Logistic Regression2

09 February 2022 20:46

Logistic Regression

EDA :- Find variables that have some relation with 'Churn'

Remove variables that are having multicollenearity ✓

→ Missing values

→ outlier (input variables)

Observation

Imbalanced dataset 1 - 14.5%

0 - 85.5%

Blanced : 1 : 30% ; 0 : 70 %

Yes No

future

Imbalanced dataset

Imbalanced learing

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots \theta_n x_n = \log\left(\frac{p}{1-p}\right)$$

$$\text{Odds ratio} = (e^{\theta})$$

Target (heart Risk) ← age, BMI, HDL, LDL, Smoker, Active

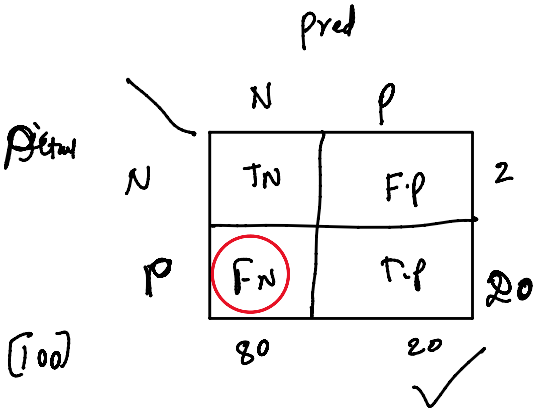
θ_1 θ_2 θ_3 θ_4 θ_5 θ_6 ↑

'Y' ↓ | Odds ratio (e^{θ}) 1.5 2.2 0.7 ↑ 1.8 5.6 ✓ 0.06↓

↑

pred

CM

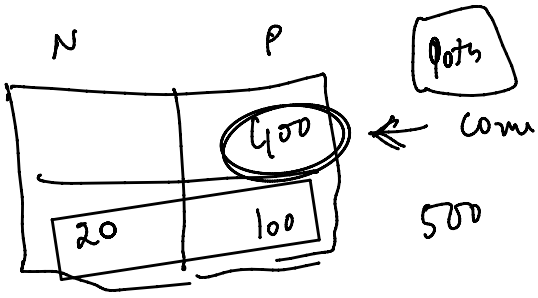


Precision $\frac{20}{22}$ → FP moderately risky

degrading Cancer

Rec $= \frac{20}{100} ? = 0.2$

FN is Not acceptable



Recall

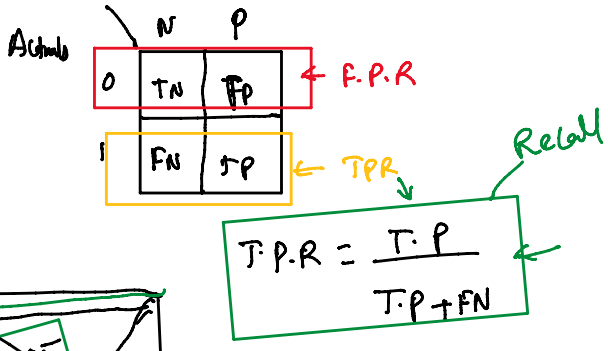
com with cost

$$\text{Accuracy} = \frac{T.P + T.N}{T.P + TN + FP + FN}$$

Precision and Recall F-1 Score $= \left(\frac{2 \cdot P \cdot R}{P+R}\right)$

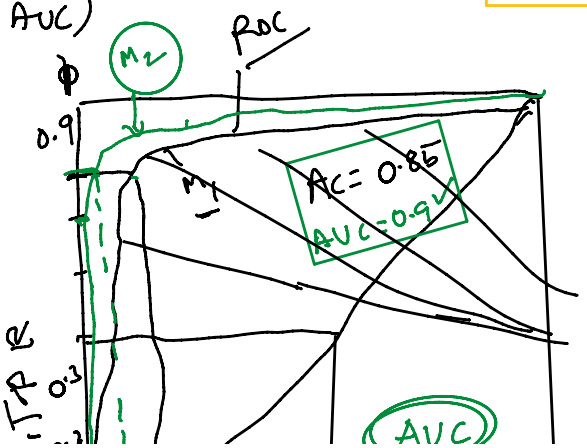
(1) Why 0.5 cut off ?

(2) How do we compare models?



ROC Curve (ROC-AUC)

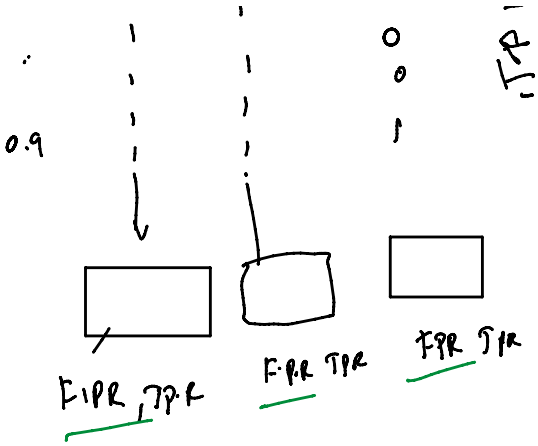
y	Cutoff 0.1 Pred	0.2 Pred	...	0.9 Pred	y
Prob					
0.1	0	0		0	
0.2	1	0		0	
.	1	1		0	
.	1	1		0	
.	1	1		0	
.	1	1		0	

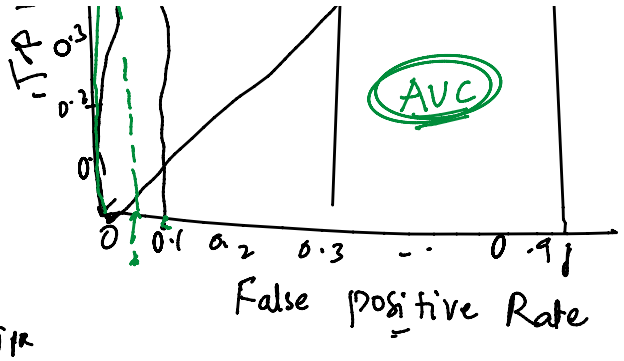


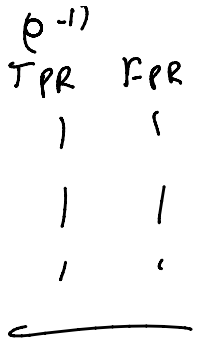
$$F.P.R = \frac{F.P}{F.P + TN}$$

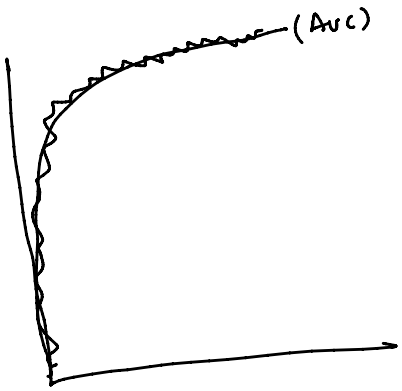
(0,1)

TPR FPR

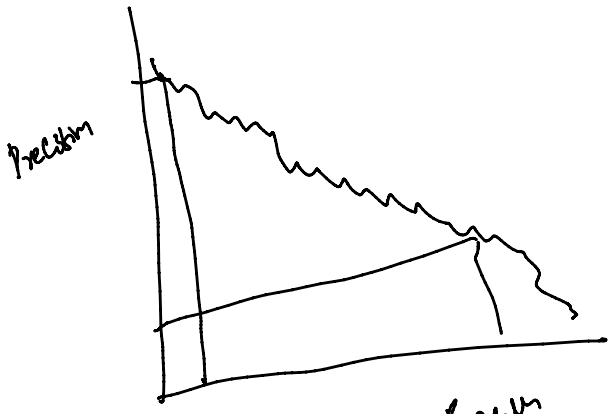


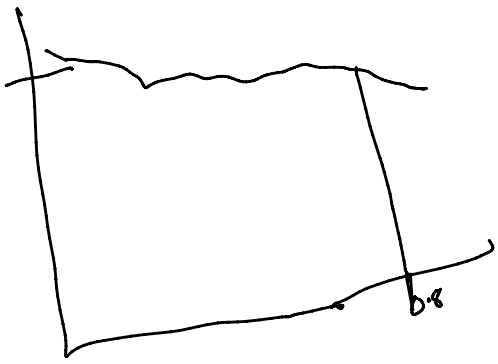




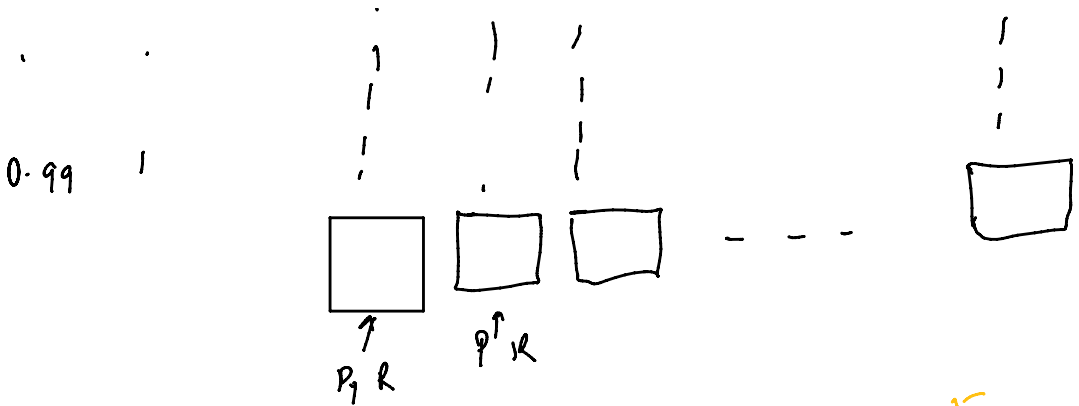


Precision vs Recall graph (PR Curve)





^ y Prob	Actuals y	Cutoff 0.05 PC	0.10 PC	0.15 PC	- - - - -	0.95
0.1	0	1	0	0		0
0.11	0	1	1	0		0
0.12	1	1	1	0		0
	0	1	1	1		0
	1	1	1	1		0
		1	1	1		1
		1	1	1		1



Cutoff | Precision | Recall

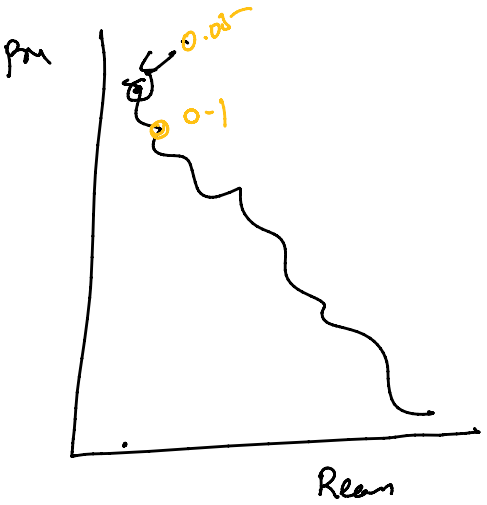
0.05 | 0.9 | 0.1

0.1

0.15

⋮

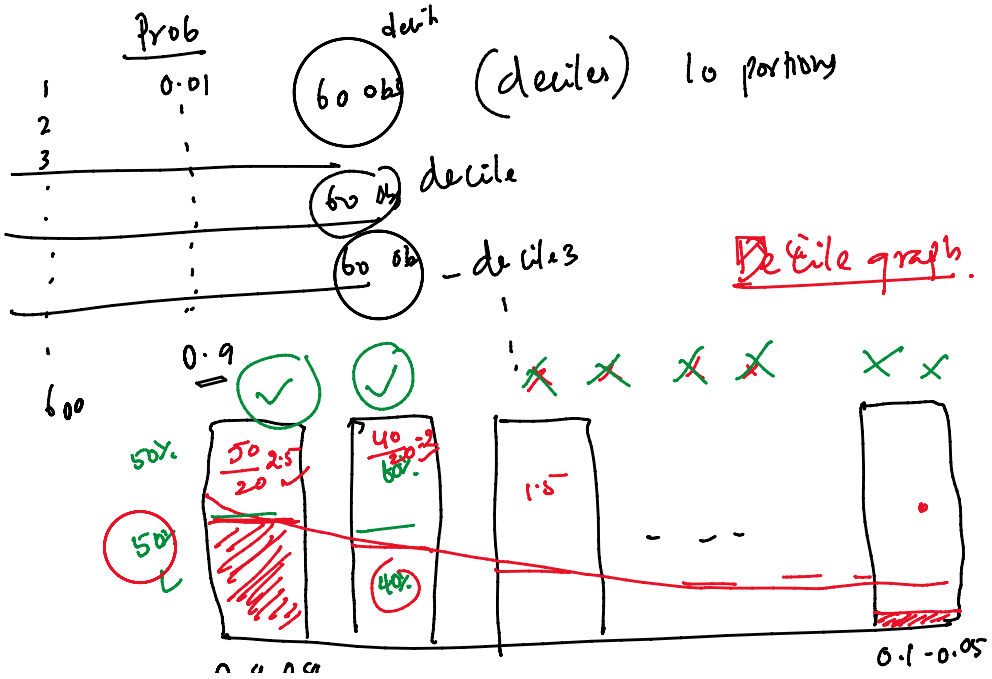
0.95



Analysis / Charts

(Benchmark)

20% Risky No. of call > 5



lift.

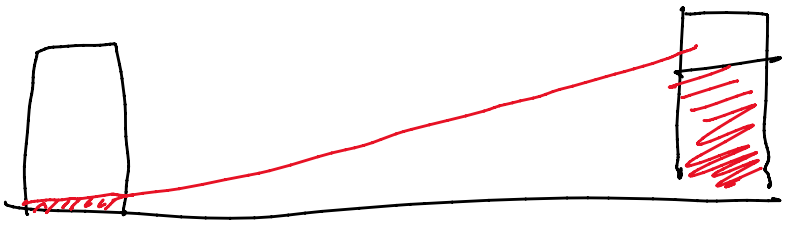
lift.

Randomly (100) → (20%)

Top decile (high prob) model

decile 100 → (50%) / 20% 2.5

decile 2 → 40% / 20% 2



	M	T	W	T	F	S	S
Pred	N	N	R	R	N	N	N
	N	N	R	N	R	N	N

P.