Logistic Regression

Properties of Exponents

1. $x^3 = x \cdot x \cdot x$

2. $x^4 \cdot x^2 = x^6$

3. $(x^2)^4 = x^8$

4. $(xy)^a = x^a y^a$

5. $\frac{x^a}{x^b} = x^{a-b}$

6. $x^0 = 1$

7. $\frac{1}{x^a} = x^{-a}$

8. $\sqrt{x} = x^{\frac{1}{2}}$

Logarithm Property

① if $3^4 = 81$
then $\log_3 81 = 4$

② $\boxed{\log_b a = x \longrightarrow b^x = a}$

③ $\log_2 (1/2) = -1 \longrightarrow 2^{-1} = \frac{1}{2}$

④ Most commonly use $\log_{10}$ or $\log_e$
$\log_e$ also written as $\ln$ or natural log.

⑤ Change of base:

$$\log_b a = \frac{\log_x a}{\log_x b} \longrightarrow \log_3 8 = \frac{\log_e 8}{\log_e 3}$$

OR $\dfrac{\log_{10} 8}{\log_{10} 3}$

⑥ $\log_{10} \approx \log$ & $\log_e = \ln$

Binary Classification

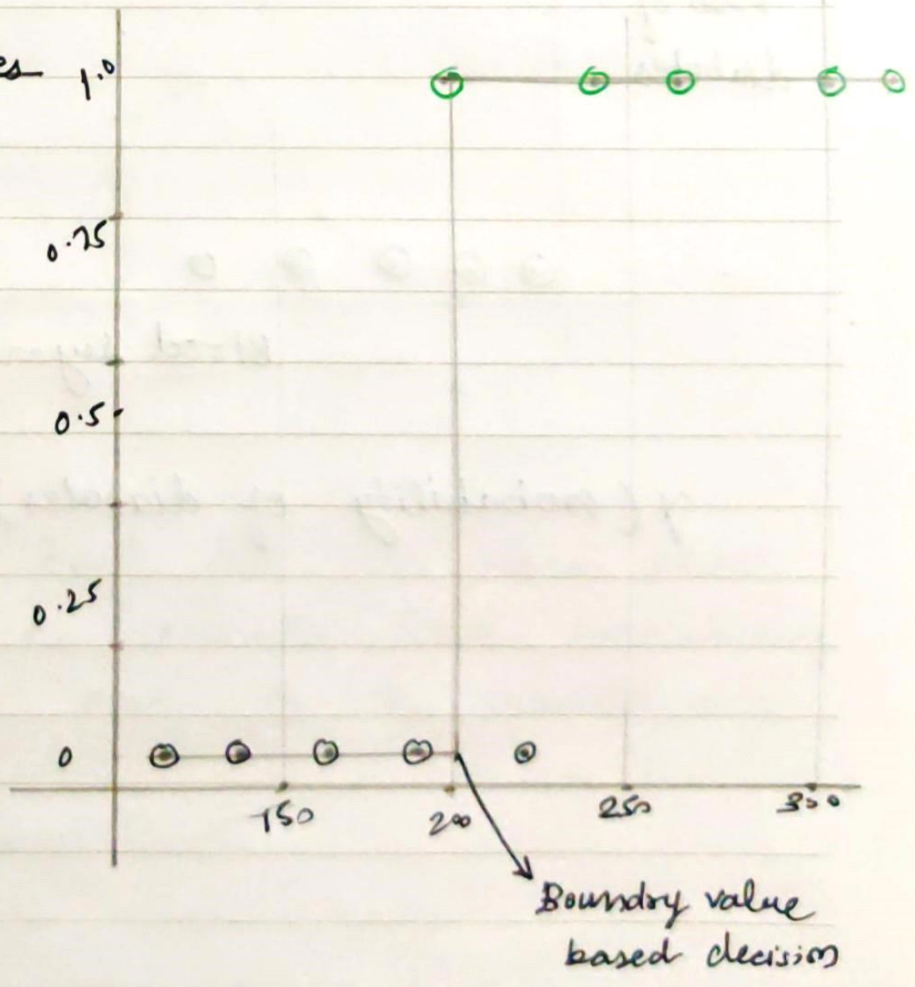
* Two possible outcomes.

Ex:- person will default in loan or not.

Email is spam or ham.

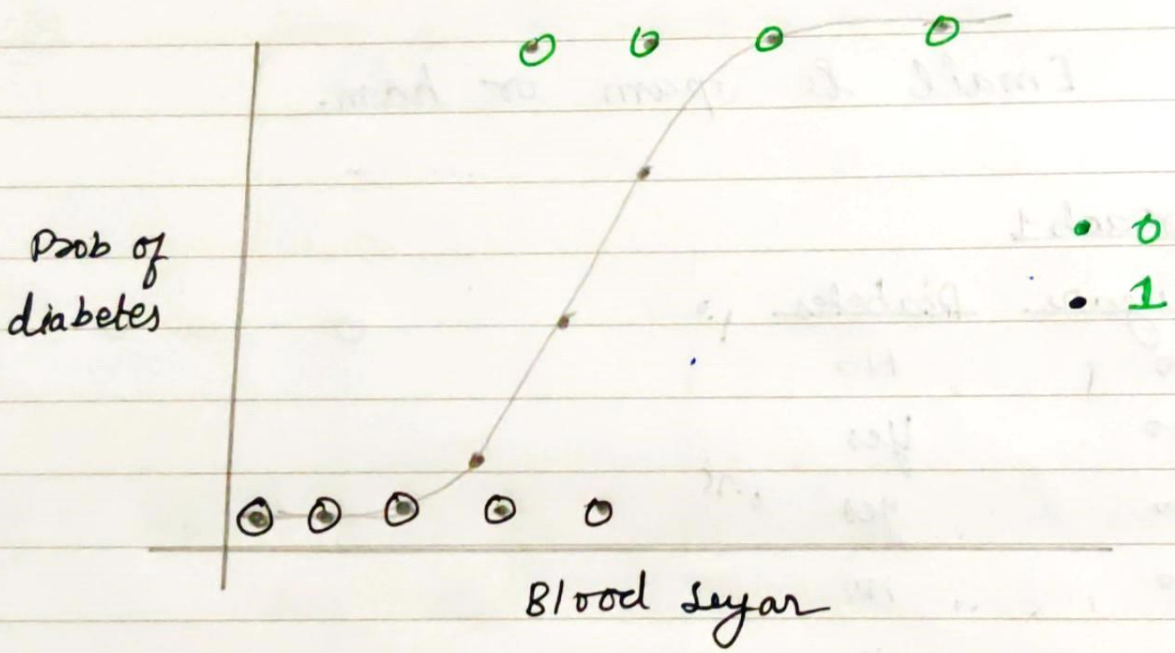
Approach 1

Blood Sugar	Diabetes
190	No
240	Yes
300	yes
160	no
200	yes
269	yes
129	no
141	no
220	no
337	yes



Sigmaid Curve

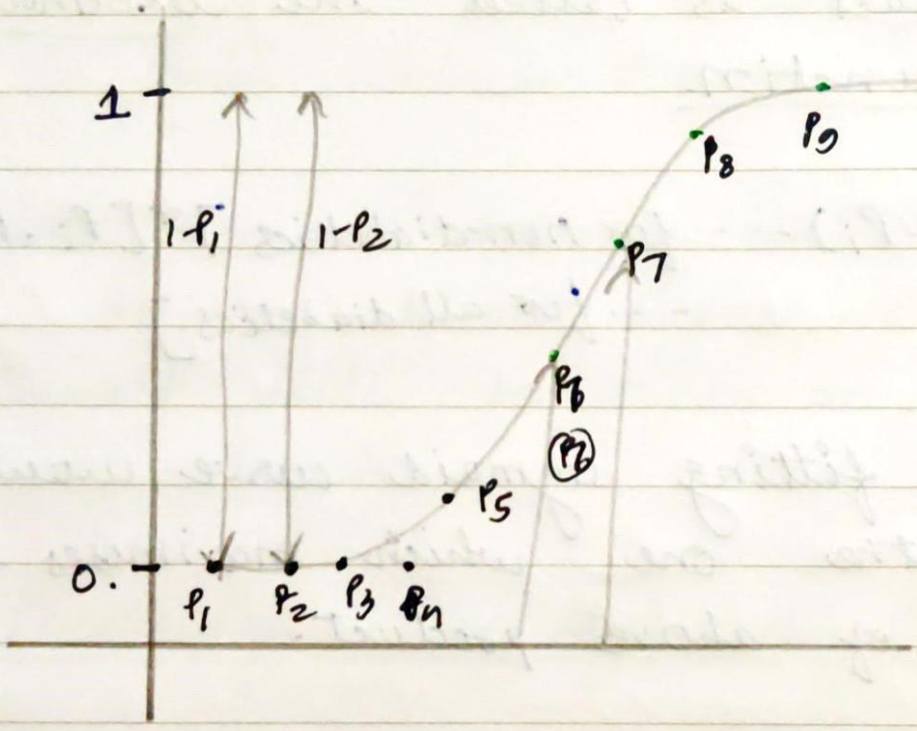
As we know simple boundry method is not accurate. so we try to make this line as curve so it will fufill our properties.



$$y(\text{probability of diabetes}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Finding Best Fit Sigmaid Curve :-

In the formula we can put multiple values of β_0 & β_1 and make many curves, but we need to find combination of β_0 & β_1 which fits the data set.



to find the best fit we need point like P_1, P_2 should have probability closer to 0 and P_7, P_8 should closer to 1.

$$\frac{P}{1-P} = e^{(\beta_0 + \beta_1 x)}$$

take log [here log = $\log_e$ = ln]

$$\boxed{\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x}$$

Interpreting :

$\frac{P}{1-P}$ = odds $\quad\quad$ $\ln\left(\frac{P}{1-P}\right)$ = log odds

Ex:- suppose $\frac{P}{1-P} = 4$

P(Diabetes) = 4 * P (No Diabetes)

Means prob of people having diabetes is 4 times higher then people not having diabetes.

P(diabetes) = 0.8 = 80%

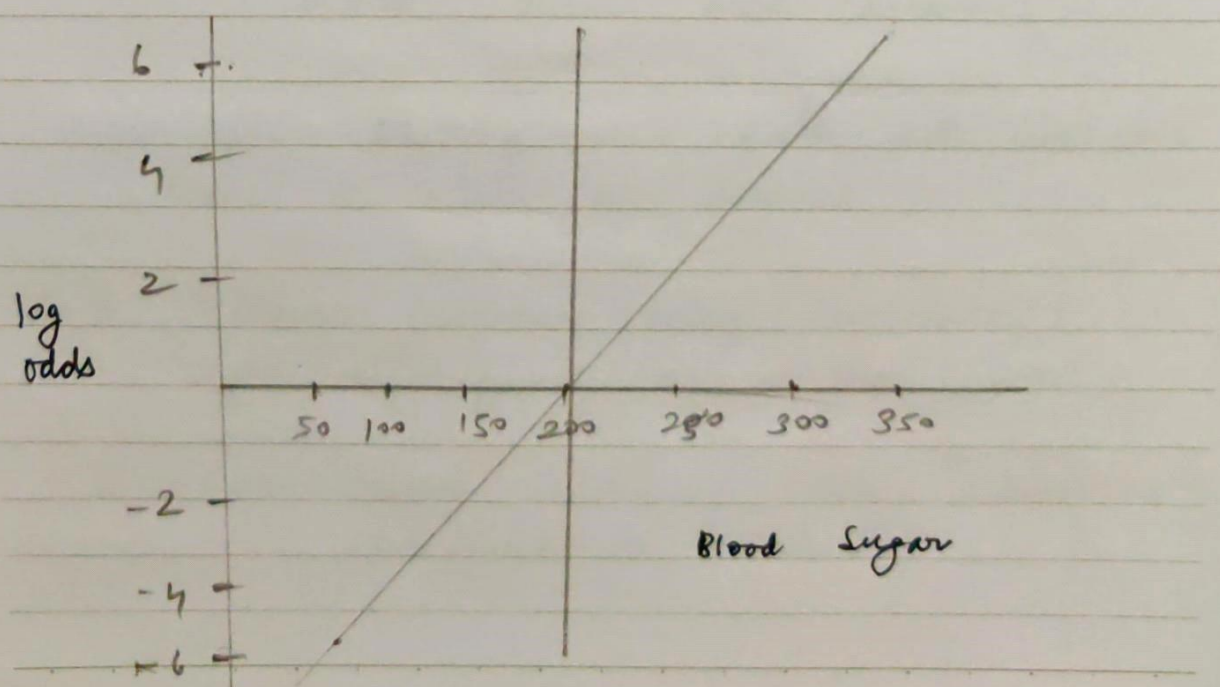
$\frac{P}{1-P} = 1 \Rightarrow P = 50\%$, (1-P) = 50%

4, 1 are odds here.

So in the end we get value of

$\beta_0 = -13.52$, $\beta_1 = 0.063$

Blood Sugar	Prob.	odds	log odds.
129	0.005	0.005	−5.30
141	0.011	0.011	−4.51
160	0.034	0.036	−3.32
190	0.194	0.241	−1.42
200	0.313	0.456	−0.79
220	0.619	1.625	0.49
240	0.853	5.803	1.76
269	0.974	37.46	3.62
300	0.996	249.0	5.52
337	0.999	999.0	6.91



$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

$$\frac{p}{1-p} = e^{(\beta_0 + \beta_1 x)}$$

Module 3 Session : 2

Multivariate Logistic Regression

Confusion Matrix

After completing the model when you will predict the answers it will not be in form of 0-1 it will probability between 0 to 1.

So what we do, is decide a threshold for example in churn case study

if < 0.5 then CHURN
> 0.5 then Non-Churn

Now after testing our data set we can have 2 class of errors:

1. 'Churn' customer being (incorrectly) classified as 'Non-Churn'

2. 'Non-Churn' customers being (incorrectly) classified as 'Churn'.

To capture these error we make something called confusion matrix.

Actual \ Predicted	No (Non-Churn)	Yes (Churn)
No (Non-Churn)	1406	143
Yes (Churn)	263	298

Accuracy: It is the percentage of correctly predicted label.

$$\text{Accuracy} = \frac{\text{Correctly Predicted Labels}}{\text{Total No. of Labels.}}$$

$$\text{Accuracy} = \frac{1406 + 298}{1406 + 143 + 263 + 298}$$

$$\approx 80.75$$

Manual Feature Elimination

$$VIF_i = \frac{1}{1 - R_i^2}$$

High value of VIF is not good.
Usually greater than 4 is bad.

Model Evaluation

Suppose a company wants to know churn & non-churn customer. So they can give offers to customers who can leave, so we need to tell customers who can be churn but if we identify a non-churn customer as churn then they will also get the offer means loss.

Suppose we get a confusion matrix like below:

		Predict	
Actual		NC	C
NC		3269	366
C		595	692

So actual churn customer = 595 + 692
Churn by correct prediction = 692

so total $\frac{692}{595+692} \approx 53\%$ churn customer

got predicted correctly, which is risky as we missed so many customers who were churn.

Hence it is very critical that you consider the overall business problem you are trying to solve to decide the metric you want to maximise or minimise.

So we use 2 commonly use metric for this.

1. Sensitivity
2. Specificity

1. Sensitivity:

$$\text{Sensitivity} = \frac{\text{No. of actual yes correctly predict}}{\text{Total No of Actual Yes}}$$

Actual/Predicted	Not churn	Churn
Not churn	3269	366
Churn	595	692

OR we can write :

Actual / Predicted	Not Churn	Churn
Not churn	True Negative	False Positive
Churn	False Negative	True Positive

Churn = 1

Not Churn = 0

True Negative : Actual Not churn predict as Not Churn.

False Positive : Actual Not churn predicted as churn.

Falge Negative : Actual ~~predict~~ churn predicted as ~~0~~ Non Churn.

True Positive : Actual churn predicted as Churn.

Sensitivity : (TRUE Positive Rate)

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$= \frac{692}{595 + 692} = \frac{692}{1287} \approx 53.7\%$$

Even though our accuracy (~80.47%) but sesitivity turned out to be quite low. (53.7%)

Specificity

$$\text{Specificity} = \frac{\text{No. of actual Nos Correctly Predicted}}{\text{Total No. of Actual Nos.}}$$

$$= \frac{TN}{TN + FP} = \frac{3269}{3269 + 366}$$

ROC Curve

True Positive Rate (TPR)

No. of positive ~~data~~ ~~by~~ correctly predicted by total no. of positives.

$$TPR = \frac{\text{True Positive}}{\text{Total no. of actual positive}}$$

$$= \frac{TP}{TP+FN} = \text{Sensitivity}.$$

False Positive Rate (FPR)

No. of false positive predicted divided by total negative.

$$FPR = \frac{\text{False Positive}}{\text{Total actual Negative}}$$

$$= \frac{FP}{TN+FP}$$

$$= 1 - \text{Specificity}$$

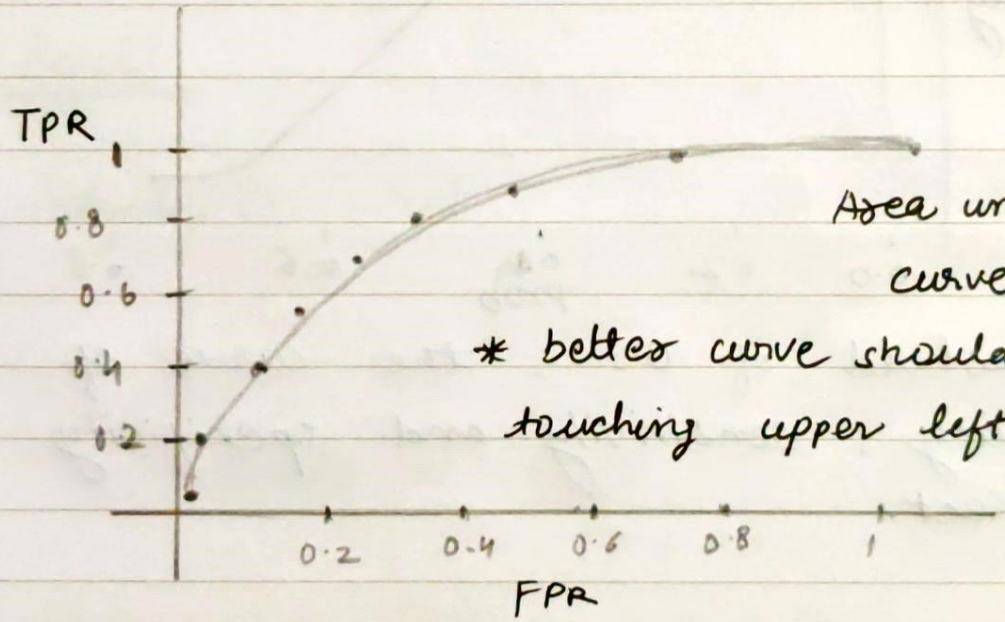
So basically to make a good model you want to maximise the TPR and minimise the FPR.

Steps For ROC Curve

1. We take multiple cutoff to predict churn and non churn.

2. With predicted value, we calculate TPR and FPR.

3. We make a graph with TPR & FPR



Area under the curve – AUC

* better curve should be touching upper left corner.

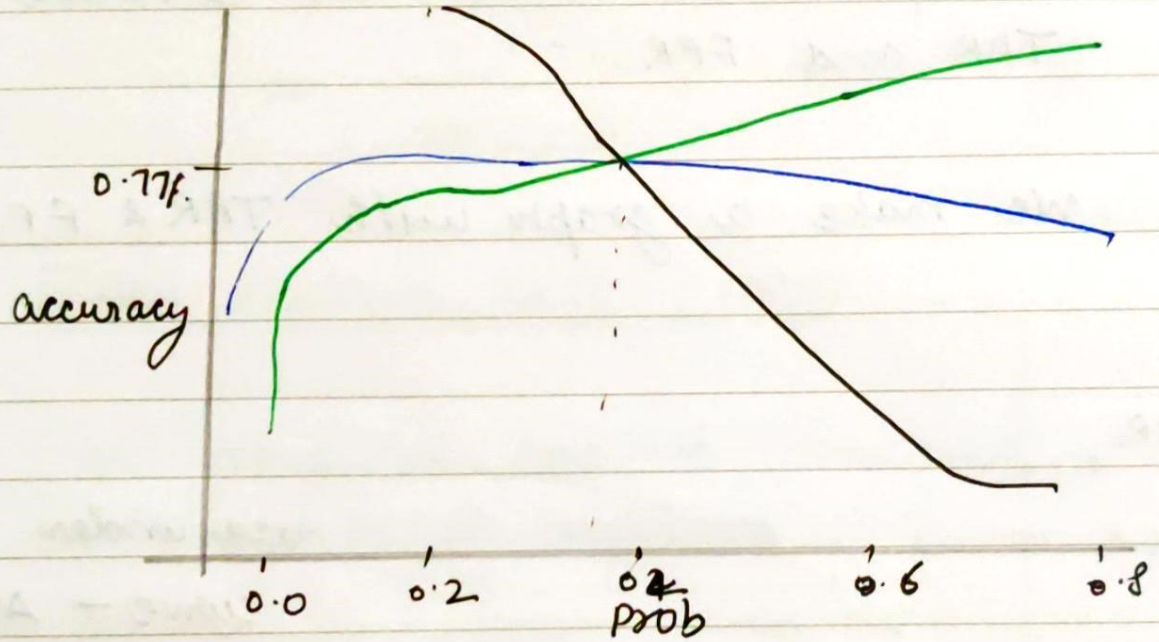
So higher the area under the curve of an ROC curve, the better is your model.

Finding the Optimal Threshold

- We make prediction on all diff cut-offs.

We calculate accuracy, ~~spesit~~ sensitivity specificity.

We make a graph with all three.



At threshold of 0.3, the curve of accuracy, sensitivity and specificity intersect.

Precision & Recall

Precision

Actual / Predicted	No	Yes
No	TN	FP
Yes	FN	TP

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision is probability that a predicted 'Yes' is actually a 'Yes'.

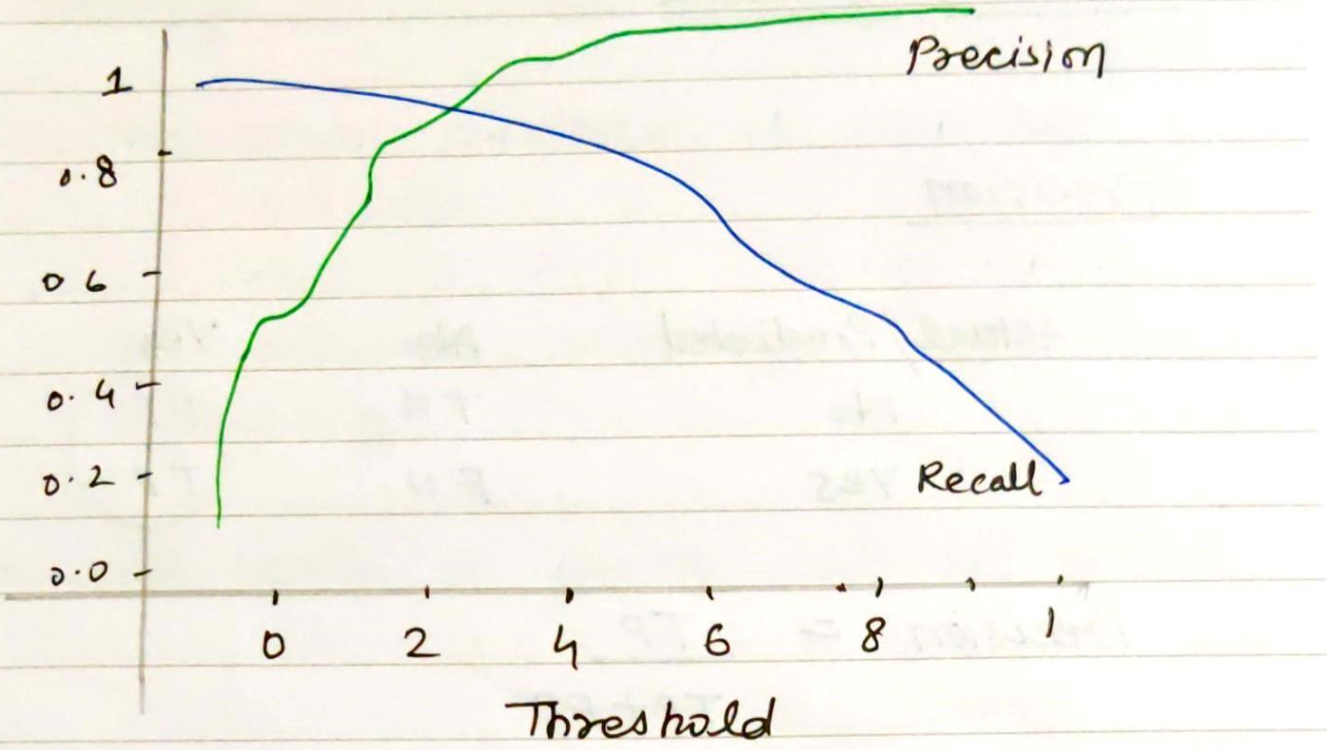
Recall

Actual / Predicted	No	Yes
No	TN	FP
Yes	FN	TP

$$\text{Recall} = \frac{TP}{TP + FN}$$

Probability that an Actual 'Yes' case is predicted correctly.

Same as sensitivity.



F-1 Score

When you want to see precision and recall together.

$$F = 2 \times \frac{\text{Precision} \times \text{recall}}{\text{Precision} + \text{recall}}$$

Industry Application

There are 2 types of logistic regression:

1. Binary logit: involve 2 level of dependent variable.

2. Multinomial logit: involve more than 2 level of dependent variable.

Sample Selection

Three types of error you should lookout for below errors.

1. Cyclic or seasonal fluctuations in the business that need to be taken care of while building sample.

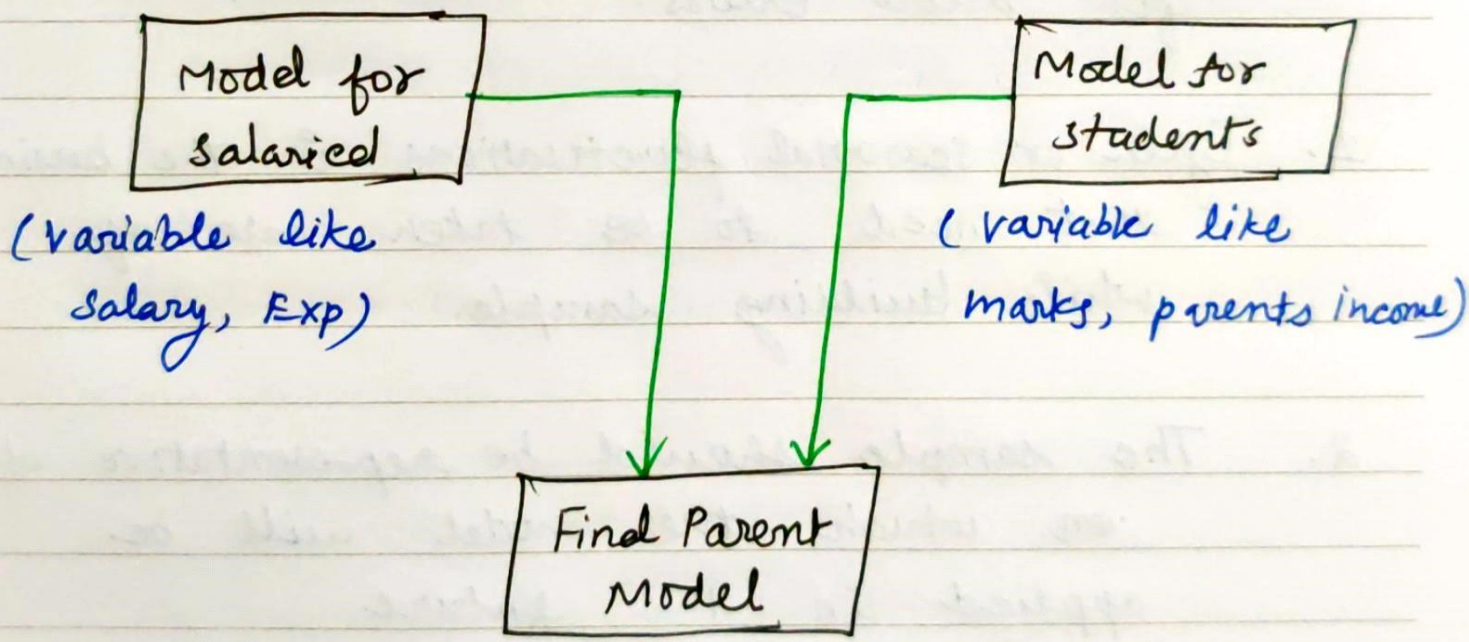
2. The sample should be representative of on which the model will be applied in the future.

3. Rare event should balanced before modelling.

Nuances of Logistic Regression: Segmentation.

Some times a single model will be not be that accurate, so you make diff model for diff type of data/objective. and then you combine them in one model to achive higher accuracy.

Ex:- ICICI want to give credit card. ealier they use to give only salaried person but now they are giving to students too. But samples and variable will be diff for them.



Data Transformation 1

Advantage of dumy variable: it make the model stable.

Disadvantage of dummy variable: If you change continuous variable to dummies all the data will be compressed into very few categories and that might result in data clumpsing.

Variable Transformation

WOE (Weight of Evidence) - Usually in risk analysis we use this type of variable transformation.

Ex:- Suppose you have data of credit card utilization which is continous but you make it several bin of that.

Credit Card Utilization	No. of goods	No. of bad	WOE	IV
Missing	12367	4797	-0.37	0.05
0%-20%	2369	127	1.61	0.07
20%-25%	6395	118	2.67	0.38
25%-45%	3356	259	0.88	0.7
⋮				
>95%			-1.04	
Total	41536	11102	X	0.77

$$WOE = \ln\left(\frac{\text{good in bucket}}{\text{Total good}}\right) - \ln\left(\frac{\text{bad in bucket}}{\text{Total bad}}\right)$$

or

$$WOE = \ln\left(\frac{\% \text{ of good}}{\% \text{ of bad}}\right)$$

You can make a graph with these value.

It is important, that WOE value should follow an increasing / decreasing trend across bin.

If trend is not monotonic then you can combine or compress the buckets / bins. (Coarse bucket)

Advantage:

1. It reflect group identity, means it captures the general trend of distribution of good and bad customers.

2. Model becomes more stable because changes in the cont. variables will not impact the input much.

Disadvantage:

You may end up doing score clumping.

Information Value (IV)

$$IV = WOE * \left[\left(\frac{\text{Good in bucket}}{\text{Total good}}\right) - \left(\frac{\text{Bad in bucket}}{\text{Total Bad}}\right)\right]$$

OR

IV = WOE * (% of good in bucket − % of bad in bucket)

It is an important indicator of predicted power.

It tells how we should do bening.

IV value should be high.

Commonly Faced Challanges

Low event rate: some time some events are less in data. Like in case of fraud detection. less fraud will be there. Then we can use biased data to solve this.

Missing Values: Some values are not missing at random. Suppose for one custome utilization is not there. Then may be customer was not worthy of ~~dat~~ card. hence we will not replace null with other value.

Model Evaluation (Second look)

Measures are:

- Discriminatory Power
- Accuracy
- Stability

Discriminatory Power (Gini)

Gini = Area Under ROC Curve

Model Validation - Stability

In sample validation
Out of time validation (some old data)
K-cross validation (K fold)

Stability

Performance stability
Variable stability