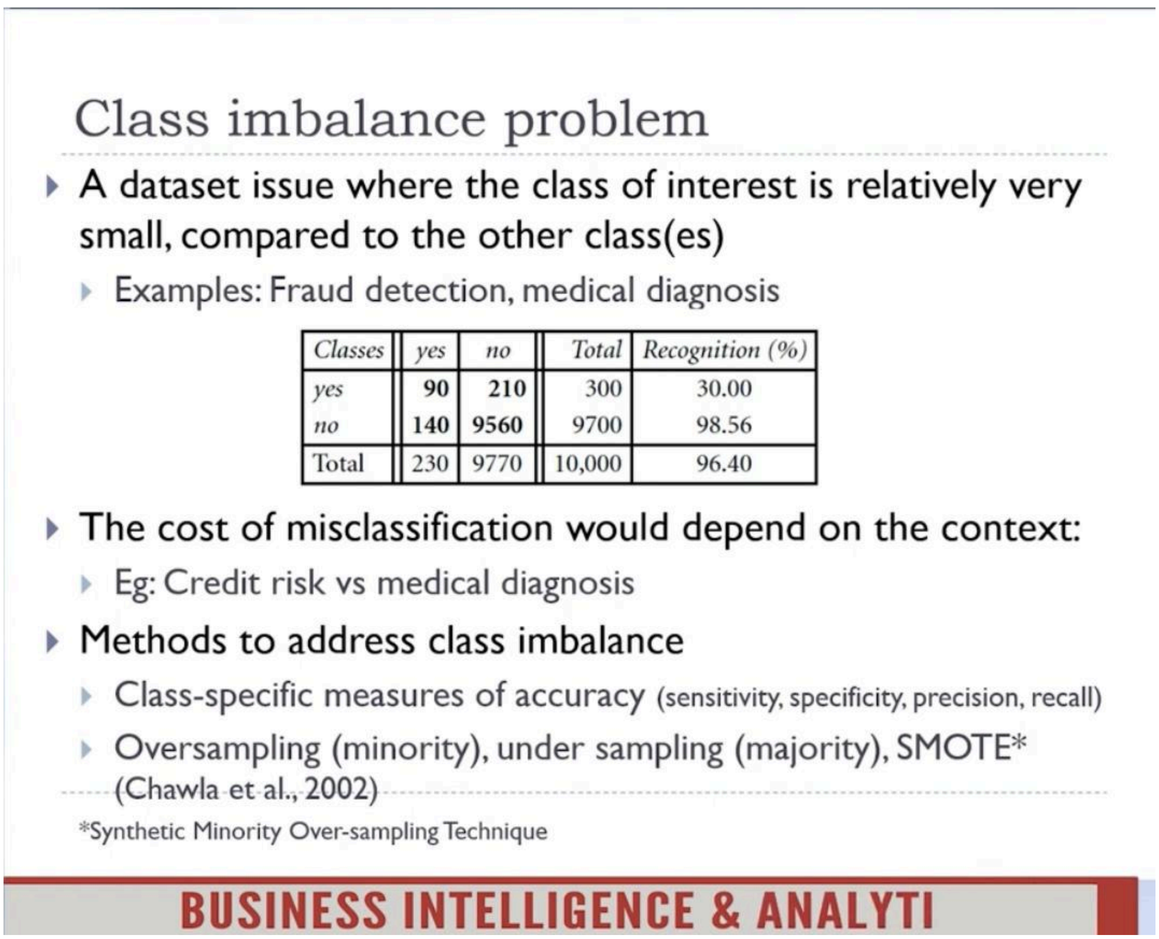
Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:06
Lecture:24

CLASSIFIER PERFORMANCE | Business Intelligence & Analytics



So, now we are going to discuss related problem in classification, after learning about various measures that are available for assessing the performance, the classification performance or the prediction performance of classifiers. So, we have seen different measures that are available and so given a set of data or set of data in the confusion matrix, how you can compute the different measures of classifier performance, that is what we have seen so far. Now, before we go on to understand how these different measures can give a researcher or an analyst different insights into the performance of the classifier, let us also be aware that in certain cases involving classification, there is a problem, there is a peculiar problem known as class imbalance problem. Class imbalance problem is a problem of data set or it is also a problem of the phenomenon that generates data set. So, in such cases, says suppose it is a binary classification problem and your target variable is binary, yes, no or positive and negative.

So, in certain cases like fraud detection, the number of fraud versus non-fraud transactions, just think about it, if it is credit card transaction, the number of fraud transactions would be relatively much lower than the number of non-fraud or number of credible transactions, number of good transactions. So, fraud transaction, the number of records that involves fraud transaction, if you take 100 records or if you take in 100 records, you may not find a fraud transaction at all. So, actually the percentage or the relative percentage of fraud transactions would be much low and this leads to a problem called class imbalance problem where fraud has very few as compared to regular or acceptable transaction. It is binary, fraud versus non-fraud.

Let me use the word non-fraud as a transaction which is not fraud, that may not be the most appropriate word, but fraud versus non-fraud, so binary. So, the fraud is a very, very rare occurrence and therefore the number of records would be very low. Therefore, you have a class imbalance problem. In such cases, if you try to build a decision tree or any classifier based on such data set, you will have a problem in getting accurate classification results for that class, for which the data is very small in size. So, that leads to a problem in assessing classifier performance.

So, look at the example that is given in the slide. So, you have a binary classification data set where the total size of the data set is 10,000. Out of which actual number of classes which are yes is 300, actual number of cases which are no or non-fraud, as I said is 9,700. 300 is to 9,700, think of that ratio. So, or out of 10,000, it is only 300 which is actually an yes or 300 divided by total is 10,000 or just 3 percent of the classes are yes or fraud or you know, having a tested positive for a particular disease etc.

So, therefore, there is a class imbalance problem and look at the recognition rate. So, when we look at the recognition rate, the true positive versus true positive divided by number of positives, this is actually TP divided by P which is 90/300 or that is just 30 percent. So, your sensitivity comes as a small or low value. But look at the other phenomenon which is your specificity or true negative recognition rate,sorry, that is not the cell I should have rounded, this is the one, 9560 out of total 9,700 which is 98.56.

What explains the difference between the true positive rate and the true negative rate or the difference between sensitivity and specificity? The simple reason is that the number of classes that are available with positives are very, very small in number proportionately. And therefore, your classifier is not trained well to perform prediction well or to perform classification well and you have that problem. If you have a more balanced data set and if the classifier was trained, based on that balanced data set, then you would expect a more comparable performance in terms of sensitivity and specificity. So, this is a problem with data or this is a problem with the phenomenon and you cannot change the phenomenon. But you can work with the data, you can work with the data and there are certain techniques that have been proposed to address this problem of class imbalance.

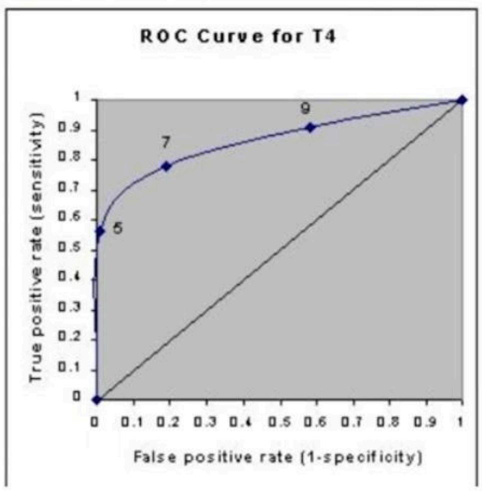
So, number one method of course, is to look at class specific measures of accuracy and that is what we have been doing. So, since in this case the specificity is measured with a much larger data set or part of the data set than sensitivity, therefore and if your interest is in true negatives and your interest is not in true positives, then you are fine. So, you do not have a problem. For example, we have already discussed that specificity is equal to true negative divided by negatives is equal to 1 minus false positive divided by negative. So, false positive is actually negative.

So, in situations where or in cases where or in problems where you are not, you are worried about false positive and false negative is not a problem. We discussed as an example, credit business yesterday. You do not want false positives, whereas in a medical diagnosis you do not want false negatives. So, depending on what is your focus, if your measure is good enough for you, because that is the measure that you focus on, then you are fine. But if not, so in this case if false negative is going to bother you more, then you do not have a good classifier here, because its recognition rate is very low or its sensitivity is very low.

So, therefore, you have the next alternative which is about working with the data set itself. So, there is actually a class of methods known as sub-sampling techniques, basically boost your data or increase the size of your data set from the main sample itself. The main sample itself is used to increase the size of the data set or from the main sample you can do two things and then you can also use a more advanced technique or more optimum approach known as synthetic minority over sampling technique. But before that, look at what are the first two techniques. One is over sampling.

So, you try to increase the size of the minority class by over sampling or boosting that particular category or the other work around is, you do under sampling, you reduce the relative size of the other class for which the data volume is high. So, either make minority more or increase the size of the minority or reduce the size of majority. And SMOTE is a compromise and it is not a compromise, it is a much more nuanced technique for minority over sampling. I am not getting into the details of how a SMOTE technique or algorithm works, but it is a method to over sample the minority class, so that there is a fair representation of both the classes in the data set itself. And after making the data set more balanced, then you build a classifier and there the classifier performance will be and the recognition rates will be better. That is the approach.

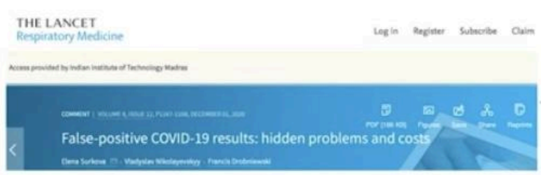
ROC curves



T4 value	Hypothyroid	Euthyroid
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93

Cutpoint	True Positives	False Positives
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58

Cutpoint	Sensitivity	Specificity
5	0.56	0.99
7	0.78	0.81
9	0.91	0.42



BUSINESS INTELLIGENCE & ANALYTICS

So, next topic related to classifier performance is the ROC curve or ROC in short form stands for receiver operating characteristics. ROC curves or receiver operating characteristics. So, you may all of a sudden wonder what is ROC, what is the receiver here. So, that is very interesting.

Actually, the ROC curves has its origin in World War II. During World War II, there is a particular incident called Pearl Harbor attack, where fighter aircrafts of Japan attacked US military base, US defense base. I am not sure whether it was military, US defense base in Pearl Harbor without any warning or they had radar. Of course, radar is a technology for sensing remote objects, when they approach an object. So they had radar, but despite having this alarm systems to announce an attack, the US defense was not able to properly detect and defend themselves. And it is a Pearl Harbor attack of Japan that actually forced US into the World War II, that is history.

But then, of course, the whole US defense establishment had to actually rethink what happened, what went wrong, why could not they actually detect the attack. So, there started, there is something wrong with the receiver or the radar. So, they recognize that an alarm need to be generated when there is a real occurrence and an alarm should not be generated when there is no occurrence. But unfortunately, this particular tuning of a device or a classifier cannot, is a trade-off, is a trade-off. You can actually make an alarm very sensitive, that is, only if there is a real attack, then only the alarm will sound.

But when you make the alarm or the receiver very sensitive, then the problem is that it may actually miss certain real incidents, because only if the incident is really positive, if it really occurrence only then the alarm will sound, but then there is a chance for false negatives. There was a real occurrence, but since the alarm is, it is like making the admission to a program very strict. You set a CAT score as 99.99. Well, CAT score has its own limitations.

So, you may get really top scorers of CAT, but you have the other problem that some students, potential students who missed it by a 0.01 may not make it, but they could have been much more promising students. So, you miss them. There are a lot of false negatives who could have been admitted, who, which could have been a real incident, but because the alarm is tuned to be extremely sensitive to true positives, it does not sometimes report an alarm. Think of the other extreme, when you tune your classifier such that even when there is a minor probability of occurrence of an incident, an alarm is reported.

Then that leads to a lot of false alarms or a lot of false positives. But there is an advantage there. You do not miss an incident if it is real, but along with the real incident, it may also report a lot of false incidents. So, all the time you may be listening to alarms and then you become insensitive to alarms. So, I think in real life, we are aware of this situation.

So, the ROC curves in one sense is depicting this phenomenon. The example given here in the three tables is basically this. Let me first try to explain this to you as sample data and then explain the ROC curve to you. But the basic idea is what I try to illustrate to you.

So, here the data set pertains to a medical diagnosis context. So, the classifier is basically a test, a test result called T4 value. You know in detecting hypothyroid, which is a problem of the thyroid, there are certain tests, blood tests that you know pathology labs do, TSH and T4 etc are useful in detecting the problem of hypothyroid. And so, they can set the T4 value, the cut off for T4 values at different levels. They can make it very strict, they can make it relatively loose. What does that mean? So, let us look at this table.

What this table depicts is that the column number 1 is about the different cut offs for T4. Column number 2 is actually the number of hypothyroid cases and the number of true hypothyroid cases that are detected for different cut offs. You know that the top, total number of hypothyroid cases which is the P is equal to 32. Total P is 32. But there are also euthyroid cases, which may also get detected as hypothyroid when you change the T4 value.

So, there are actually 93 euthyroid cases or the number of, no, euthyroid cases are not hypothyroid cases. So, look at the results now. When you keep T4 value cut off at 5 or less, of the 32 cases, 18 get detected. So, the true positive is 18.

The true positive is 18. What does that mean? 7, 4, 3, etc should have been detected or the 14 cases should have been detected. But they were not reported as true positives, they became false negatives. So, you can see that when the 18 is detected, so 18 by 32, which is 0.56. 0.56 is the true positive rate. When the cut off is 5, but when the cut off is 5, there is also a false positive. A particular case which is actually euthyroid also got reported as hypothyroid. So, you have one case, which is false positive. 1 divided by 93. 1 divided by 93. This is 18 divided by 32. So, that is the false positive rate. Now, look at what happens when you change the cut off. When you make the T4 value more relaxed, just like changing the CAT score from 99.99, you make it 95. You see that more true positive cases got detected. 18 + 7, 35, which makes it, sorry, which makes it 35 divided by 32. I am sorry, 25. I am sorry, this is 25. 18 + 7, which is 25. 25 by 32 got detected or that is 78 percent got detected. But what happens? When you relaxed, some cases which were not hypothyroid also got detected as hypothyroid because you relaxed the norm, you relaxed the cut off. So, 17 + 1, 18 cases now become actually false positives or 18 by 93. That is a false positive rate. And similarly, you further relax it, you make it 9 as the cut off, then another 4 gets added.

So, 91 percent of the, 91 percent of the positives get detected. But alongside you can see false positives. These cases which are not actually truly hypothyroid also get detected as hypothyroid because you are relaxing the cut off. So, you can see this reverse trend. When you increase the true positive rate, when you increase the true positive rate, the false positive rate is also increasing.

When you increase, try to increase the false positive, true positive rate, the false positive rate is also increasing. When you make any cut off relaxed, you will get more positives. But along with that, you will get more false positives also, because your condition is relaxed. That is a simple rational here. And you can again see that when sensitivity increases, specificity decreases.

There is an inverse relationship between sensitivity and specificity. Specificity here is nothing but 1 minus false positive divided by negative or 1 - 0.01, which is, this is nothing but false positive divided by N. We have seen that already. So, when sensitivity increases, specificity decreases. This is another relationship.

When true positives increases, false positives also increase. And that is the reason why specificity is decreasing. Let me say that again, when true positives increase, false positives also increase. And therefore, specificity decrease because specificity is 1 minus false positive divided by N. And it is the same phenomenon or it is the same observation from the data that is visualized in the form of an ROC curve, where x axis is the false positive rate and y axis is the true positive rate.

And what is this 5, 7 and 9 ? They are the cut offs. You cannot understand an ROC curve without understanding the real phenomenon behind it. What changes the true positive rate and the false positive rate is the change in the cut offs. Change in the cut offs or you decide the sensitivity. When you change sensitivity, it affects your false positive rate. Or when you, these interventions or deciding whether the cut off should be a 5 or a 7 or a 9 will have an influence on the false positive rate and in turn on specificity.

Look at a case when your cut off is 9. When your cut off is 9, you can see that relatively your true positive rate is high. You have reached a true positive rate, which is close to 0.91. This is 0.91. You reach 0.91. But what happens to the false positive rate? The false positive rate is 0.58. We have seen that already, 0.58. So, when true positive rate is increased, false positive rate is increased. Or when you increase the cut off or when you increase the sensitivity, basically, your true positive rate is increasing here or your sensitivity has increased. So, when you increase the sensitivity by changing the cut off, your false positive rate also increases. But look at a true positive rate or sensitivity here.

The sensitivity here or the true positive rate here is 0.56. It is 0.56. That is the cut off which is 5. It is 0.56. And what about the false positive rate? It is very low, 0.01. Your sensitivity is low, then your false positive rate is also low. But you, we already know that when you make the true positive rate low, whatever you get are really positives.

You do not get false positives. You are very sure about what you are getting. But you are missing out a lot of potential positives. You are missing out this range who are actually positive, but you are not able to detect them, because you have set your sensitivity at a low value, which is 5. Now, you can imagine how to apply this in different cases. For example, if it is medical diagnosis, where you do not want false negatives, you do not want false negatives, but you can still manage false positives.

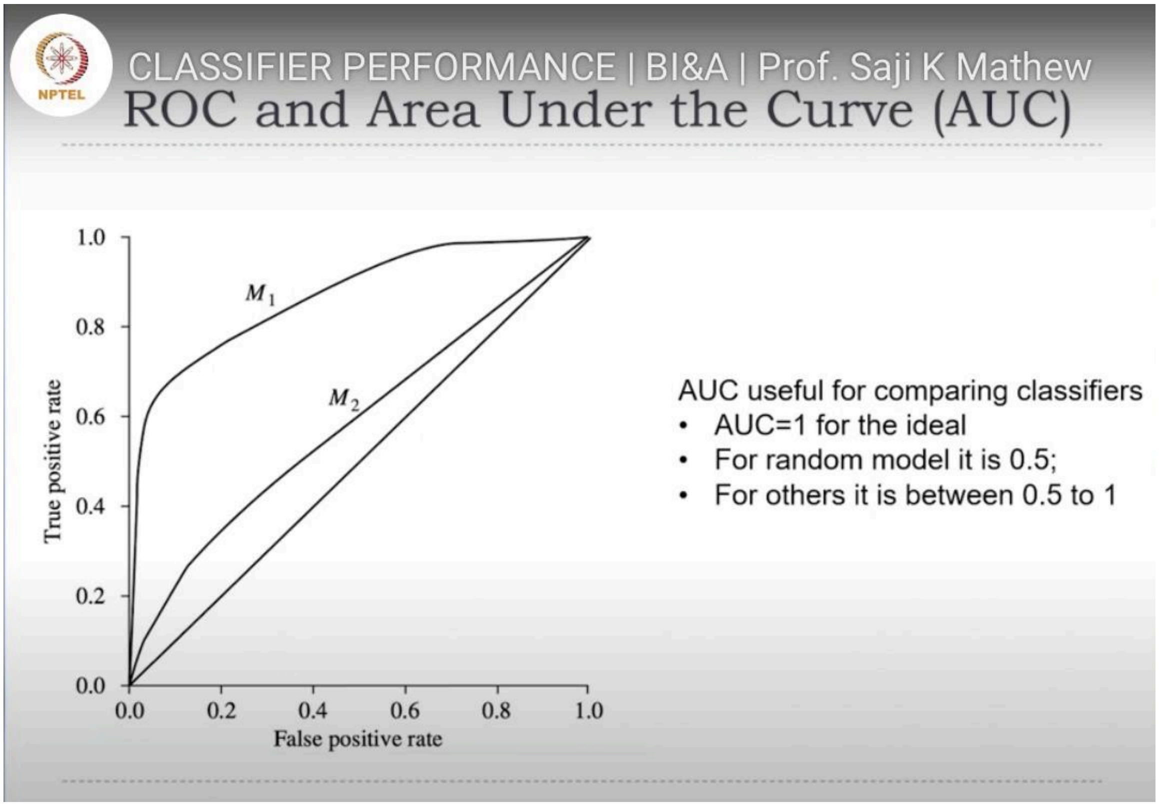
If someone is actually sick, but diagnosis did not report it. It is a false negative. And you want to keep the false negative very low. But false positive is acceptable, or you want to cover as many true positives as possible. That is a converse of it.

You do not want false negatives, or you want all the true positives. You would like to keep your cut off close to 9. You will move here. The only problem is, as we have already seen that that will result in a lot of false positives. Those who do not have sickness may actually get reported as having a particular disease or sickness. And you see that the Lancet during the COVID season, published an article.

Generally, in medical field, it is believed that false negatives are more costly than false positives. But Lancet reports suggested that false positive is also very costly, particularly in the case of COVID. And therefore, one may not go to the cut off at nine in such a case, one may actually find a more optimum point for COVID test cut off. We are familiar with this, with this test in recent past.

So, that is one application of this curve. You know, you visualize this. But in cases where false positives are more costly, as in the case of credit business we discussed, we may actually go towards here because the number of false positives are very low here. So, you may actually try to keep the cut off very low or you may try to keep the sensitivity very low, so that you have minimum false positives. So, the cut off is something that a researcher or an analyst or a decision maker can decide. Where do you want to tune your classifier? Where do you want to set the sensitivity depending on the application? And that is the insight a curve like the ROC curves provide us.

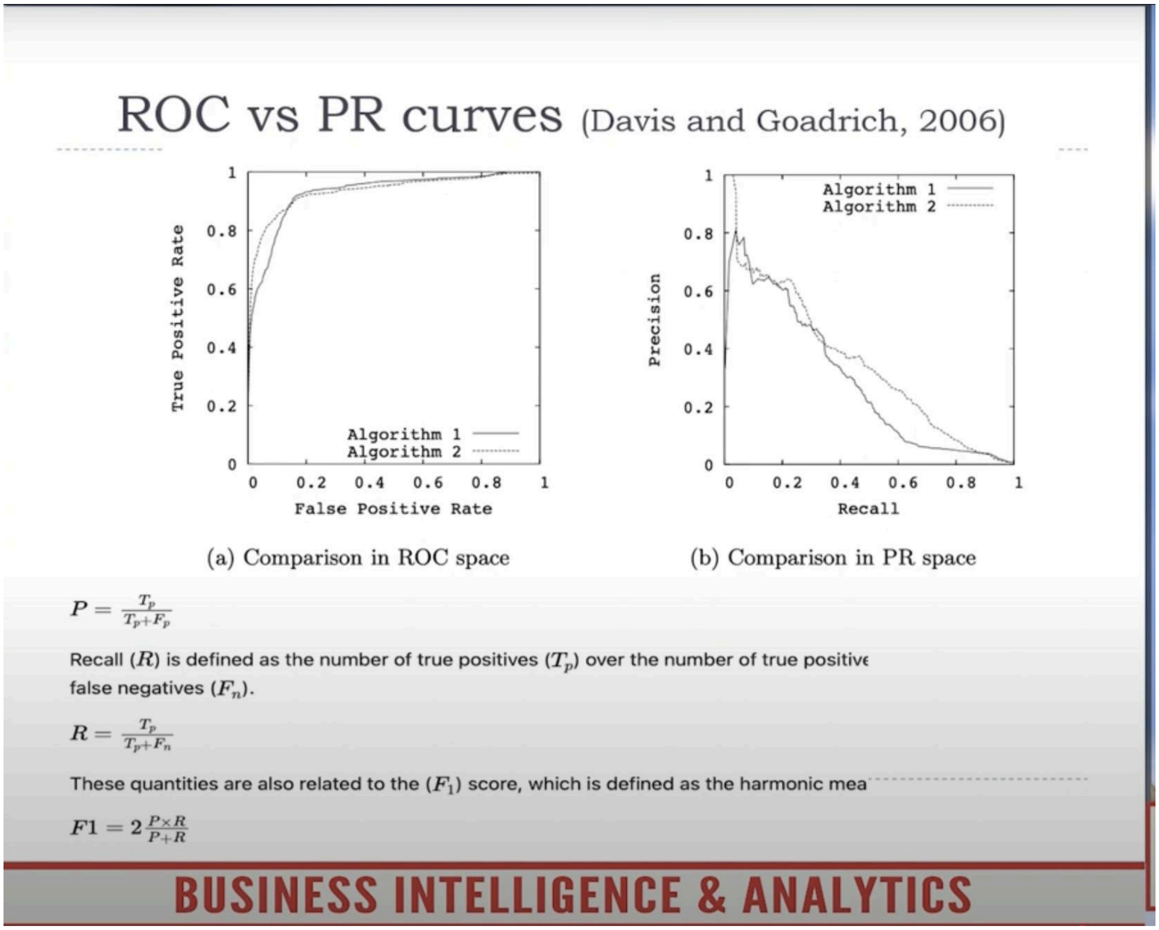
Now, here is visualization of ROC curves. And this represent the ROC curves of two classifiers M_1 and M_2. So, now looking at this graph, one should be able to recognize which classifier is performing better relatively. And of course, you can see a straight line, which is actually a random model or there is no model. If you actually classify randomly, you know, this is the true positive rate and false positive rate relationship that you see, because it is, it does not have any specific pattern. But if you use classifierM_2 , you can see that for a given false positive rate, this performs better because your true positive rate is much higher than that of a random model.



But look at M_1 for a given false positive rate, your M_1 performs much better in terms of true positive rate than M_2. What would be an ideal curve? An ideal curve would be that it will recognize all the true positives, without having any false positive, which shows that it actually comes here. And it remains there, even if the true positives, false positives increase by changing the sensitivity, but the sensitivity at the first instance itself is 1. And now, one way to have an objective measure of the performance of a classifier bor relative performance of a classifier is to use a measure called area under the curve, area under the curve or AUC. So, AUC for a random model, you can see is just half of this, just half of this square, this is a square because the scales are 1.

So, just half or let us take 1×1 and the half of it is 0.5, this is random. You see it here. AUC will be 1 for an ideal model, when for 0 false positive rate, it gives 1 true positive rate. And therefore, the ideal one is AUC - 1.

Any model, for any reasonable model, the area under the curve will vary between 0.5 and 1. And the more it is tending towards 1, the better the model is. So obviously, by looking at the curves, we can see that M_1 performs better than M_2. This is one way of commenting on the performance of classifiers using the ROC curves by comparing only the true positive rate and the false positive rate.



Now, we have another set of graph, which compares precision and recall along with ROC. So, the graph on the left side is the ROC curve, the graph on the right side is the is the precision, precision versus recall, PR stands for precision versus recall. The x axis is recall, the y axis is precision. In ROC, the x axis is false positive rate and y axis is true positive rate. And these two graphs when they are put together gives us more insight about how the algorithm or the classifier performs, taken from a research paper of Davis and Goadrichon ROC versus PR curves.

So, you can see this comparison. And we already learned that an ideal curve, are as far as ROC is concerned would be tending towards the north west, it should tend towards the northwest and that is more ideal. And the curve that is falling towards the middle of the square is not desirable. Now, when we come to precision versus recall, it is a different dynamics. What is precision? We have already learned that precision is true positive divided by true positive plus false positive.

Precision is true positive divided by true positive plus false positive. So, we also learned that when false positives increase, precision decreases. When false positives increases, precision decreases. When false positives increases, we have already seen that sensitivity increases, but precision decreases. ROC curves informs us about sensitivity and how this is related to the false positive rate. And we know that if you are willing to increase the cut off so that we let in more false positives, obviously the sensitivity increases or true positive rate increases, but that is not enough.

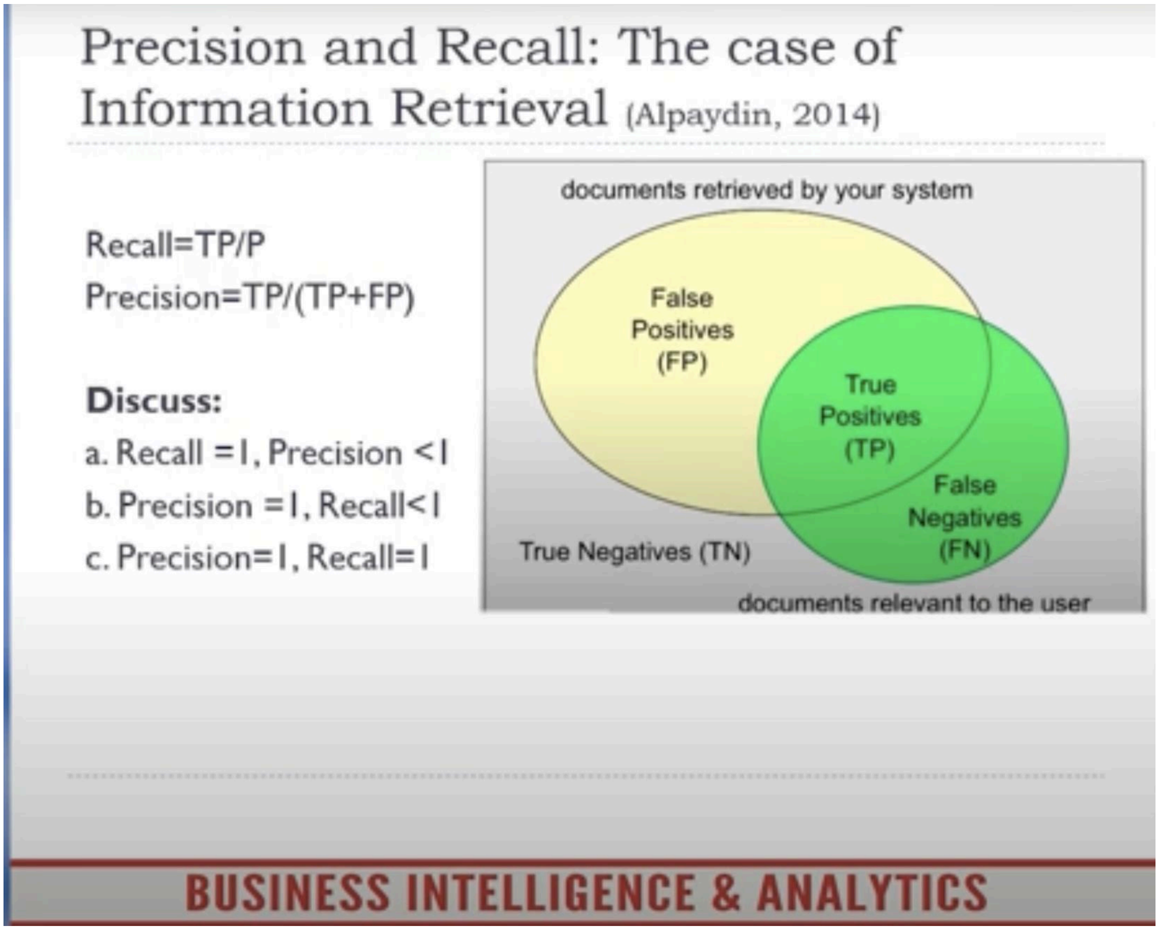
We also look at a countermeasure which suggests that alongside the false positives has also increased and that measure is precision. So, recall is TP by TP plus false negative, which is same as TP divided by positive. Positive is nothing but the sum of true positives and false negatives.

And this is nothing but sensitivity. All right, we have seen this already. So, what is actually brought it here in the x axis is sensitivity, which is recall. So, it is called precision recall curves, it is not generally called precision sensitivity curves, but conceptually it is the same. So, recall may be increasing or sensitivity may be increasing, but what happens when sensitivity increases, precision decreases.

When sensitivity increases, precision decreases. Why? Because false positive increases. You are increasing the false positive, when you are increasing the true positive rate or recall. So therefore, what is a more desirable curve? We do not want precision to go down. We do not want true positive, false positive rate to go up to classify true positive rate or sensitivity. So therefore, the more a graph is towards the north east, the better the graph is because you are getting higher precision, when the sensitivity increases. But if a graph is tending towards that, then it shows that when you increase sensitivity, precision is falling or false positives are increasing too much.

But along with sensitivity, if false positives are not increasing that much, you get a better shape if the graph is towards the north and east. And that is the sort of insight that you get on the trade-off between sensitivity and precision or precision and recall. So, the precision and recall are inversely related. You can see that. We saw that from the data already a while ago, but the graph is depicting that. So, these graphs together informs us about the trade-off between the different measures and depending on the application, one could actually figure how different classifiers are performing by plotting their ROC curves and PR curves. And this is very insightful.

And there is also a measure known as F1 score, which is the harmonic mean of precision and recall $2 \times PR/(P + R)$, which is another measure, objective measure which informs about the overall relationship between P and R. Now, let me give you another example, before I close this discussion on classification.



Classification is a very important technique and an important problem. And it has diverse applications and the kind of objectives in classifications could also be different. And that is why we have different measures. This is something that we have already discussed.

Let me actually present you a context that is that is presented by Alpaydin in his textbook. And that is the case of information retrieval. Information retrieval is about searching documents from large databases. You are a researcher and you are searching for research papers relevant to your research topics. So, what do you do? In systematic literature review, for example, what you do is, you design keywords. And based on the keywords, you run queries or you run search in databases, in research databases.

And when you run these queries, you get a lot of results. You may get 10s and 1000s of papers. Sometimes the papers run to 100,000 or more. You have a problem there because of the large numbers. So, which papers are relevant, which papers are not relevant. And finalizing, going through the results and finalizing the papers that you really need to read and review is a huge challenge.

So, it is that problem that is depicted here in the form of information retrieval. The problem I described to you is actually a problem of information retrieval. So, we have already seen what is recall or same as sensitivity and precision, PR. Now, the diagram, the Venn diagram in the slide actually helps you imagine different scenarios.

Now, there are two categories of results. The first category is documents relevant to the user. I am searching for a particular set of research papers related to a topic. And documents related only to that topic is relevant to me. But my search results may have a lot of results, which are not relevant to me. They are actually false positives. My interest is only true positive. I want papers that are relevant to me and that should be retrieved by the system also.

Documents retrieved by your system is actually this circle. Documents relevant to the user is the green circle. The green circle is documents relevant to me. The yellow circle is document that is actually retrieved. And I can see that document that is retrieved and relevant, which is the intersection and that is a true positive, is a subset of all the documents that is relevant to me. And it is also a subset of documents that is retrieved. Now in this case, looking at this, there are three scenarios. And I want you to work with me in understanding and appreciating what each of the scenarios represent.

Number 1, recall is equal to 1, but precision is less than 1. Recall is equal to 1. That is sensitivity is equal to 1. What does that mean? Sensitivity is equal to 1 when TP is equal to P. Sensitivity is equal to 1 or recall is equal to 1 when TP is equal to P, meaning all the positives, all the documents relevant to me have been retrieved. So, it is exhaustive, collectively exhaustive or the green circle has gone inside the yellow circle.

The retrieved documents have all the relevant documents or all positives have been retrieved. Therefore, sensitivity or recall is equal to 1. But there can be a situation when precision is less than 1. When does that happen? You have a lot of documents retrieved, but the relevant documents is this, but a lot of false positives.

This is complete positive. So, the results contain all my relevant documents, nothing is left out. But I now have to skim through the entire large number of relevant documents that have come which are false positives to get my positives. That is not, that is not a very desirable situation. So, this is the case 1, when all true positives are exhaustive, but lot of false positives. The other case is when precision is equal to 1 and recall is less than 1.

What is that situation? Precision equals 1 only means that FP = 0. Correct? There is no false positive. There is no false positive. It is TP by TP. So, all positives that have been identified are positives, are truly positive.

There is no false positive. There is nothing that is reported as false positive. So, what happens here? The search result is not exhaustive. Why do I say that? Recall is less than 1? Recall less than 1 means TP is less than P, TP is less than P or the search did not accomplish the full results.

Whatever it has reported, they are all positive. There is no false positive. But the search is not exhaustive. It is like the circles moving the other way around. There is no false positive. And therefore, there is only one circle and that circle is, that circle is the circle of positives, positives.

But not all the relevant documents have been identified. Not all the relevant documents have been identified. And therefore, lot of, lot of false negatives, lot of false negatives, false negatives is high. That is why you have a situation when true positive is less than P. Correct? Now, what is the ideal condition? You do not want a search where you get a lot of junk, along with useful material. You do not want a result where there is no junk, but the result that is obtained is not exhaustive.

You want your results to be sound, where precision should be 1 and recall should be 1. And that is, that is the ideal expectation. But the search results are satisfying or information retrieval is satisfying when both precision and recall tend towards 1. Not just the 1, not just 1 alone. Just because you have precision 1 or there is no false positive is not completely satisfying, if recall is low. Because you are missing lot of positives.

And on the other hand, if you have a lot of irrelevant materials, that is again a problem. It adds to the complexity. So, this case again enables you to appreciate the need for these two measures, recall and precision. And the need to look at both the measures together to understand how the classifier is performing.

With that, I close this lecture on classifier and classifier performance. And we would now move on to understand how a classifier algorithm works. And for that, I am going to take decision trees as a special case or a special group of algorithms. Within decision trees, there are different types of algorithms. So, we will discuss the broad concepts related to decision trees. And then also apply decision trees to solve a real life problem with a data set. And that is our effort in the subsequent sessions.