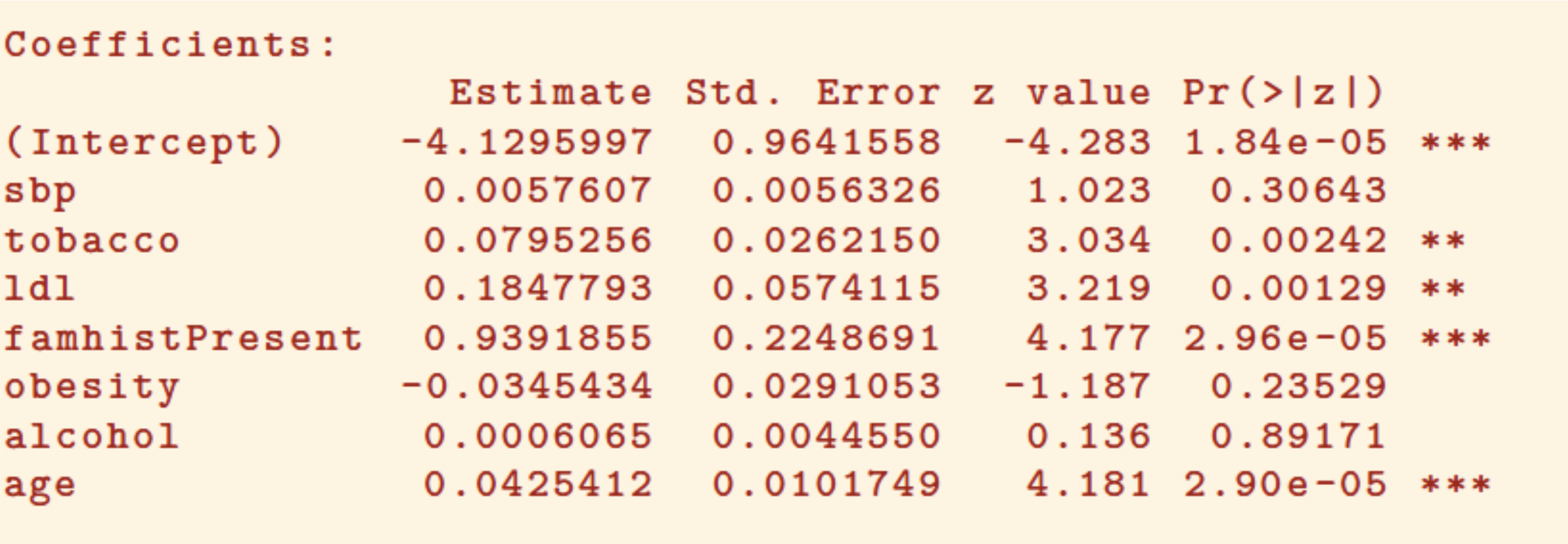
Lecture 9 – Logistic Regression (Part 2)

Instructor: Hamed Hasheminia

Agenda

- Unbalanced observations and Logistic Regression
- FP/FN/TP/TN/FPR/TPR/FNR
- The effect of changing Threshold
- ROC curves
- Area Under Curve
- How to compare classification algorithms

Quiz - How do we interpret this?



```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp             0.0057607  0.0056326   1.023  0.30643
tobacco         0.0795256  0.0262150   3.034  0.00242 **
ldl             0.1847793  0.0574115   3.219  0.00129 **
famhistPresent  0.9391855  0.2248691   4.177 2.96e-05 ***
obesity        -0.0345434  0.0291053  -1.187  0.23529
alcohol         0.0006065  0.0044550   0.136  0.89171
age             0.0425412  0.0101749   4.181 2.90e-05 ***
```

Case-Control Sampling and Logistic Regression

- In South African data, there are 160 cases, 302 controls — $\tilde{\pi} = 0.35$ are cases. Yet the prevalence of MI in this region is $\pi = 0.05$.
- With case-control samples, we can estimate the regression parameters β_j accurately (if our model is correct); the constant term β_0 is incorrect.
- We can correct the estimated intercept by a simple transformation

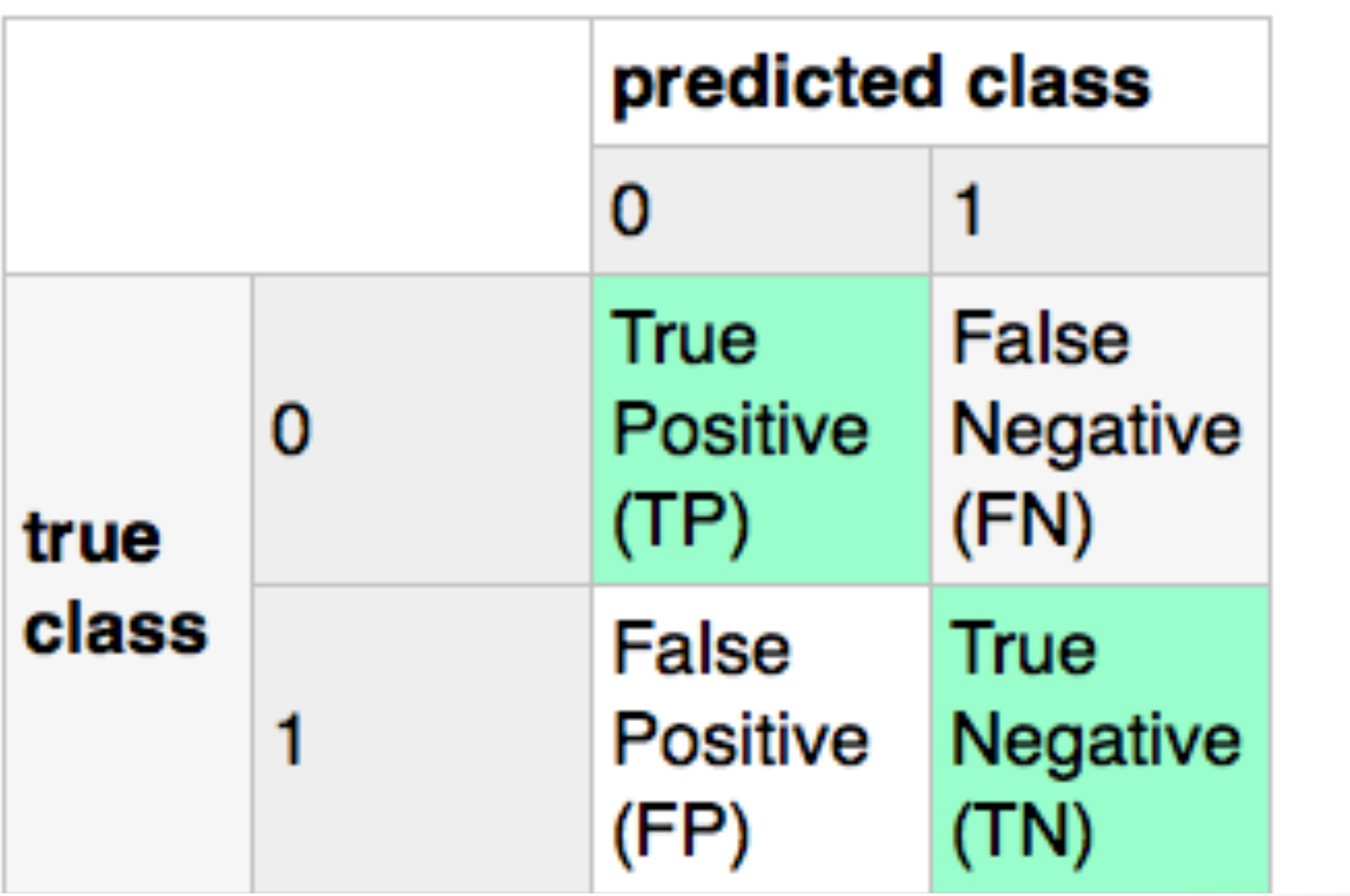
$$\hat{\beta}_0^* = \hat{\beta}_0 + \log \frac{\pi}{1-\pi} - \log \frac{\tilde{\pi}}{1-\tilde{\pi}}$$

Credit Data

		Predicted Default Status		
		No	Yes	Total
True Default Status	No	9644	23	9667
	Yes	252	81	333
	Total	9896	104	10000

- (23 + 252)/10000 errors – a 2.75% misclassification rate! Is it good?
- Some caveats
  - This is training error, and we may be over-fitting. But this is not a big concern in this case since n = 10000 and we only used 4 parameters.
  - If we classify everything as No – then we make only 3.33% error.
  - Of the true No's, we make 23/9667 = 0.2% errors; of the true Yes's, we make 252/333 = 75.7% errors!

TP/FP/FN/TN



		predicted class	
		0	1
true class	0	True Positive (TP)	False Negative (FN)
	1	False Positive (FP)	True Negative (TN)

Error / Accuracy / False Positive Rate / True Positive Rate / Precision / Recall

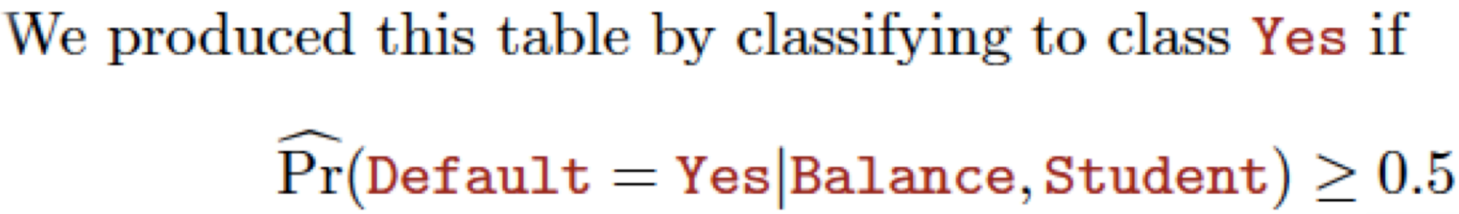
- ERR = (FP + FN) / (FP + FN + TP + TN)
- ACC = (TP + TN)/ (FP + FN + TP + TN) = 1 – ERR

- False Positive Rate (FPR)
  - FPR = FP / (Total Negatives) = FP / (FP + TN)
- True Positive Rate (TPR) – Also Called Recall
  - TPR = TP / (Total Positives) = TP / (TP + FN)
- False Negative Rate
  - FNR = FN / (Total Positives) = FN / (TP + FN) = 1 - TPR
- Precision (PRE)
  - PRE = TP / (TP + FP)

Credit Data - continues

		Predicted Default Status		
		No	***Yes***	***Total***
True Default Status	***No***	***9644***	***23***	***9667***
	Yes	***252***	***81***	***333***
	Total	***9896***	***104***	***10000***

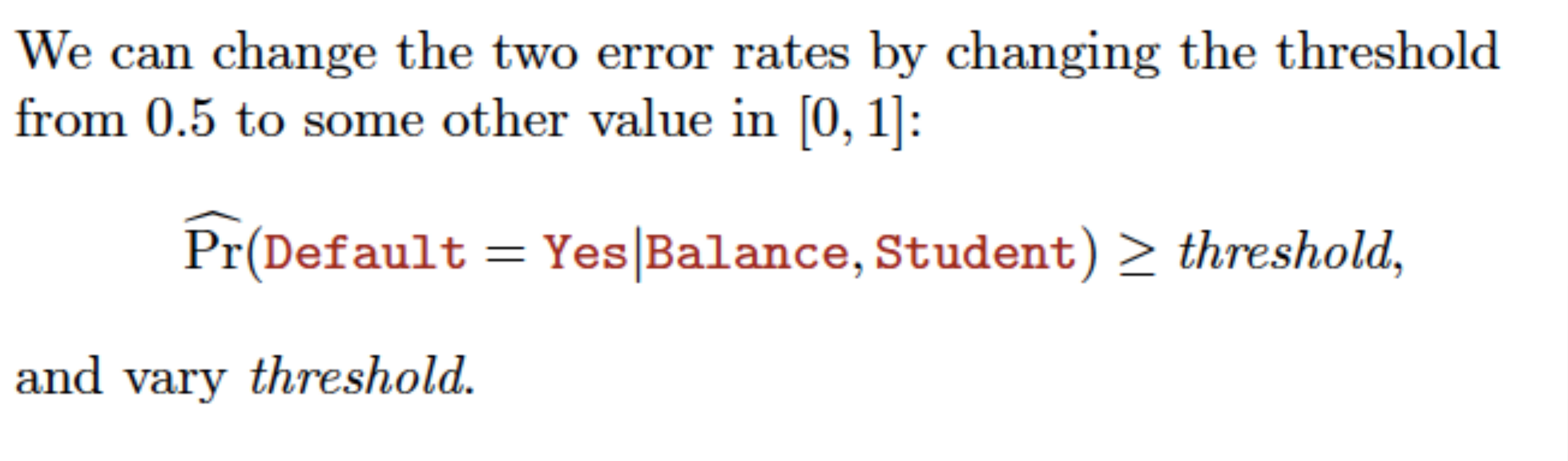
(From those who did default 75.6% mistakenly predicted that they would not default)
(From Those who did not default, only 0.24% were mistakenly predicted to default.)



We produced this table by classifying to class Yes if

$$\widehat{\Pr}(\texttt{Default} = \texttt{Yes}|\texttt{Balance}, \texttt{Student}) \geq 0.5$$

Changing Threshold

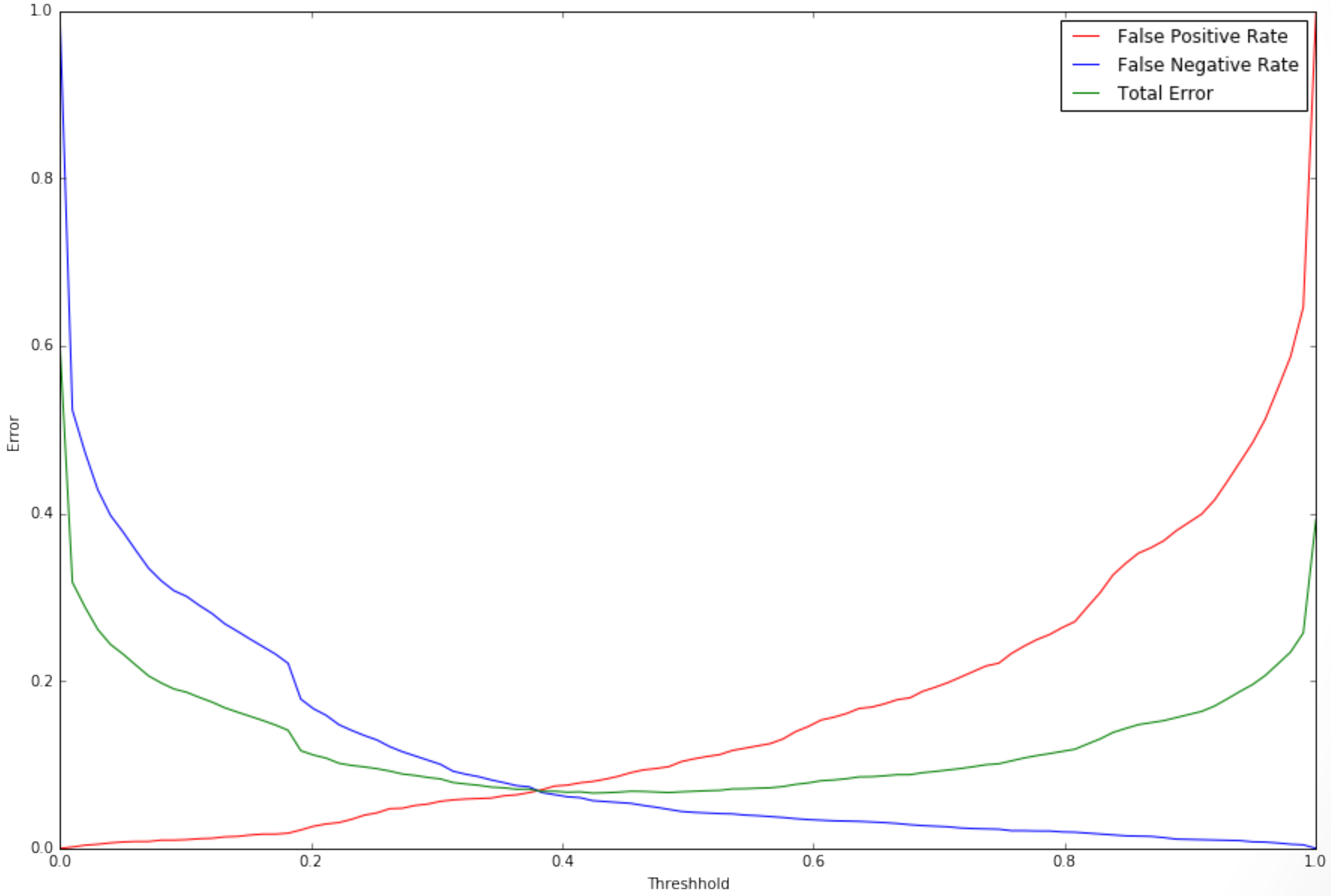


We can change the two error rates by changing the threshold from 0.5 to some other value in $[0, 1]$:

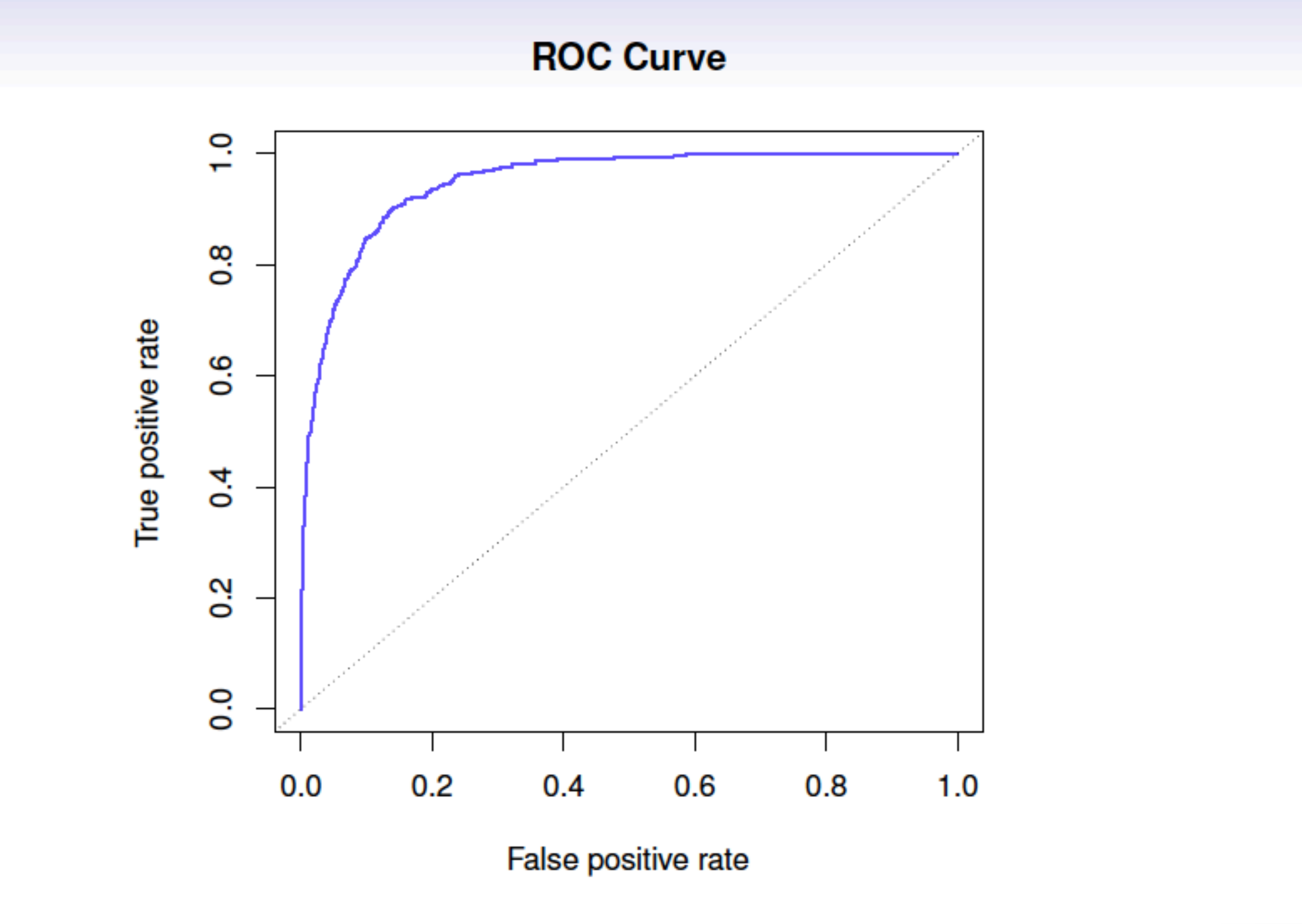
$$\widehat{\Pr}(\texttt{Default} = \texttt{Yes}|\texttt{Balance}, \texttt{Student}) \geq \mathit{threshold},$$

and vary *threshold*.

Varying the Threshold (Spam/Ham Example)



ROC Curve

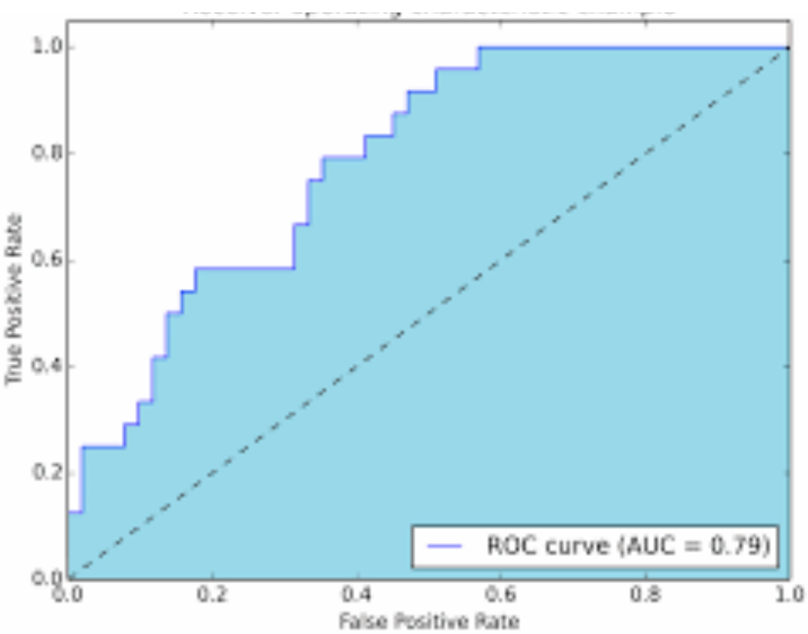

ROC Curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate

Let's explore ROC

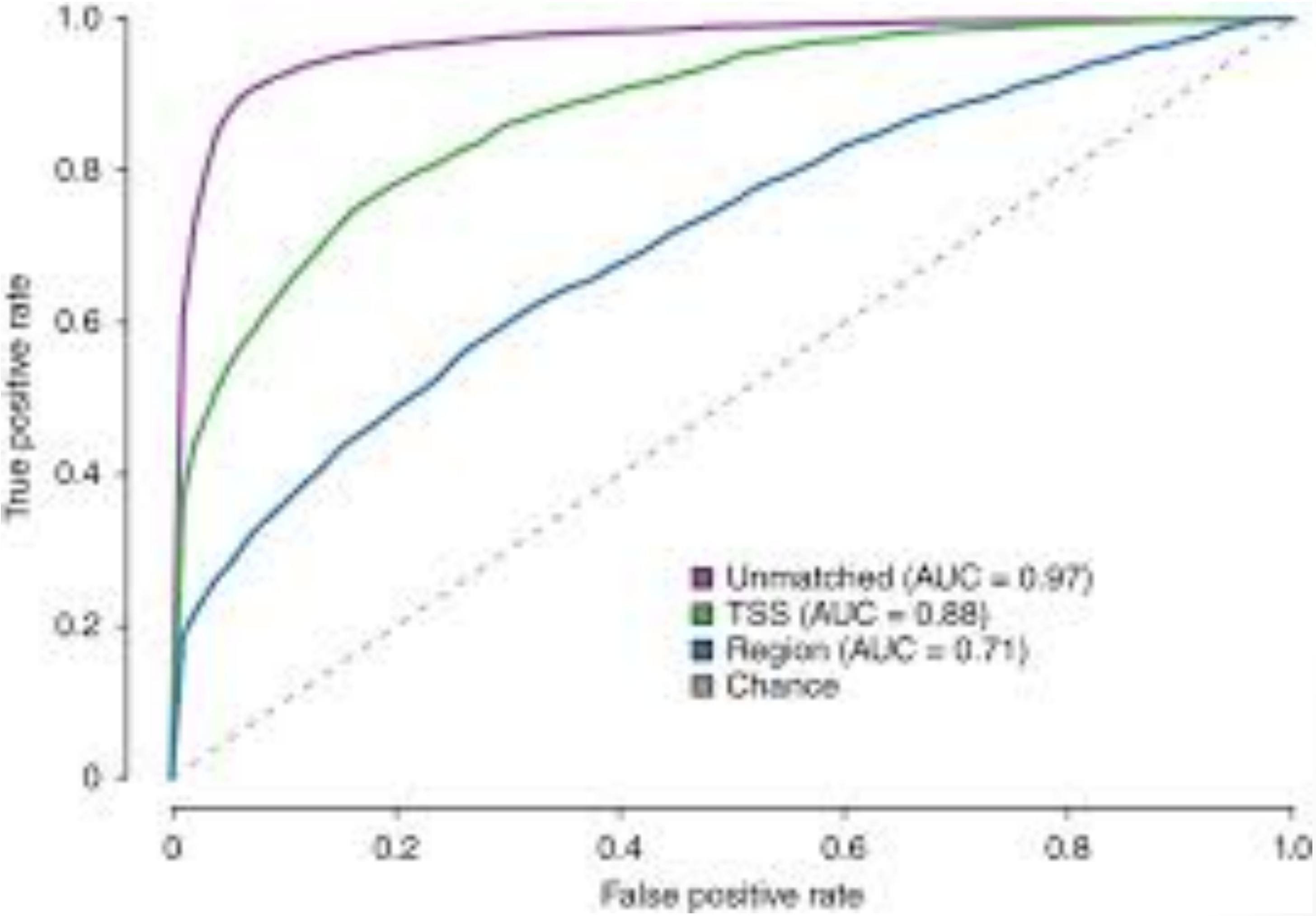
- http://www.navan.name/roc/
- After Instructor's instructions please team-up with with another student and explain ROC to each other.

Area Under Curve (AUC)

- One of the measures used to evaluate Classification Algorithms is Area Under Cure (AUC) of ROC.
- Usually the model which has the largest AUC is considered the best classification model.
- AUC is a number between (0.5 and 1). Why couldn't it be less than 0.5?



Which Classification Model is better?



Summary

- How to adjust Logistic Regression coefficients for unbalanced data
- FP/FN/TP/TN/FPR/TPR
- How changing Threshold can change FPR/TPR/FNR/TNR
- What ROC curves mean
- How to calculate Area Under Curve
- How to compare classification algorithms using AUC