CAP CURVE

Let's say you're a data scientist at a store which sells clothes and your store has a total of 1,00,000 customers.

And you know that from experience whenever you send an offer like an e-mail to all your customers or to any random sample of your customers, approximately 10 percent of them respond and purchase the product.

So, we've got an offer that we want to send and we want to see how many customers are going to purchase the product.

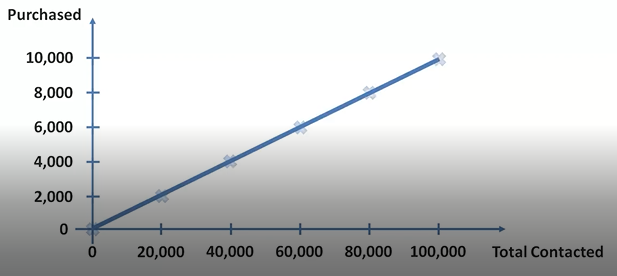
If we send to 0 -> zero respones

20,000 -> 2,000 responses

40,000 -> 4,000

and so on...

Here we can draw line which will represent this random selection.



Can we somehow improve this experience get more customers to respond to offers. And how about instead of sending these offers randomly to 20,000 customers, we pick and choose the customers we send these offers to.

For this we will build a model.

We can take a group of customers before we send out the offer and then look back and see who purchased, a male or a female, which country were they in, what age, were they browsing on mobile or via computer and all of these factors we can take them into account, measure them put them into a logistic regression and get a model which will help us assess the likelihood of certain types of customers purchasing based on their characteristics or the general demographic status other characteristics.

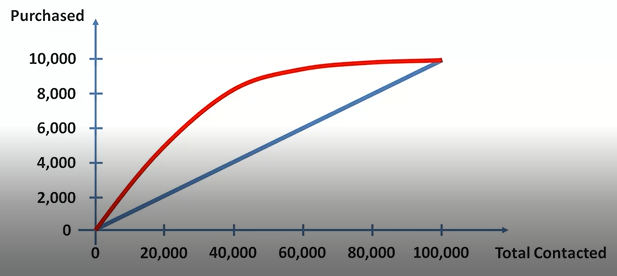
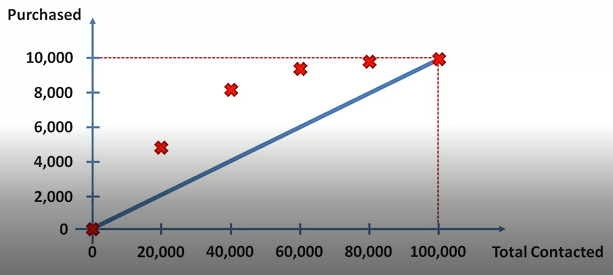
And once we've built this model we apply it to select customers we will send the offer to. That model will rank our customers who give them a probability to purchase a product. And then we can use that probability to actually contact our customers.

Now if we contact 0 customers -> 0 responses

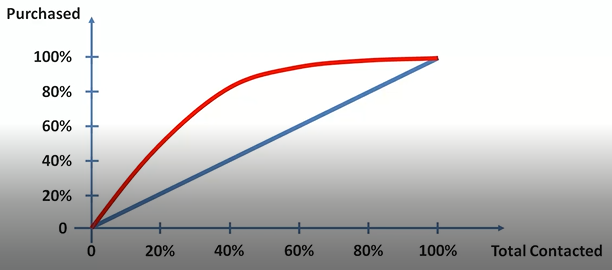
20,000 -> higher than 2,000

60,000 -> close to 10,000

1,00,000 -> 10,000



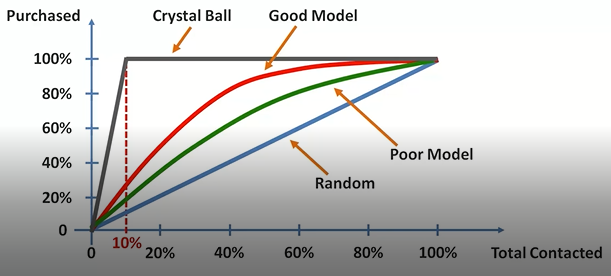
The next step we want to do is convert these axes from absolute values to percentages because this is how we generally represent a cap curve.



This red line is called cumulative accuracy profile of your model.

The better the model, larger the area between these lines.

CAP CURVE ANALYSIS



There are three lines that are important on the cap curve.

Blue line which is the random line when you select your samples at random.

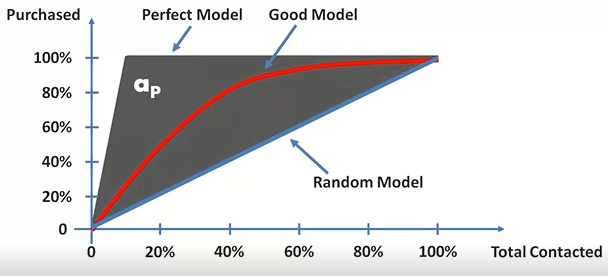
Red line which is our model line and different models will have different red lines.

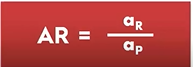
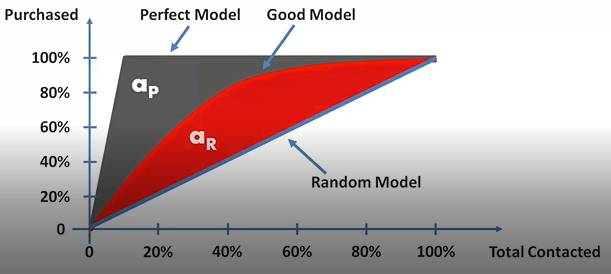
Gray line which is the perfect model or when you have a crystal ball when you can select all of the future turners or purchaser's or whatever action takers and you can select them right away on the dot before even selecting one single person that you don't want to select.

What can you derive from this curve?

-> The closer your red line is to your gray line, the better your model is.

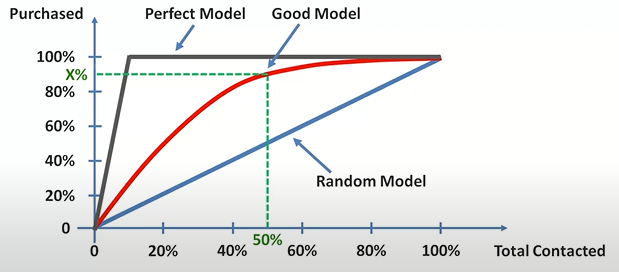
The closer your red line is to your blue line, the worse your model is.





Better approach than calculating AR:

Look at the 50% line and where it crosses the model.



If:

x< 60% Rubbish Model  
60% <x< 70% Poor Model  
70% <x< 80% Good Model  
80% <x< 90% Very Good Model  
x > 90% Too Good to be True (over fitting can be there)