Profit Curves and Imbalanced Classes

Darren Reger Lecture for Galvanize DSI

I have 5 models, which is the best?

- Model 1 with Accuracy 0.977
- Model 2 with Accuracy 0.02
- Model 3 with Accuracy 0.98
- Model 4 with Accuracy 0.88
- Model 5 with Accuracy 0.748

What about now?

- Model 1 with Precision 0.44 and Recall 0.6
- Model 2 with Precision 0.02 and Recall 1.0
- Model 3 with Precision 0 and Recall 0
- Model 4 with Precision 0.115 and Recall 0.75
- Model 5 with Precision 0.0672 and Recall 0.9

Does this help?

	Predicted: Yes	Predicted: No
Actual: Yes	12	15
Actual: No	8	965

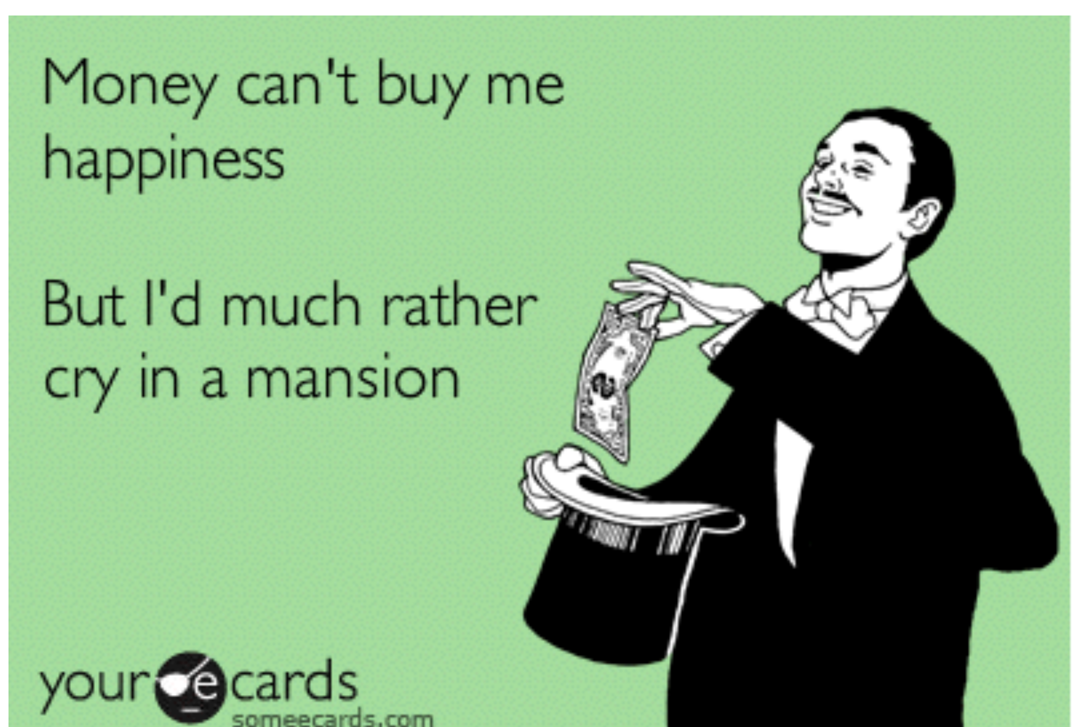
	Predicted: Yes	Predicted: No
Actual: Yes	20	980
Actual: No	0	0

	Predicted: Yes	Predicted: No
Actual: Yes	0	0
Actual: No	20	980

	Predicted: Yes	Predicted: No
Actual: Yes	15	115
Actual: No	5	865

	Predicted: Yes	Predicted: No
Actual: Yes	18	250
Actual: No	2	730

Discussion of Business Applications



Revisiting Confusion Matrix

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \rightarrow \begin{pmatrix} \frac{TP}{TP+FN} & \frac{FP}{FP+TN} \\ \frac{FN}{TP+FN} & \frac{TN}{FP+TN} \end{pmatrix}$$

$$\begin{pmatrix} \frac{TP}{TP+FN} P_{+} & \frac{FP}{FP+TN} P_{-} \\ \frac{FN}{TP+FN} P_{+} & \frac{TN}{FP+TN} P_{-} \end{pmatrix}$$

$$\text{Profit Matrix} = \begin{pmatrix} B_{P_+} & C_{P_+} \\ C_{P_-} & B_{P_-} \end{pmatrix}$$

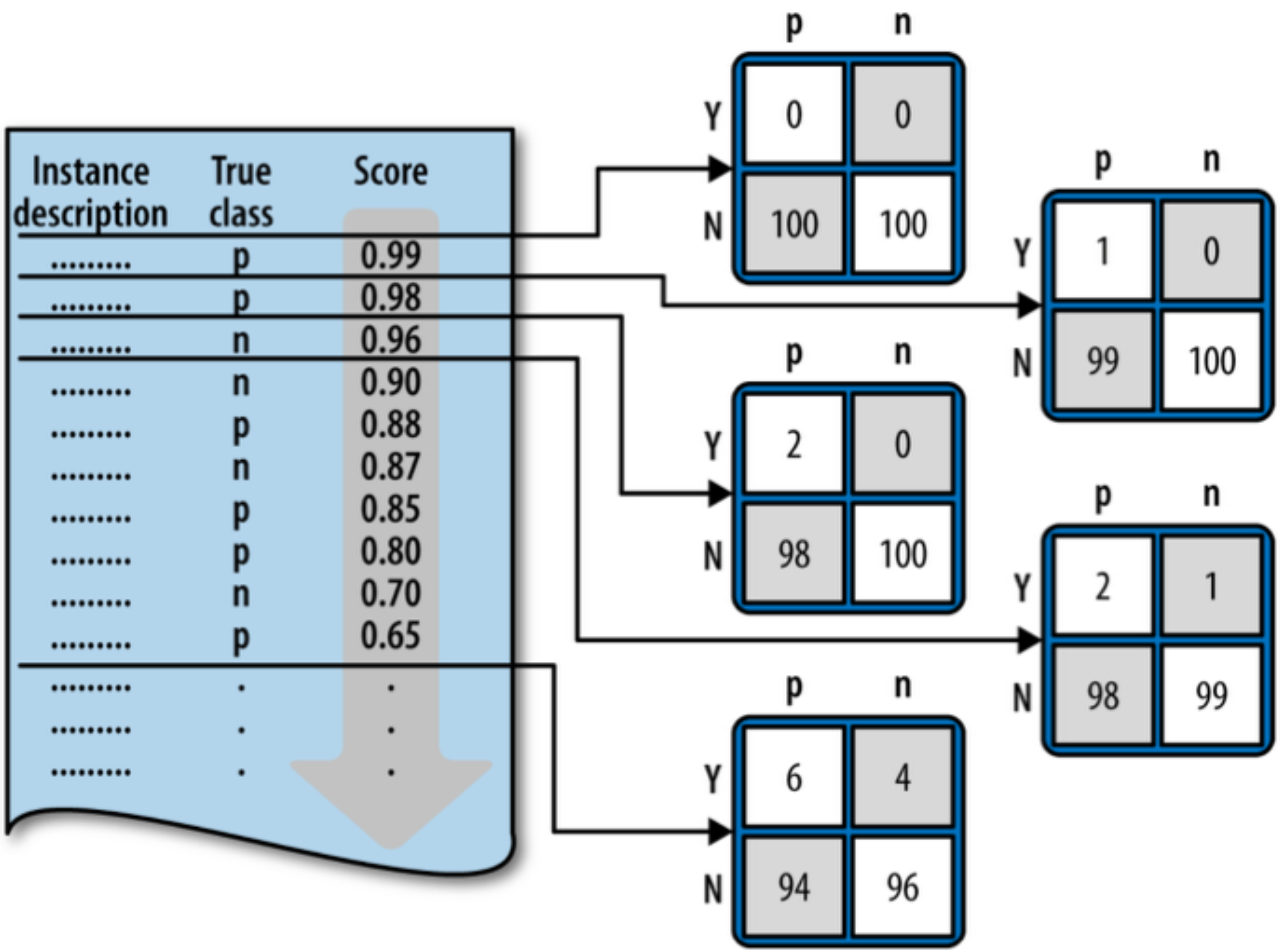
Profit Calculation

Combining information from the Confusion matrix and the Cost-Benefit matrix we can calculate Expected Profit!

		Actual	
		p	n
Predicted	Y	b(Y,p)	c(Y,n)
	N	c(N,p)	b(N,n)

$$
\begin{aligned}
E[Profit] &= P(Y,p)\cdot b(Y,p) + P(Y,n)\cdot c(Y,n) + \\
&\quad P(N,p)\cdot c(N,p) + P(N,n)\cdot b(N,n) \\
&= P(Y|p)\cdot P(p)\cdot b(Y,p) + P(Y|n)\cdot P(n)\cdot c(Y,n) + \\
&\quad P(N|p)\cdot P(p)\cdot c(N,p) + P(N|n)\cdot P(n)\cdot b(N,n) \\
&= P(p)\cdot [P(Y|p)\cdot b(Y,p) + P(N|p)\cdot c(N,p)] + \\
&\quad P(n)\cdot [P(Y|n)\cdot c(Y,n) + P(N|n)\cdot b(N,n)]
\end{aligned}
$$

Let's Make a Curve



Making Profit Curves

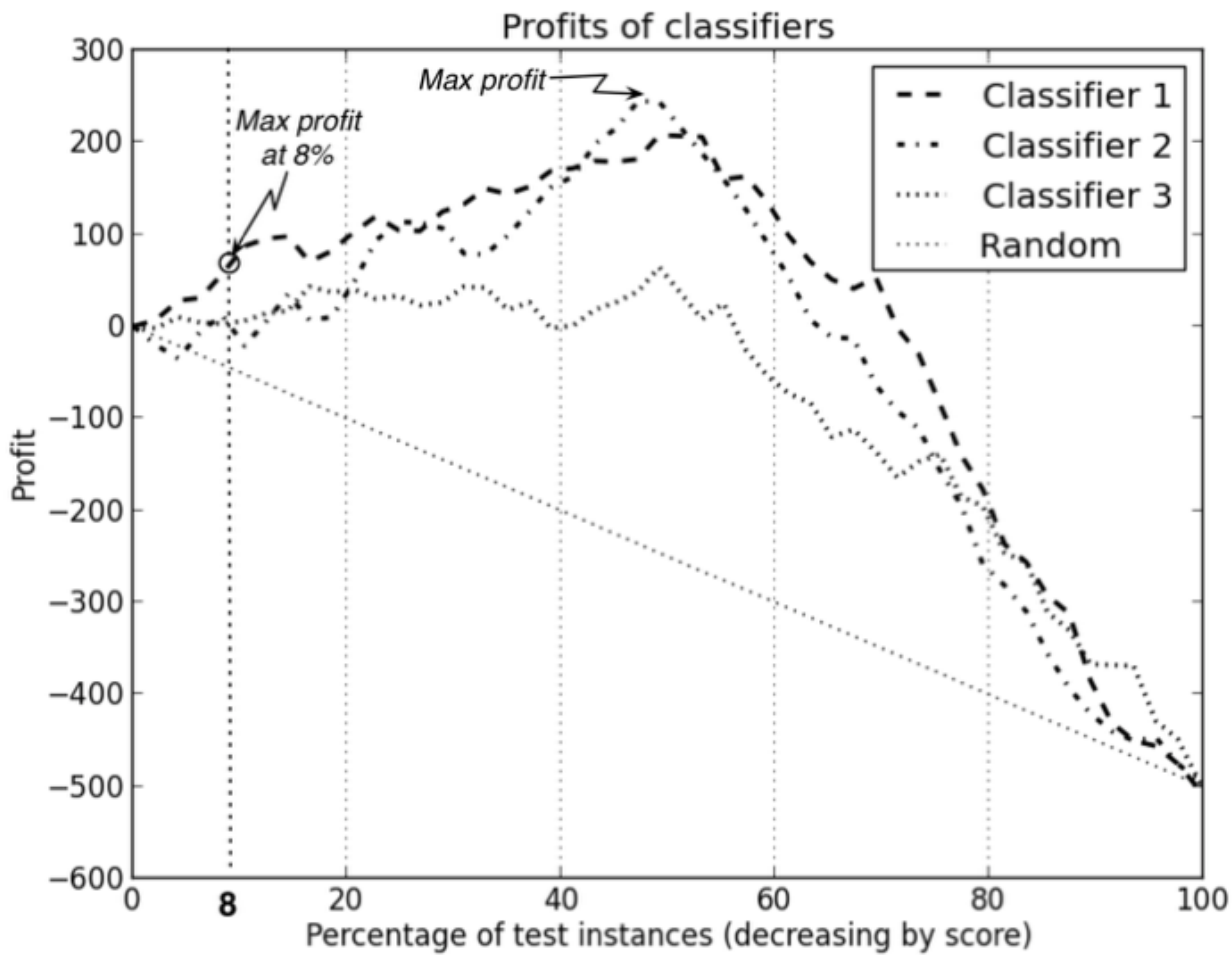
For a given model f, each threshold value T gives a point on the Profit Curve

Model score is the threshold probability classifying $+$ vs $-$

1. Allow T to be the maximum score
2. $TP = 0, FP = 0$
3. Calculate $E[Profit]$
4. For each observation, i:
   - If $\hat{\pi}_i > T \longrightarrow$ increment TP
   - Else $\longrightarrow$ increment FP
5. Add point (% Test Instances predicted Positive, $E[Profit]$) to the Profit Graph

Increment T from max-score to min-score, repeating steps 1-4

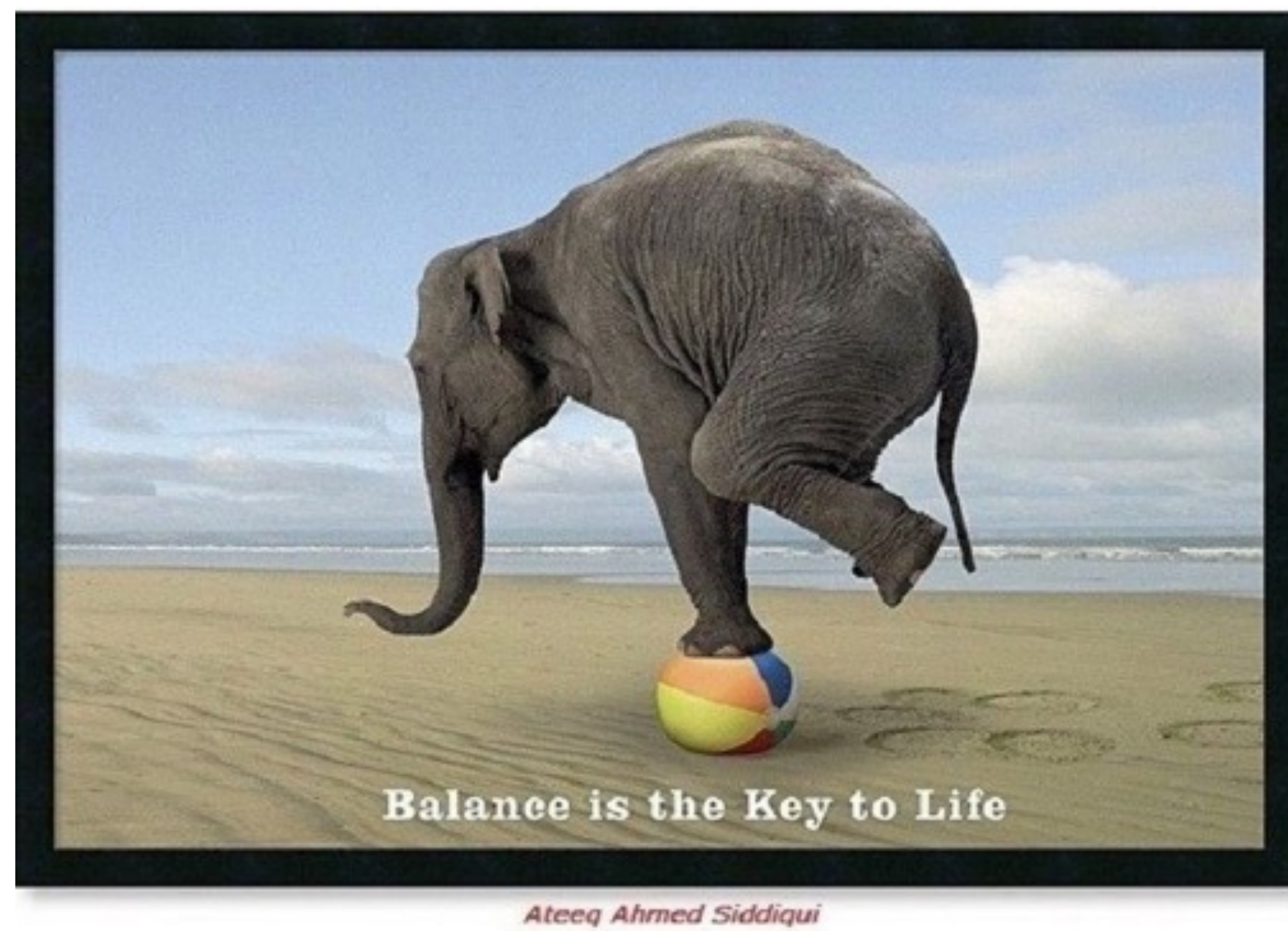
$$$$$$$$$



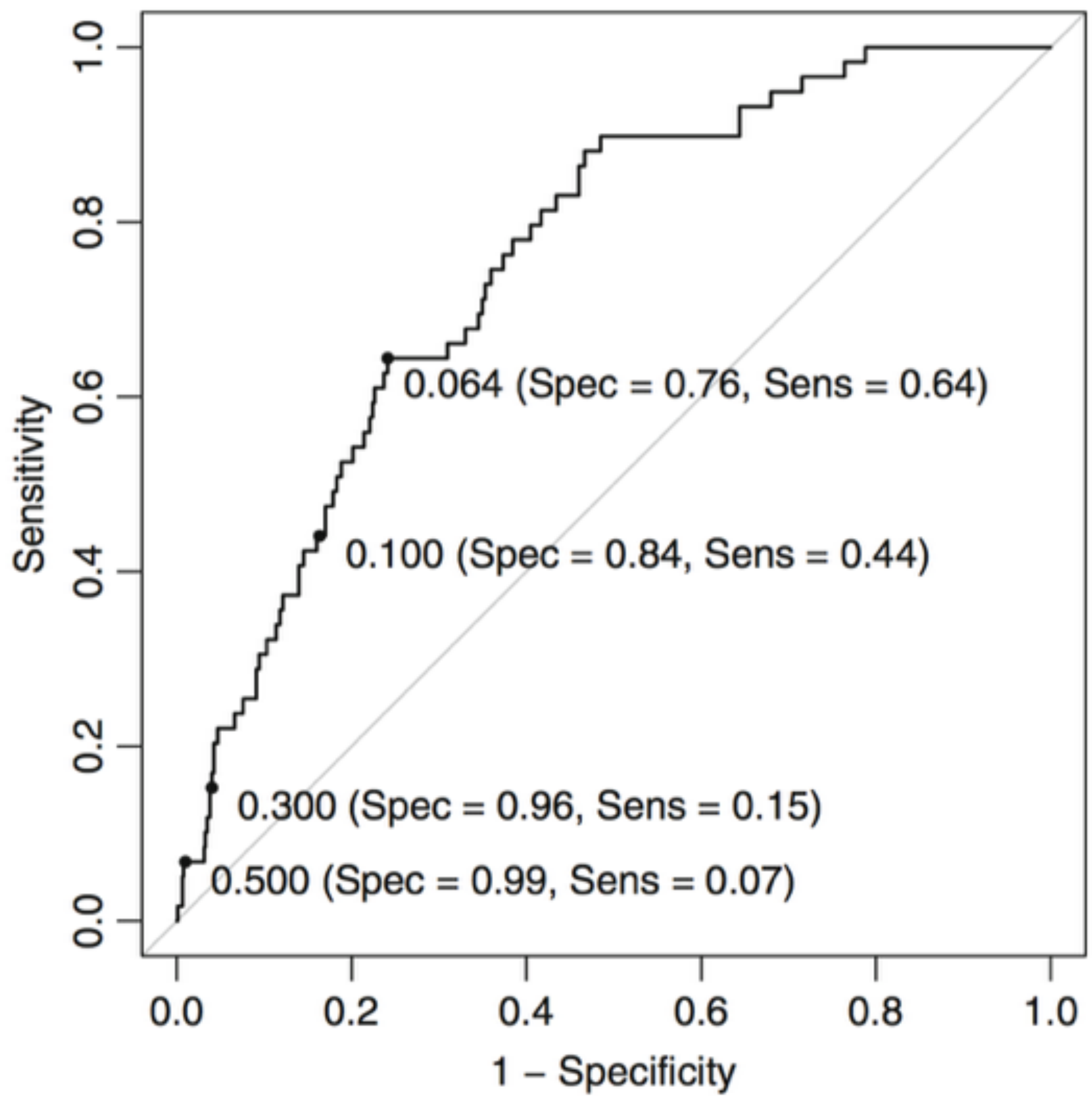
Profits of classifiers
Max profit
Max profit
at 8%
Classifier 1
Classifier 2
Classifier 3
Random
300
200
100
0
−100
−200
−300
−400
−500
−600
Profit
0
8
20
40
60
80
100
Percentage of test instances (decreasing by score)

Imbalanced Data

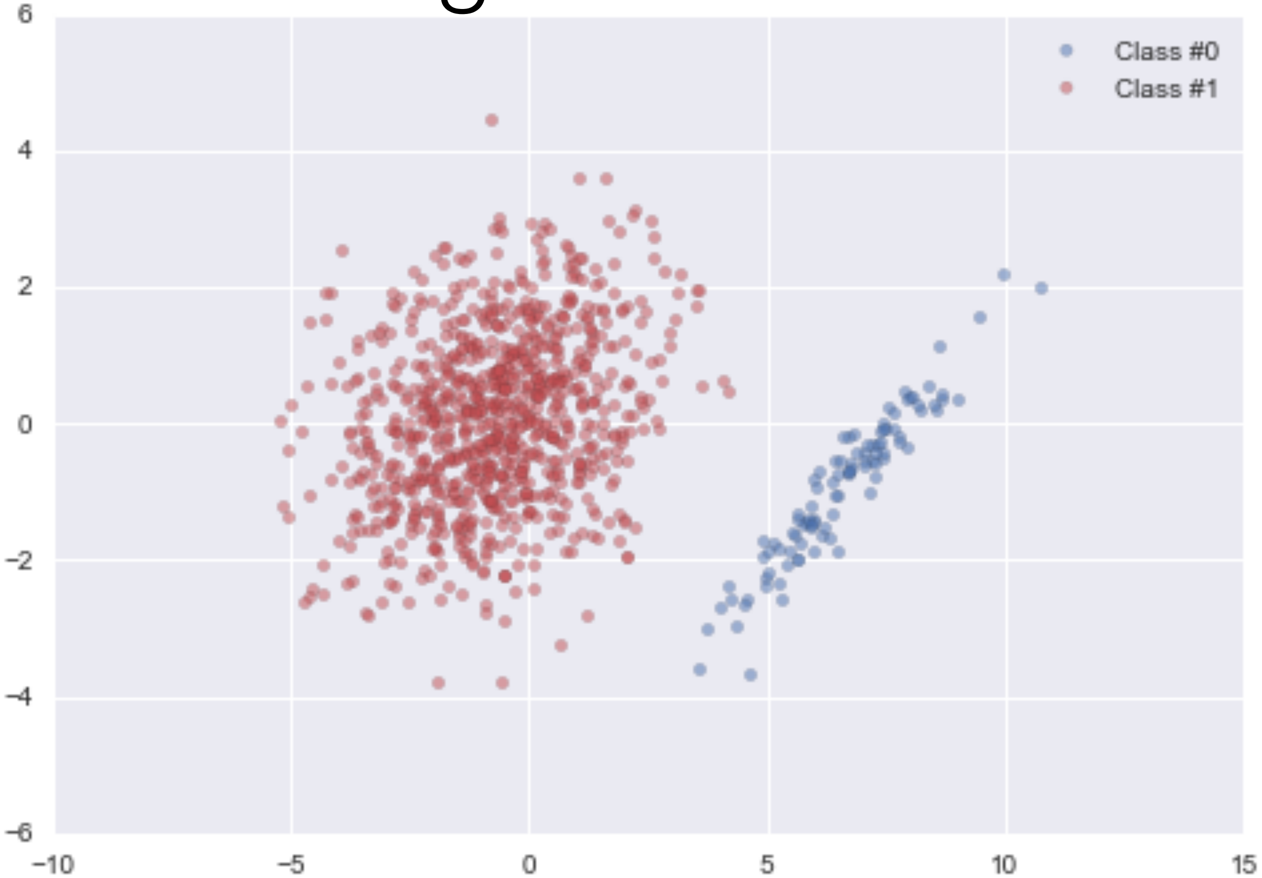
- Where do we see it?
- Why is it bad?
- What can we do?
  - Assign Weights
  - Balancing Classes



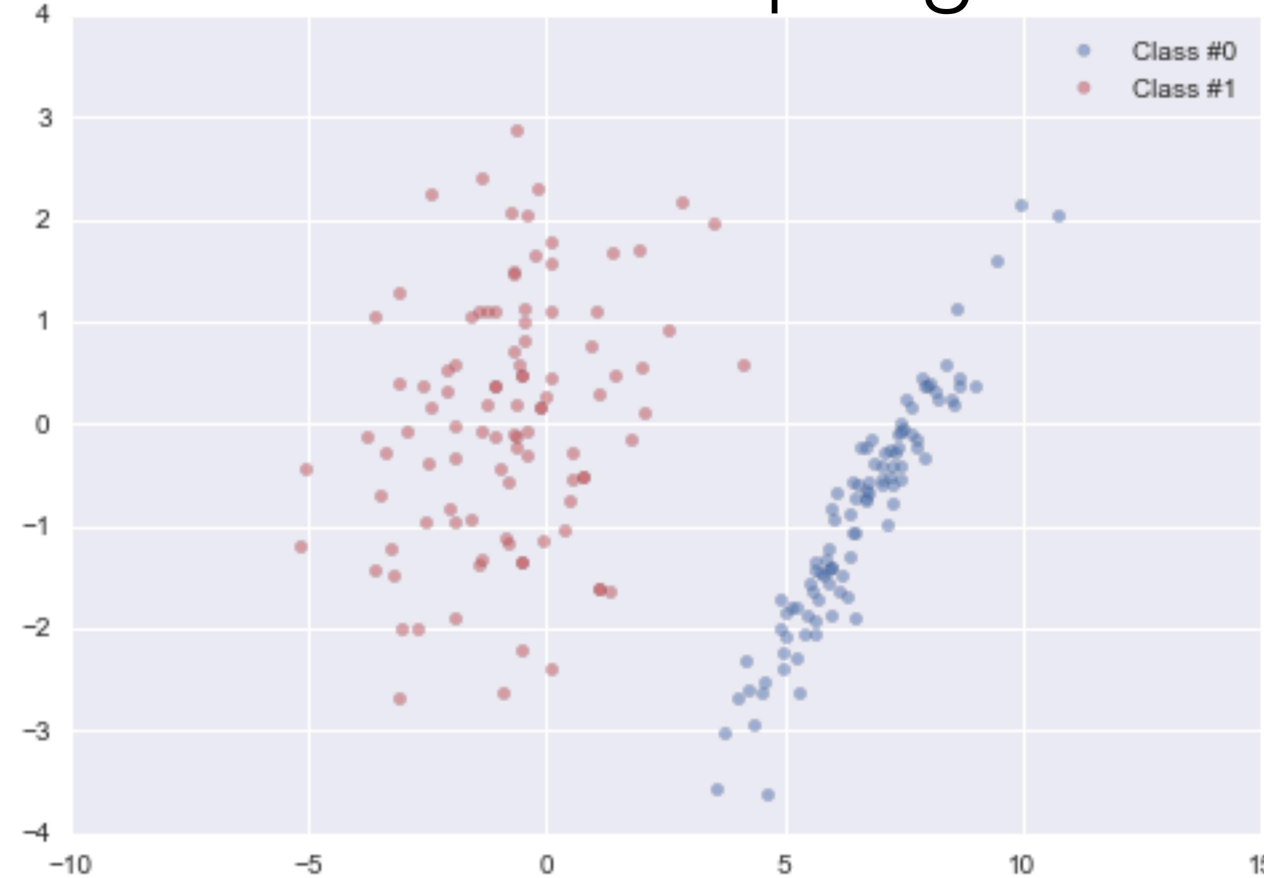
Choosing Cutoff Point



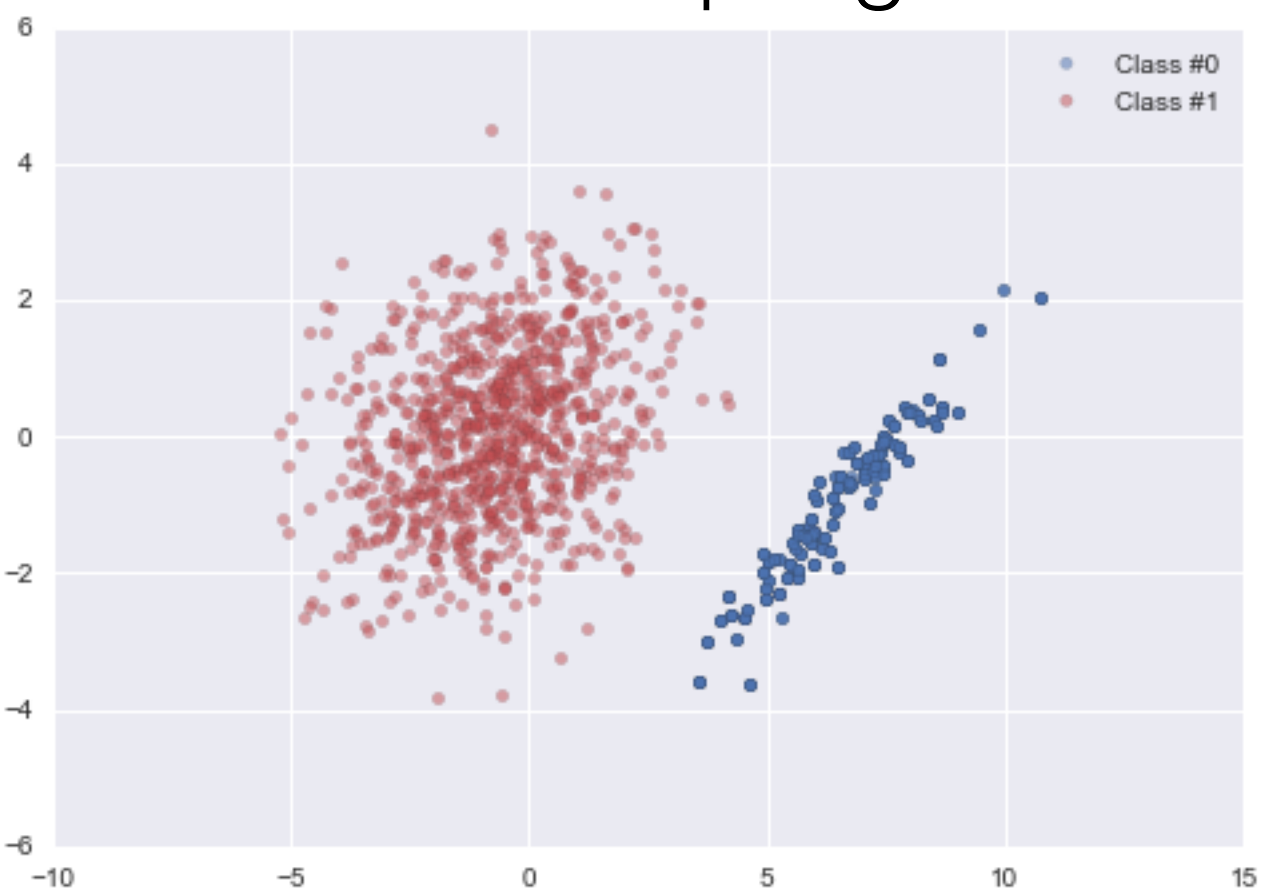
Original Data



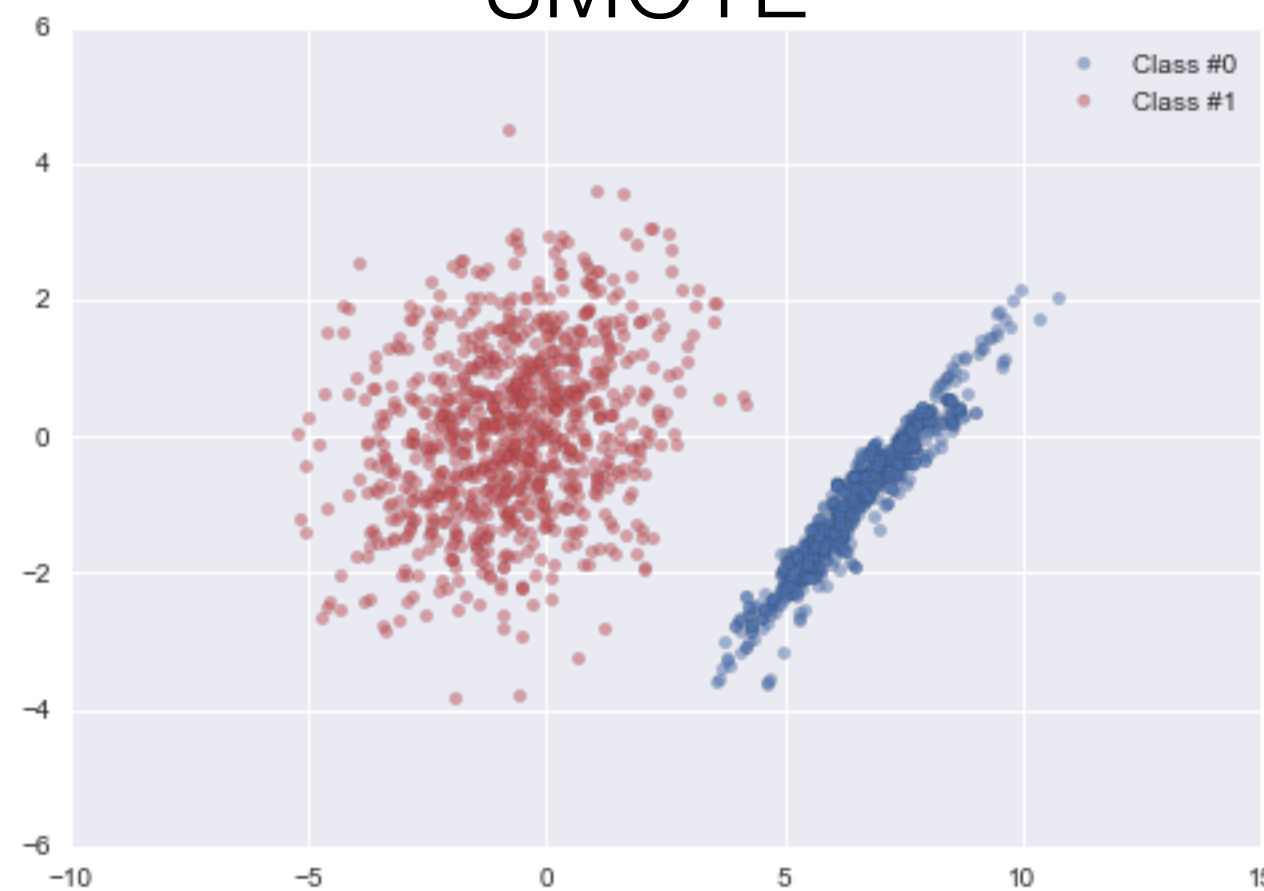
Undersampling



Oversampling

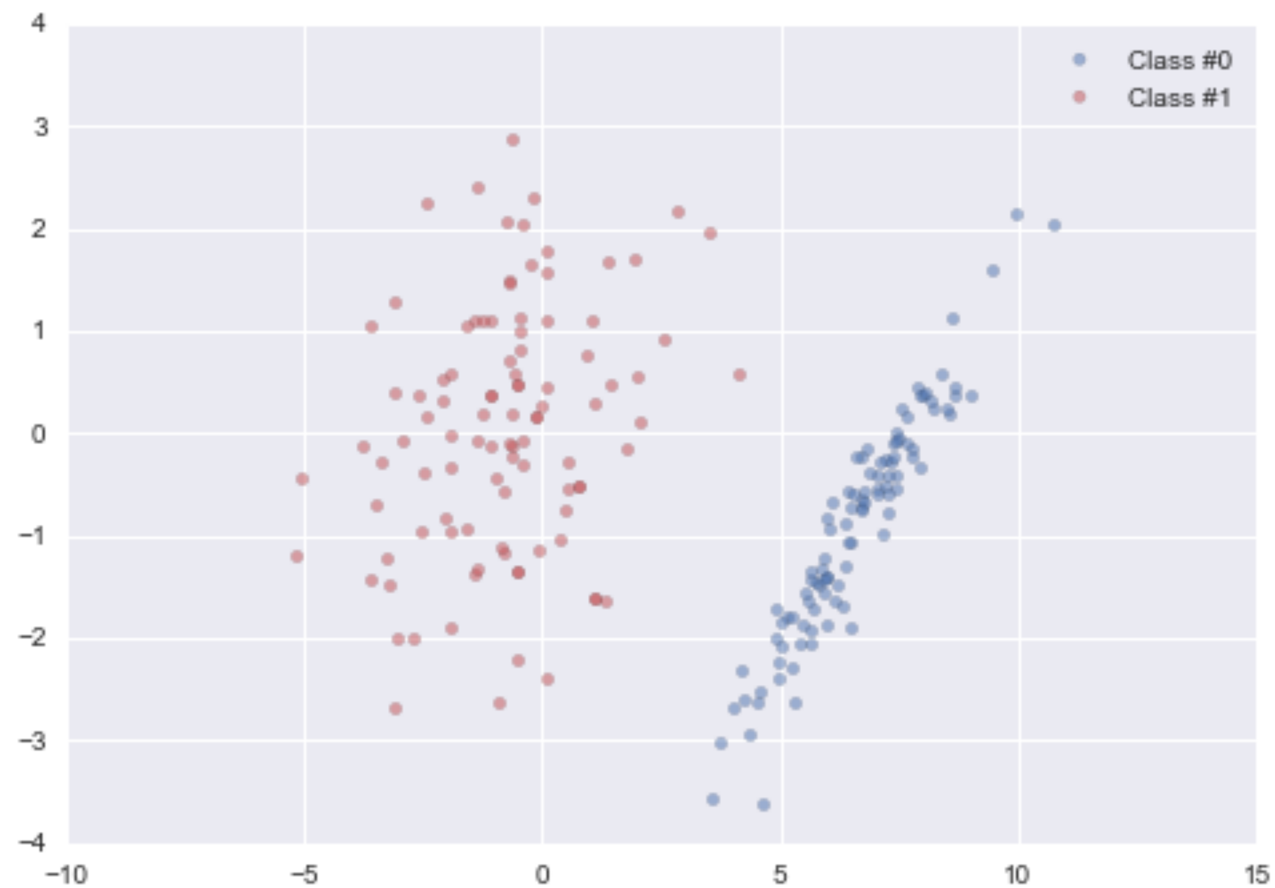


SMOTE



Undersampling

- Randomly discards majority class observations
- Pros: Makes calculations way faster
- Cons: Throwing out data :(

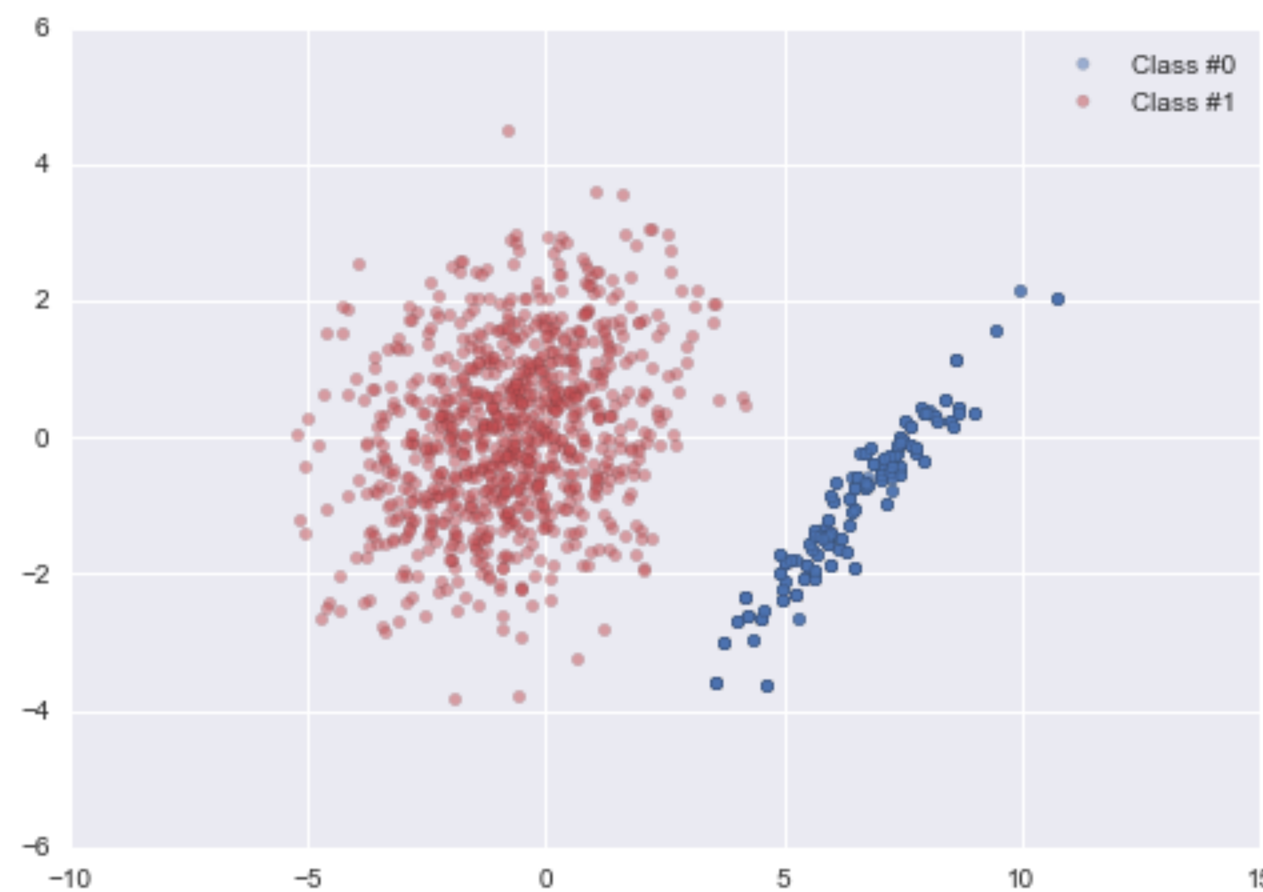


Oversampling

- Replicates minority class observations

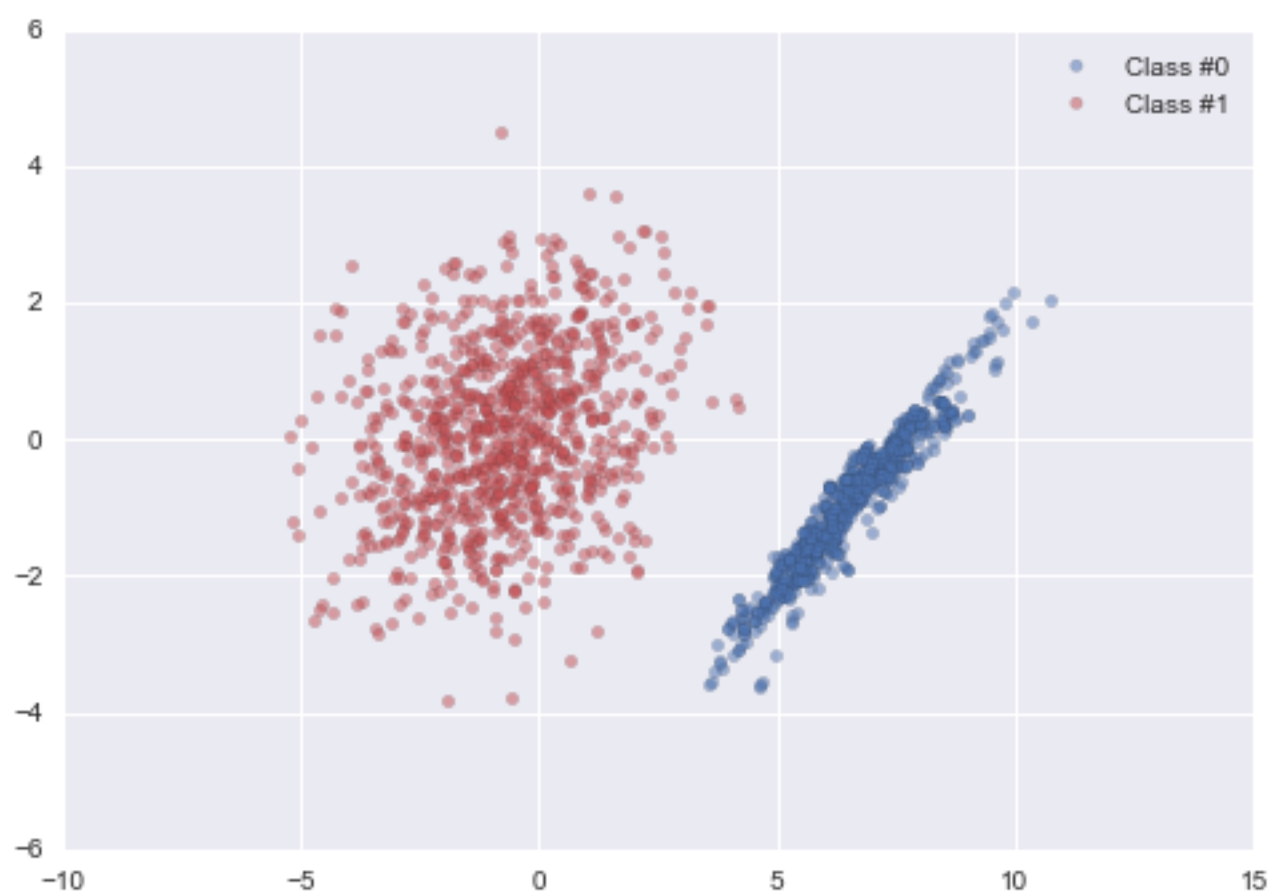
- Pros: Doesn't discard info

- Cons: Overfitting likely :(

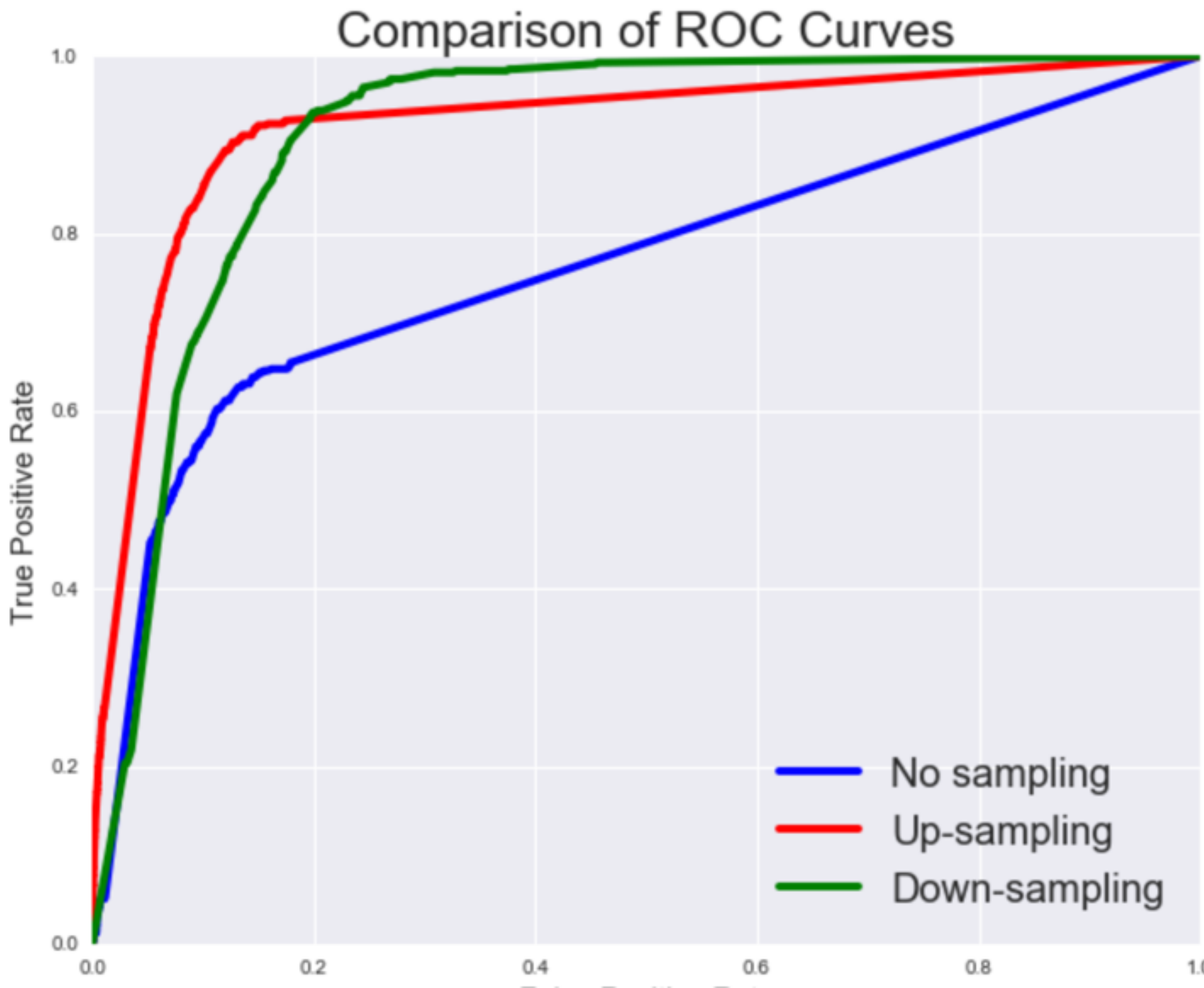


SMOTE

- Generates synthetic minority class observations
  Let's look at the original paper's pseudocode:
  https://www.jair.org/media/953/live-953-2037-jair.pdf



How do we pick?



Changing Cost Function

- Models with explicit cost function can be modified to incorporate classification cost
  - e.g. SVM, logistic $$\frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i \longrightarrow \frac{\|w\|^2}{2} + C^+ \sum_{\{i|yi=+1\}}^{n_+}\xi_i + C^- \sum_{\{j|yj=-1\}}^{n_-}\xi_j$$
- Can affect the optimization
  - ex. cost sensitive logistic regression no longer convex
- Not possible for all models