time of training. When you build a model, what you focus on is the fit, the goodness of fit of the model for the given data set or the pattern in the given data set and we have measures like R^2 in regression, which actually gives us a sense of how well the model fits. And model fit is very important. It is necessary, but this is not enough for a model to be applied in practice.

Then therefore, the next step is when you are actually using models for prediction in prediction scenario, for example, a classifier is built and classifier algorithm is developed or built using data, using training data. The next step is to make the model predict and we discussed this in connection with prediction performance using test data. You have different test procedures or cross validation techniques available for testing the predictive performance of models. For example, the MSE,, the RMSE and so on.

This is also important, but this, so far we are in the laboratory. When you build a model, when you actually test a model etc., we are in the laboratory. This is lab. And here, you are going to deploy the model or apply the model or use the model. For example, a classifier to determine whom a marketing firm is going to target.



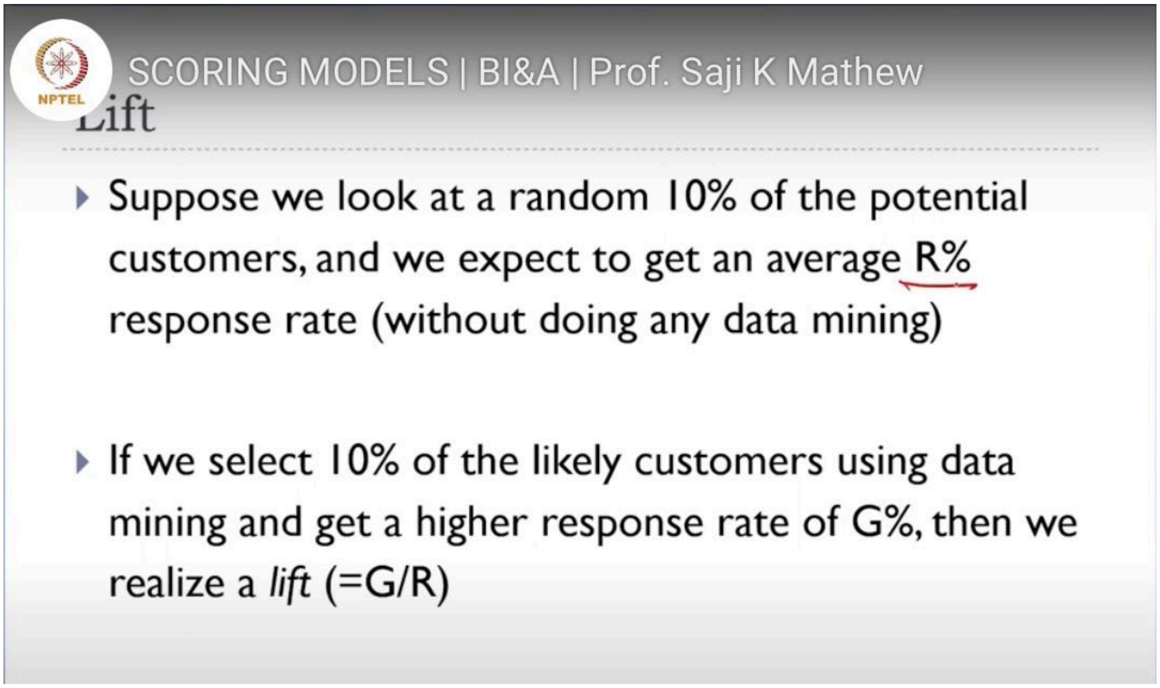
Evaluating a scoring model

- Fit
  - How well the model fits the data ($R^{2,}$)
- Prediction Performance
  - MSE, RMSE
- Field Performance
  - How useful is the model for action (Lift)

BUSINESS INTELLIGENCE & ANALYTICS

And that is a choice, that is an application. And when you deploy the model, how will the model performs in terms of business performance measures? What is important to a business is not R^2. It is not MSE or RMSE that they care about. What they care about is the response rate as we saw in the previous example. What matters to business is business value.

Response rate is a proxy for business performance or business value, you know that can be created using an algorithmic intervention. So a measure of the performance or the field performance of an algorithm is a measure called lift. And let us see what is lift in a specific context like what we are discussing now. In order to illustrate what is lift, I am taking an example. Suppose we look at a random 10 percent of the potential customers, and we expect to get an average R percent response rate without doing any data mining or analytics or applying any techniques.


SCORING MODELS | BI&A | Prof. Saji K Mathew

Lift

- Suppose we look at a random 10% of the potential customers, and we expect to get an average R% response rate (without doing any data mining)
- If we select 10% of the likely customers using data mining and get a higher response rate of G%, then we realize a *lift* (=G/R)

So what does that mean? You have a data set of n prospects, but your budget is limited to 10 percent. This is the database of say 10 million or 20 million. So 10 percent is what you want to choose. You can just pick that 10 percent randomly, just randomly. It could be the, you know, just pick based on no order, the 10 percent.

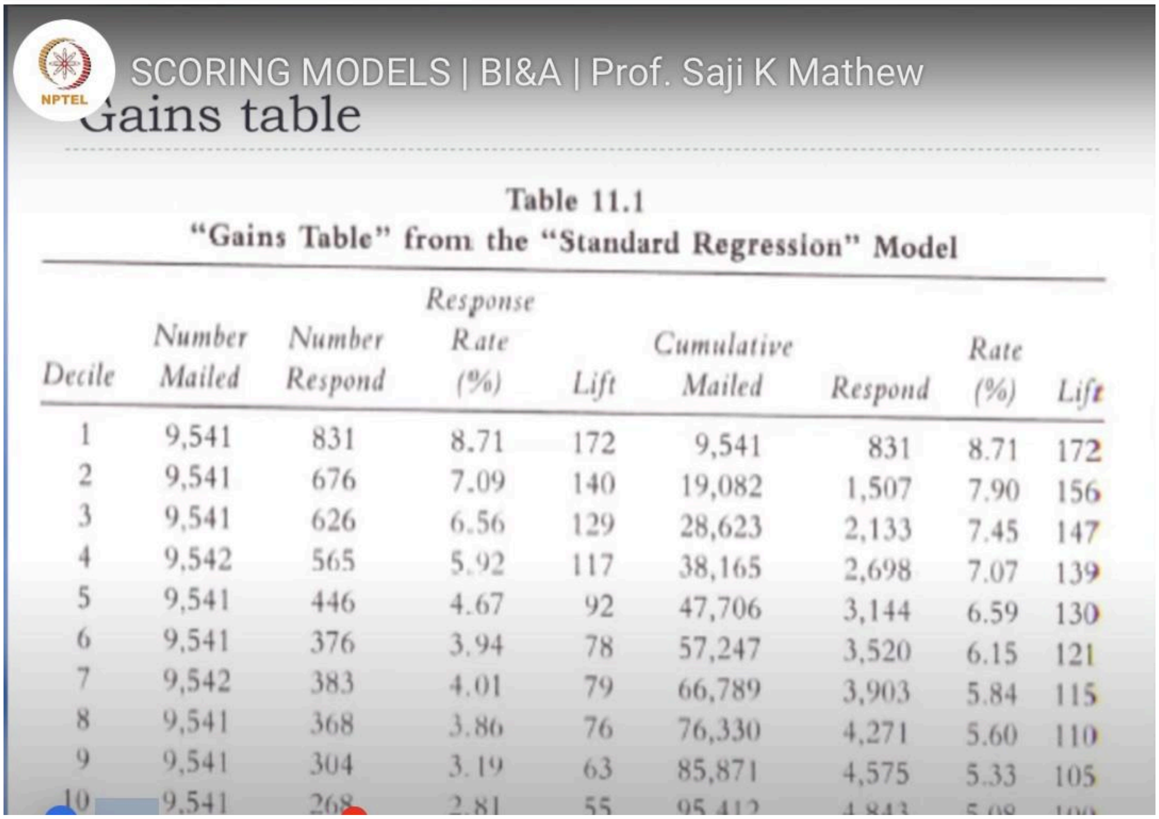
And then send your promotional package to them and suppose that response rate is R, R upon or the R is the denominator or is the reference. Now suppose we select the 10 percent instead of doing it randomly, we develop some algorithm like the bizocity score, which is a scoring model. And suppose we select the top 10 percent who are likely to respond and then send your promotional packets to those customers and the performance now or the response rate now is G. Then G/ R, G divided by R is the lift. Lift is a ratio, lift is a measure of the improvement in performance of a particular business activity or a business because of a algorithmic intervention.

And this is specific to a context. I am explaining the concept of lift specific to a targeting problem. When it comes to different problem context, lift is measured differently or the formulae could be different, but the concept remains the same. How much improvement happens because of an algorithmic intervention? That is what is lift is. And since it is a ratio, you can imagine different values for the ratio.

Suppose lift is equal to greater than 1, say 1.8, will you be happy? Yes, a business will be happy because compared to the random selection, there is a improvement which is 1.8 times. So you got response which is 1.8 times the random response rate.

And if you can make it 1.9 or 2 using better algorithms, you are gaining further. Suppose your lift is equal to 1.0 or just 1, what does that mean? You put lot of efforts, but no improvement, absolutely no improvement. The algorithmic intervention does not pay anything, but there is a cost.

So actually, the business value is negative because there is a cost involved in the development. So this is not desirable. And suppose we can imagine even worst cases where it becomes less than 1, then this is complete damage to the business. And therefore, you would have better left the business as it is, instead of applying your knowledge in analytics or data mining. So what it essentially means is that lift is a very intuitive measure which gives you an idea how the business is performing.



SCORING MODELS | BI&A | Prof. Saji K Mathew

Gains table

Table 11.1
"Gains Table" from the "Standard Regression" Model

Decile	Number Mailed	Number Respond	Response Rate (%)	Lift	Cumulative Mailed	Respond	Rate (%)	Lift
1	9,541	831	8.71	172	9,541	831	8.71	172
2	9,541	676	7.09	140	19,082	1,507	7.90	156
3	9,541	626	6.56	129	28,623	2,133	7.45	147
4	9,542	565	5.92	117	38,165	2,698	7.07	139
5	9,541	446	4.67	92	47,706	3,144	6.59	130
6	9,541	376	3.94	78	57,247	3,520	6.15	121
7	9,542	383	4.01	79	66,789	3,903	5.84	115
8	9,541	368	3.86	76	76,330	4,271	5.60	110
9	9,541	304	3.19	63	85,871	4,575	5.33	105
10	9,541	268	2.81	55	95,412	4,843	5.08	100

So going further, let me demonstrate to you how the concept of lift and business performance can be tracked when you apply algorithms. There is something called gains table, which if you do analytics intervention systematically, you can track how the algorithm has performed and then what changes needs to be done or what strategy needs to be followed in targeting etc using a gains table. So this particular table demonstrates the responses that were obtained when they sent their, when they did direct marketing based on a regression model, a standard regression model they developed for targeting customers. And based on the score obtained by each prospect, they classified prospects into deciles, 10 deciles you can see. Of course, the top decile has the highest values or highest scores based on regression and they fall into the top category and when they were targeted, you know, it is 9541.

So obviously, you can see that the total number of prospects that were targeted is 95412. So the first decile has 9541 and when they, out of them 831 responded. So the response rate is 8.71, which is 831/ 9541. You can see that the response rate is the highest in the first decile, obviously because they got the highest score based on the regression model or they were most likely to respond.

And then you can also calculate lift here. You can see the lift is calculated in percentage, of course, so it is 172. And what is lift? Lift is equal to, we know that it is G/ R. And what is G/ R? So the random response rate is the denominator and the numerator is the response rate based on algorithm.

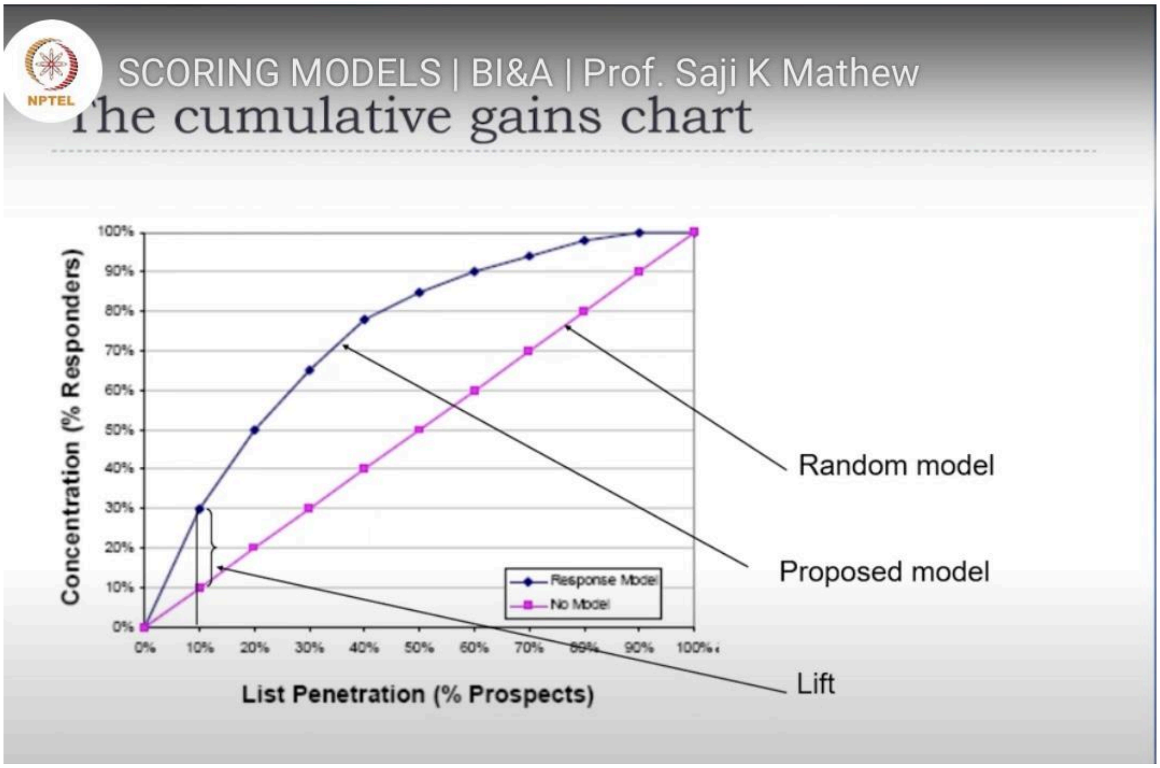
And you can see that 8.71 is G. So 8.71 is G. And what would be the random response rate? Random response rate is that if you send your material to all the people that is in your target group or in your selected group, what would be the response rate? And that is calculated here based on the cumulative value, here you can see that.

So 5.08, that is 1.72 × 100 becomes 172. So gains table is useful in evaluating the performance of different deciles. And you can also see that when you move from the fourth decile to the fifth decile, your lift is actually going below 1. And therefore, obviously you can see that, you know, one should not be targeting those deciles at all. So this table that way is very insightful as to how you can select from a population or a larger group, you can actually target a specific set of targets whom you would actually reach out to.

That is the gains table. And what is displayed here is the same concept in the form of a graph or it is visualized here. And let me explain this graph, the cumulative gains chart to you. First the axis, any graph, the first thing to do is to look at the axis, the x axis and the y axis. Without that one should not look at the shape of the graph at all, otherwise you will be fooled. There is a book, Fooled by randomness. And that is worth reading.

So you can see the x axis is the list penetration or percentage of prospects. Suppose there are 10 million prospects whom you are going to target. 10 percent shows that 10 percent of 10 million, 20 percent shows that 20 percent of 10 million, 100 percent is actually the 10 million. That is the x axis, percentage of the group, the larger group whom you are actually going to target. Now, what is the y axis? Concentration or responders.

Now, you know that when you target or when you reach out to 10 million, which is your target group, of course all the 10 million will not become your customers. The actual number of customers hidden there would be a subset of this 10 million, which you do not know in advance. If you do know that number very specifically, you would have only reached out to them, but you only know a probability of response and therefore, you are reaching out to a larger number. So it, suppose this is 10, out of 10 million, suppose this is 1 million or even less. The y axis is the 1 million who are actually going to respond.



SCORING MODELS | BI&A | Prof. Saji K Mathew

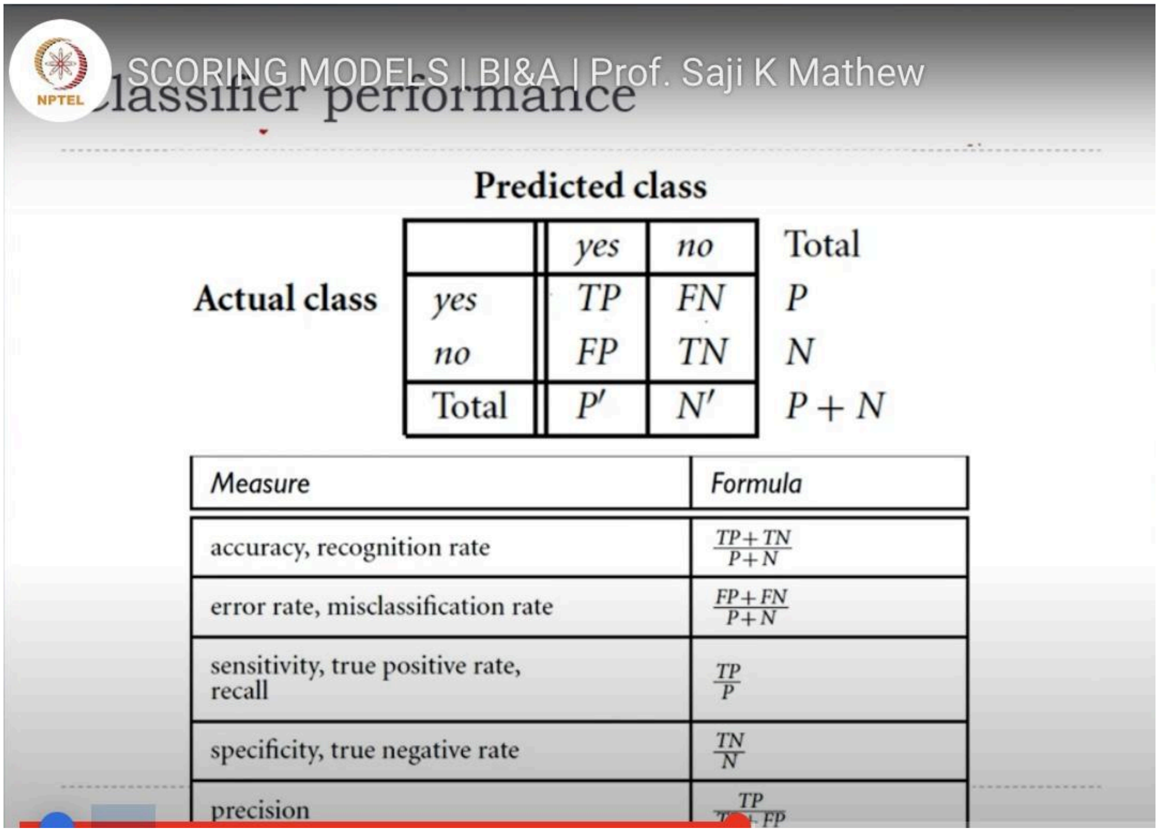
The cumulative gains chart

And each graduation of the y axis like the 10 percent, 20 percent and 30 percent is the percentage of the actual respondents. And now you can interpret this graph. Look at the 10 percent here. For example, you are sending your packets to the first 10 percent, that first 10 percent you can choose randomly, then the random response is again, you will reach to 10 percent of the respondents who are actually going to respond because there is no rule.

So it is just 10 percent. But suppose you used an algorithm to determine what is the first 10 percent I am going to reach out to chances are that by just reaching out to the 10 percent, you have reached out to 30 percent of the actual respondents because you know what rules determine a respondent. So this is an improvement over the random approach. And suppose you reach out to the 20 percent, randomly you reach out to the 20 percent who are actually going to respond, but using the algorithm, you reach out to a much higher proportion of the actual respondents. Of course, when you reach out to the whole 100 percent, all the 100 percent of the respondents do respond.

But you can stop at some point instead of going for all the target group you have. For example, by reaching out to about 40, 42 percent, you have reached out to actual 80 percent of respondents and you do not have to actually spend your money here. That is what this graph indicates. And suppose an algorithm perform and performs differently and it is shaped like this. You can see that by reaching out to much lesser, maybe reaching out to 25 percent or spending much less money, you are able to reach the 80 percent of the respondents.

And therefore, this algorithm performs better than this algorithm. L1, this is A1 and suppose this is A2, A1 is better than A2. So, there is also the concept of area under the curve, which we will see very soon. But this graph is very useful in understanding performance of algorithms when they are applied to business and that is known as cumulative gains chart. Now, we are coming back to a very important concept related to classification, which is performance of a classifier.



SCORING MODELS | BI&A | Prof. Saji K Mathew

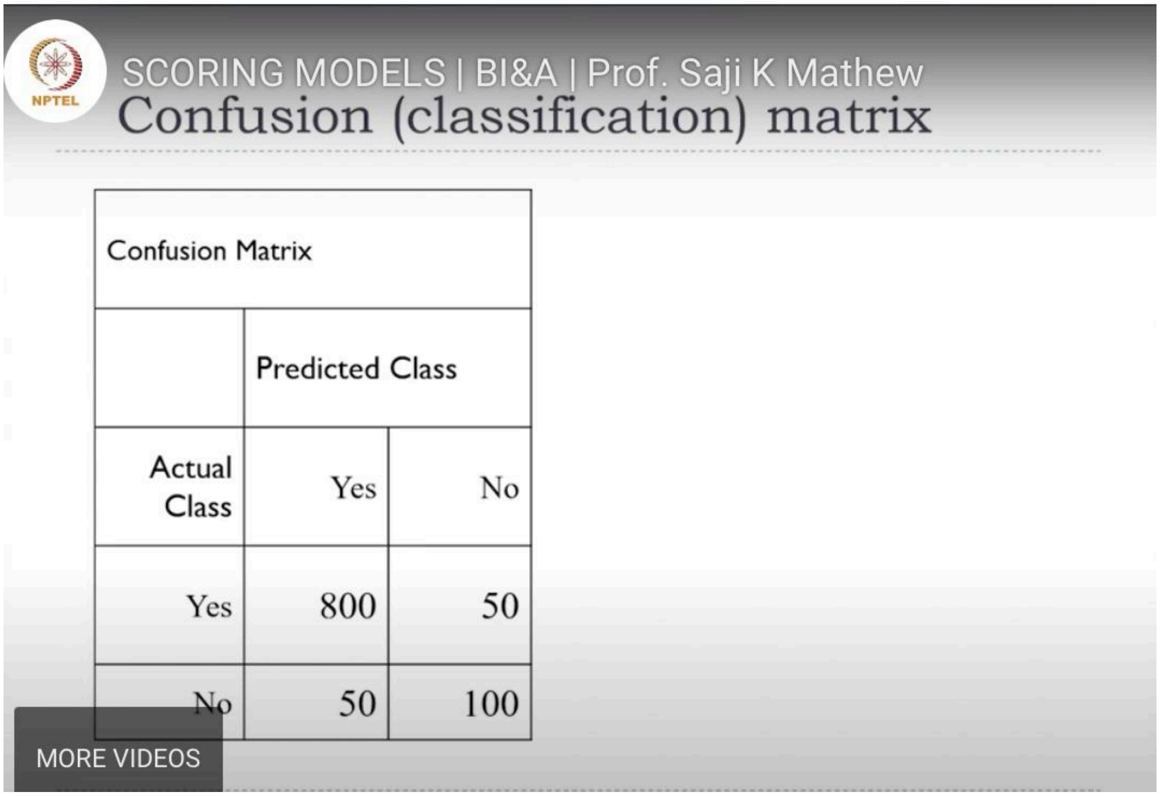
lassifier performance

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	*P′*	*N′*	*P + N*

Measure	*Formula*
accuracy, recognition rate	$\frac{TP+TN}{P+N}$
error rate, misclassification rate	$\frac{FP+FN}{P+N}$
sensitivity, true positive rate, recall	$\frac{TP}{P}$
specificity, true negative rate	$\frac{TN}{N}$
precision	$\frac{TP}{TP+FP}$

Classifier performance. In connection with regression or a general predictive model, we discussed how to do validation or cross-validation. So using cross-validation technique, you would measure how well a model predicts. So that is, that was a generic discussion and the example I took was more related to, say a context of regression or continuous valued variables, where you use an MSE or a similar measure to measure the prediction performance of a model. But when it comes to classification, you cannot use an MSE value. You can use, you cannot use mean square error as the measure because this is categorical variable.

So a target variable in classification is a categorical variable. It is not continuous valued. Categories are just categories. They do not have an order. And therefore, those measures do not apply here.



SCORING MODELS | BI&A | Prof. Saji K Mathew

Confusion (classification) matrix

Confusion Matrix		
	Predicted Class	
Actual Class	Yes	No
Yes	800	50
No	50	100

And therefore, you have a different set of measures to measure the performance of classifiers. And the starting point is a matrix known as confusion matrix, which is displayed here. The starting point is a confusion matrix. What is a confusion matrix? A confusion matrix actually shows you how well the model has classified. How well the model has assigned objects to different classes correctly.

Now the starting point is here, the test data. You of course, you built a model using training data, that is already done. That is your training data. So now you have the test data. And suppose your test data set has a particular size.

Suppose that is n, small n. That test data set consists of objects which belong to different classes. And here we are going to consider a very simple and intuitive case where the target variable is binary. Or it is categorical, but only two states, either yes or no, or either positive or negative. Yes, no, or positive or negative. So target variable is binary, either customer or non-customer, either respondent, non-respondent, male, female, low risk, high risk.

So there are different contexts in which binary classification is very useful. So that is what we are going to discuss first. So therefore, your given test data set consists of say n objects, but the objects are actually either p, either positive or negative. P + N is the total number of records or tuples or number of objects in your test data set. So, now you built a classifier already and the classifier is ready to predict and then you apply your test data set to the classifier one by one.

Suppose you take one object or one particular record, which is actually yes, the outcome variable or the target variable is actually yes. When you input the complete record to the classifier, the classifier could also classify that record or that object as positive, in which case we call it a true positive. Suppose a given object is actually yes, but the classifier classifies it as negative or no, that is the case of false negative. So true positive and false negative. Now one interesting observation here is, false negative is actually positive.

It is falsely classified as negative, but actually it is positive. So therefore, if you add TP and FN, P is equal to TP + FN, TP + FN, FN is actually positive. That is the total number of positives in your test data set. Now let us use the same logic in the next class label, which is no. Suppose an object is actually no and the classifier also classifies it as no, then that is a case of true negative.

And it is also possible that a object is actually no, but the classifier says it is positive, that is false and therefore, it is false positive. Now you can again see that a false positive is falsely classified. It is wrongly classified. It is actually not positive, it is negative. And therefore, n total number of negatives, actual negatives is true negatives classified by the classifier and false positives, which is wrongly identified.

So true negative plus false positive is the actual number of negatives in your data set. So please get these concepts correct. So total number of objects in the data set as we saw in the beginning, is P + N. So they get classified one diagonal as true positive and true negative, other diagonal as false positive and false negative. Now, with this, as the basic source for developing measures of classified performance, we have different measures for classified performance.

And that is what is shown in the next table. The measure on, the name of the measure on one column and the formula for it on the next column. So there is a measure of overall accuracy or recognition rate, overall accuracy or recognition rate, which is true positive plus true negative divided by the total number of objects. True positives and true negatives actually indicate how many classifications were done accurately. So that ratio provides an overall measure of performance of the model. And in similar terms, the opposite of it, the false positive plus false negative divided by P + N is the overall measure of error or misclassification rate.

And if I call this as accuracy, so this will be equal to 1 - a, obviously. So that is, these are two measures which actually, which actually provides a understanding of overall performance of the model. Now there are more specific measures and that is more interesting and useful. Which are they? There is a measure called sensitivity. It is also known as true positive rate, rate because it is a ratio and it is also known as recall.This is also known as recall.

And what is true positive rate, true positive rate or sensitivity? It is true positives divided by number of positives. It is very meaningful. Suppose the number of positives is 100, but classifier classified only 80 as positives. That is 80 percent.

The sensitivity is 80 percent or recall is 80 percent. But what happened to the remaining? What happened to the remaining positives? They were not identified. So there is a problem that the classifier could not identify some positives accurately or correctly. And we know that, we know from this formula that P = TP + FN or TP = P - FN. So if I extend that formula here, this is actually TP = (P - FN) / P or 1- false negative/ P.

The sensitivity is 1 minus false negative rate. Now, imagine there is no false negative. All positives in the data set have been correctly classified as positive. In which case TP becomes, TP = P or there is FN = 0. When FN becomes 0 or when there is no false negative, sensitivity or recall is 1. All positives have been completely identified. But that does not mean that there is no false positive. We only say that all true positives have been identified. And there is no false negative at all.

So look at the next ratio, which is specificity or true negative rate, which is TN/ N and extending the same logic, we can see that TN/ N is nothing but (1 - false positive)/ N.

Because false positive is actually negative. And again, we would argue here that when there is no false positive, the specificity becomes 1. Or all negatives have been correctly identified. TN = N or there is no false positive at all. No negatives actually went as false positive.

So that is the understanding that we can have here. Now, there is a final measure known as precision. Precision is a very interesting and important measure, which is true positives divided by true positives plus false positive. So the denominator is how many positives have been reported by the classifier of which how many are actually truly positive. That is a measure called precision. And you can again see here that when there is no false positive, when there is no false positive, then precision is equal to 1. Then when precision is equal to 1.

But there can be a lot of true positives. There can be all true positives covered in the classification or TP can be equal to P. But can there be false positives? Yes, because false positives are actually negatives. It identified all the positives, but it also identified certain negatives also as positives. Now can you imagine why we have multiple measures in classification? The reason why we need this multiple measures is because when you apply classification, depending on the context of application, sometimes a false positive is more costly than a false negative. And sometimes in certain context, a false negative is more costly than a false positive.

And therefore, the performance of the model in misclassification or correct classification with respect to positive or negative individually is important, because that is what is more important, not the overall accuracy. So let me give you a scenario. One scenario is medical diagnosis. In medical diagnosis, what is more costly, a false positive or a false negative? Generally a false negative. Somebody is actually sick, somebody has actually COVID, but the diagnosis which is actually a classification says the patient is healthy and does not have sickness. It is a false negative. So a false negative is more important than a false positive in medical diagnosis. And you see what is a measure which actually highlights the false negatives. You can see sensitivity has that included in it. When false negative is high, the sensitivity value comes down.

And therefore, you need to focus on the sensitivity. There are cases when a false positive is more costly than a false negative. Because it is a credit business, you give credit or you give money or assets to potential applicants. Of course, you evaluate the risk of the applicant, you do the credit worthiness evaluation through formal processes. But suppose a credit worthy customer, credit non-worthy customer is classified as a credit worthy customer, what you do is you give a huge sum or a loan to the customer and the customer does not pay back.

So a bad customer actually takes off huge sums of money and leads to huge loss as compared to a false negative. A false negative is actually a good customer, but you could not actually get that customer. Well, you lost some profit, but here you lose huge money. So the cost of misclassification in credit business is more critical when it comes to false positives. And therefore, what is a measure which focuses on false positives? You can see that it is specificity.

And precision actually gives you a sense of the performance of the model which also covers the false positive. For example, as I said the TP can be equal to P. In a formula like this the TP can be equal to P. All true positives are identified, but it has identified a lot of false positives also.

In which case the classifier actually is not doing well. So that aspect also should be evaluated or you have to look at multiple measures to evaluate a classifier. And that is the idea in having multiple measures for classifier performance.

Here is a confusion matrix as an example. And this confusion matrix actually enables you to do the necessary computations to calculate accuracy. Accuracy is equal to (TP + TN) / P + N. P + N is nothing but, P + N is the total number which is equal to 1000, which is equal to 1000.

850 + 150. So which is equal to 1000. So 1000 upon TP is 800, TN is 100. So therefore 900 / 1000 is the accuracy and what is the error? Error is the opposite. False positive plus false negative which is (50 + 50) /1000 is equal to you know 100/ 1000, 0.1. So that is the error rate. And now what is sensitivity? It is equal to 1, sorry sensitivity is equal to true positive divided by number of positives.

True positives is 800. Number of positives is 800 + 50. The 50 is also actually positive but falsely represented as negative. So 800/ 850. And what is specificity? True negative divided by number of negatives.

So true negative is actually 100. And actual number of negatives is 100/(100 + 50). And this is what is specificity and what is actually the last measure which is precision. Precision is equal to true positive divided by true positive plus false positive. Very interesting measure. True positive is equal to 800, divided by 800 plus, how many false positives? This is a false positive.

Actually no, but the predictor or the classifier called it yes, so 50. This is precision.

You can easily do the calculation of these measures from the data. And, but what is more important is how do you interpret this model? And how do you apply this model in practice to given situations or context? That is more important. Thank you. Thank you.