**Evaluating Classification Model Performance**

1. **False Positives and False Negatives –**

* False Positive (Type I Error) – It means that the outcome predicted to be true (for example in logistic regression) was indeed false.
* False negative (Type II Error) – It means that the outcome predicted to be false (for example in a logistic regression) was indeed true.

1. **Confusion Matrix –**

y^ Predicted Dependent Variable

y (Actual Dependent Variable)

|  |  |  |
| --- | --- | --- |
|  | Negative | Positive |
| Negative | Count of negative values predicted to be negative  (True Negative) | Count of negative values predicted to be positive  (Type I error – False Positive) |
| Positive | Count of positive values predicted to be negative  (Type II error – False Negative) | Count of positive values predicted to be negative  (True Positive) |

* We can calculate two values using the confusion matrix:
* Accuracy Rate = True Positive + True Negative/Total
* Error Rate = False Positive + False Negative/Total

1. **Cumulative Accuracy Profile –**

* Let’s say you are a data scientist at a store which sells clothes.
* The store has a customer base of 100,000 customers, and you know that, from experience, when you send a communication email to a random sample of customer approximately 10% of the customers respond to that email and purchase the product.
* So, we have got an offer that we want to send, and we want to see how many customers will purchase the product once we send it off.
* If we send it to 0 customers, we will obviously get 0 responses. But what will happen if we send it to 20,000 customers? Because this is a random sample, and about 10% will respond, then about 2,000 will respond, and so on.
* So, this is a random selection process.
* Here, we can draw a line that will represent the random selection process.
* The slope of the line will equal to the 10% that we know respond to the email communication that we send.

Chart, line chart

Description automatically generated

* But the question is, can we somehow improve this experience? Can we get more customers to respond to our offers when we send out our letter? Can we somehow target our customers more appropriately to get a more better response rate?
* How about instead of sending out the offer email to random 20,000 customers, how about we pick and choose the customers we send the offers too?
* How do we pick and choose customers to send the emails to? To start off with, we can build a model – basically a customer segmentation model – which will predict whether they will purchase the product or not.
* The model will basically classify the customers based on various factors – age, gender, geographic location, etc. – and calculate the probability of the user to buy respond to the offer we send.
* So, if we contact the customers based on the model we make, we will get a higher rate than 10% because we are picking out the customers that are at the highest risk of responding to the email.
* The graph below shows the response rate of the customers that we picked using the model.

A picture containing chart

Description automatically generated

* The red line shown in the graph, is the cumulative accuracy profile of the model.
* And as you can imagine, the better your model the larger will be your area between two lines, i.e., as your model gets better, the area between the red and the blue line will increase, but if your model gets worse, the blue line will be closer to the red line.
* And if a model goes below the blue line (below the outcome of the randomly selected customers), then that model is doing you a disservice.

(Note – A lot of people get confused between CAP (Cumulative Accuracy Profile) and ROC (Receiver Operating Characteristics), but they are not the same thing.)

1. **CAP Analysis –**

* As we have discussed there are three lines that are important on the CAP curve – the model developed from the randomly selected data from the data set, the good model developed from a data set tailored fit to deliver result, and perfect model to deliver the most efficient result.

A picture containing diagram

Description automatically generated

* What insights can we derive from this CAP curve?
* We know that the closer our good model is to the perfect model curve, the better it is.
* There is a standard approach to calculate the accuracy ratio – to calculate the accuracy ratio, you need to take the area under the perfect model, called ap, and then you need to take the area under the red line, called ar.

Chart

Description automatically generated

* In the next step you divide ar by ap, and the ratio of that would be between 0 and 1. The closer the ratio is to 1, the better is your model.
* It can be really complicated to calculate the area under the two, statical tools can do it for you, but how can you assess the CAP curve by looking at it.
* There is a second approach – we can analyze the CAP curve based on the outcome we will have if we select 50% of your population for the model we developed.

A picture containing diagram

Description automatically generated