

**Tera**

# Módulo 4:

## Estatística e Modelagem de Dados

### Aula 23: Aplicando em negócios

### Cost Matrix e Profit Curve





# Instrutora

## Cristiane Rodrigues

- **Bacharel em Matemática – UNESP Rio Claro.**
- **Mestre em Estatística – USP Piracicaba**
- **Experiências Profissionais:**
  - Modelagem de Credito para PF e PJ – Banco Bradesco
  - Experiência com Segmentação e Análise de Series temporais – Atento
  - Consultora Analítica no SAS Institute Brasil
  - Professora do curso SAS Academy for Data Science



# Índice

- Revisão Matriz de Confusão e ROC
- Motivação
- Exemplos
- Valor Esperado - para estruturar o uso da classificação
- Valor Esperado - para estruturar o valor da classificação
- Matriz de Custo
- Lucro Esperado
- Curva de Lucro
- Comparação: ROC X Curva Lucro
- Desafio



# Revisão: Matriz de Confusão

		Classificação Predita		
		0	1	
Classificação Real	0	TN	FP	Actual Negative
	1	FN	TP	Actual Positive
		Predicted Negative	Predicted Positive	

**TN:** True Negative

**FP:** False Positive

**FN:** False Negative

**TP:** True Positive

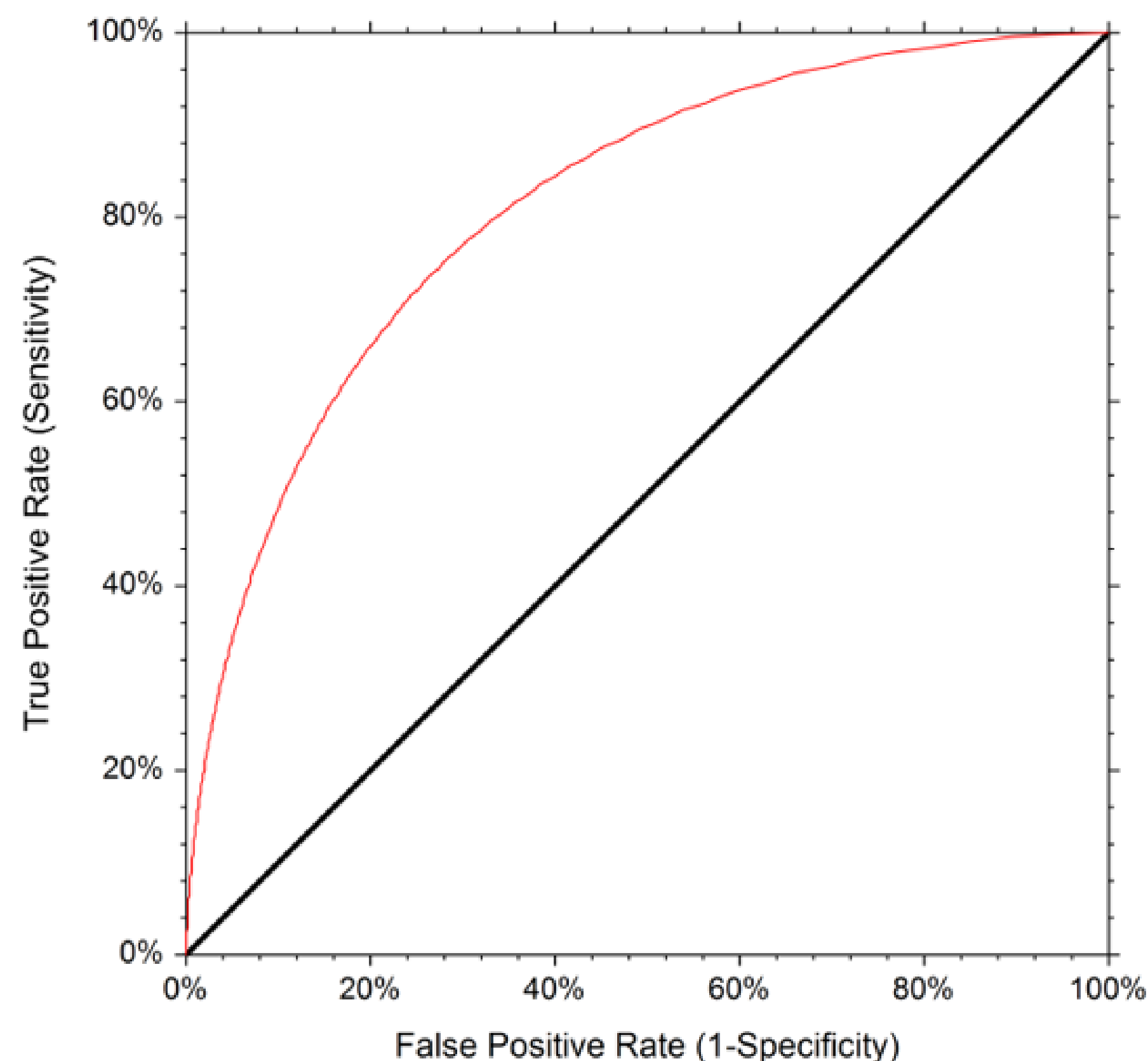
**Métricas** para avaliar a qualidade do ajuste do modelo

- **Missclassification** =  $\frac{FP+FN}{Total\ de\ casos}$  (% de erros)
- **Acurácia** =  $\frac{TN+TP}{Total\ de\ casos}$  (% de acertos)
- **Sensibilidade** =  $\frac{TP}{TP+FN}$
- **Especificidade** =  $\frac{TN}{TN+FP}$

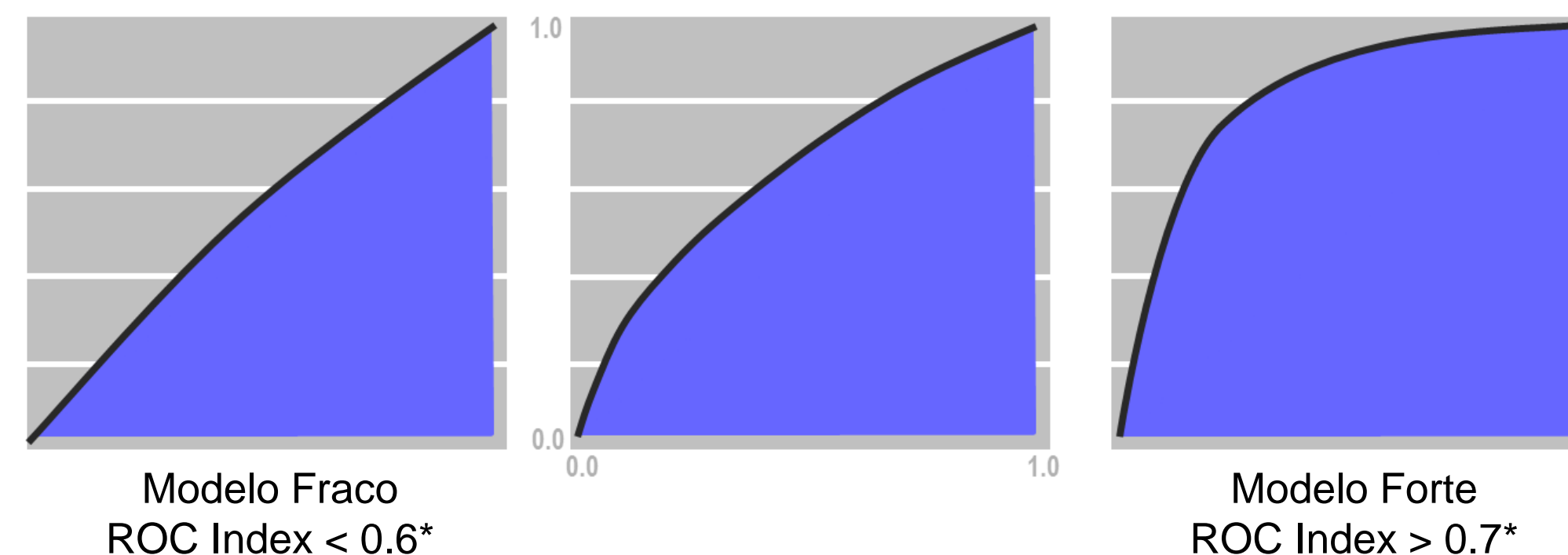


# Revisão: Curva ROC

- A curva ROC, mede, fração a fração, quantos 1's foram capturados (taxa de true positive) versus quantos 0's foram capturados (taxa de false positive).



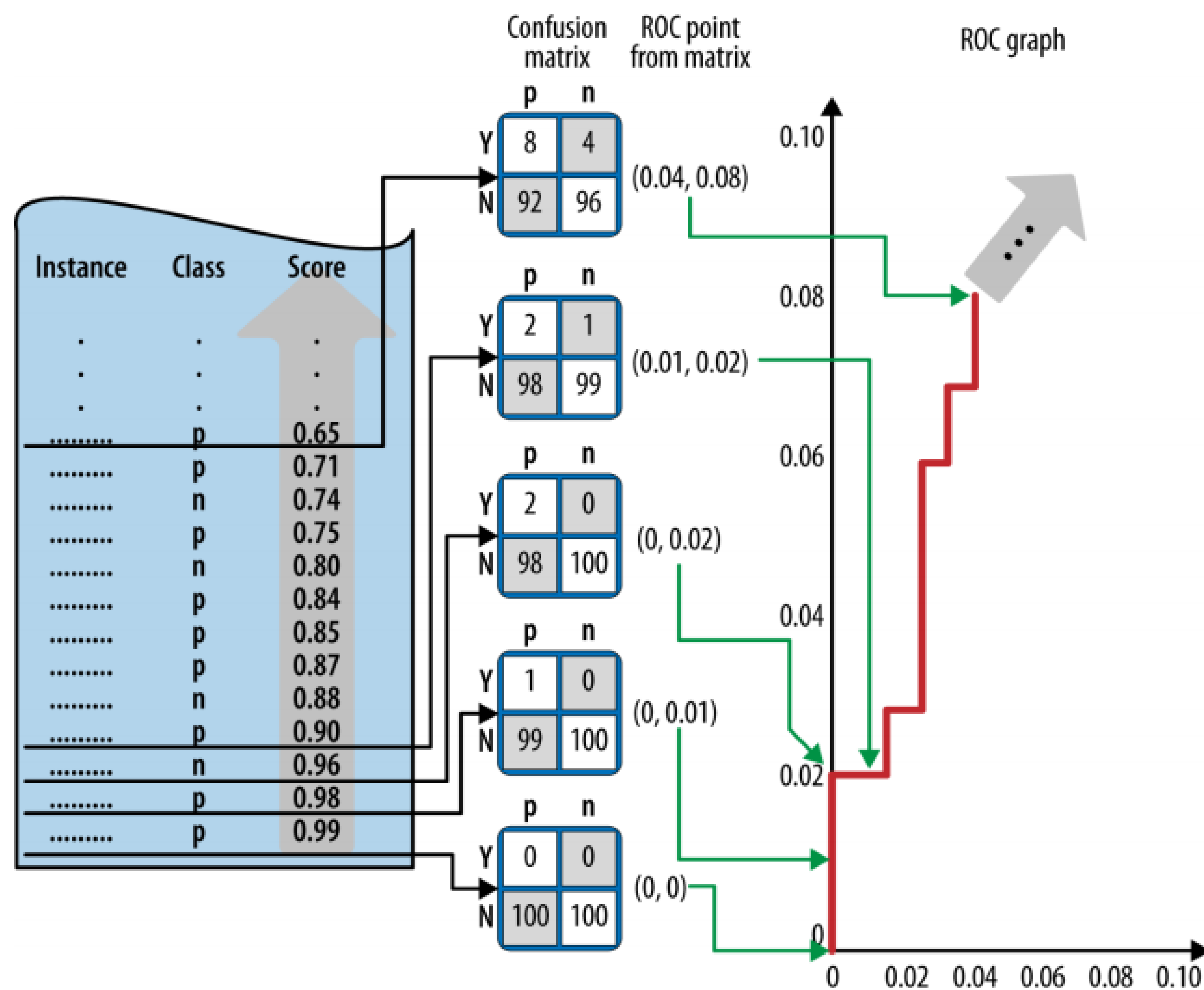
Quanto maior a área sob a curva melhor é o modelo ajustado



\* Regras de bolso sempre são perigosas, o modelo ideal depende sempre do problema modelado.



# Revisão: Curva ROC - Construção



# Estudo de Caso

## Matrix de Custo e Curva de Lucro no Python

**Fonte da dados:** `kaggle`

**Link:** <https://www.kaggle.com/kost13/us-income-logistic-regression/data>

**Resumo:** Dados do Censo Americano referentes a renda dos cidadãos e variáveis explicativas como Idade, Educação, raça, as quais podem influenciar nessa renda.

**Objetivo:** Ajustar um modelo para prever a renda da cidadão americano ( $\leq 50k-0$ ,  $>50k-1$ ), definir a matriz de custos e plotar a curva de lucro.





# Estudo de Caso

## Matrix de Custo e Curva de Lucro no Python

**Parte 1: Ajustar um modelo de regressão logística para prever a renda**  
**Calcular a matriz de confusão e plotar a curva ROC**



# Motivação

- Classificação regular
  - Objetivo: minimizar a taxa de missclassification
  - Todos os tipos de erros de classificação são considerados igualmente graves.
  - Problema: Cada tipo de erro pode ter um custo associado, ou seja, a classificação pode ser sensível aos custos
- Classificação por custo
  - Objetivo: maximizar os lucros ou minimizar os custos esperados.
  - Alguns erros de classificação podem ser são mais graves do que outros
  - Os custos podem depender da classe onde os casos foram classificados
  - Exemplo: Detecção de fraude - os custos não dependem apenas da fraude prevista, mas também da quantidade de dinheiro envolvida em cada caso.

		Classificação Preditada		
		0	1	
Classificação Real	0	TN	FP	Actual Negative
	1	FN	TP	Actual Positive
		Predicted Negative	Predicted Positive	



# Motivação

- A função de classificação tenta evitar erros de classificação com um alto peso.
- O trade-off de evitar erros de classificação "caros" é um número aumentado de erros de classificação "baratos".



**Número de Erros**



**Custo**

em comparação com a mesma classificação sem uma matriz de custos.

- Você pode atribuir pesos aos erros de classificação especificando uma matriz de custos.



# Exemplos – erros com pesos diferentes

- Marketing:

Objetivo: Encontrar clientes mais prováveis a aderir a uma promoção

Target: Se o cliente aderiu ou não a alguma promoção passada

Esforço: Contato telefônico com o cliente mais propenso a comprar

		Predicted	
		0	1
Actual	0	\$0	-\$1
	1	\$0	\$99

- Credit Scoring:

Objetivo: Verificar a probabilidade do cliente entrar em default

Target: Se o funcionário entrou ou não e default nos últimos 90 dias

		Predicted	
		0	1
Actual	0	TN	Deixa de Ganhar
	1	Perde	TP

**Ideia: Utilizar o fato dos erros terem pesos diferentes, ponderar essas diferenças e calcular qual o melhor retorno para o problema abordado.**



# Valor Esperado (VE)

- Agora estamos prontos para discutir uma ferramenta conceitual bastante útil para ajudar o pensamento analítico de dados: valor esperado.
- Valor esperado é a média ponderada dos **valores dos diferentes resultados possíveis**, onde o peso atribuído a cada valor é **a probabilidade de ocorrência**.  
Um cálculo de lucro esperado dá pesos maiores para os níveis de lucro mais prováveis, enquanto os níveis de lucro improváveis recebem pesos menores.
- A ideia é maximizar o lucro esperado.



# Valor Esperado

- *Cálculo do valor esperado*

$$VE = p(o1) * v(o1) + p(o2) * v(o2) + p(o3) * v(o3) + \dots$$

- $o_i$ : é um possível resultado de decisão;
- $p(o_i)$  é a sua probabilidade;
- $v(o_i)$  é o seu valor.

Vamos ilustrar o uso do valor esperado como uma estrutura analítica com dois cenários de ciência de dados diferentes.

1. Para estruturar o **uso** da classificação
2. Para estruturar o **valor** da classificação





# 1. VE para estruturar o uso da classificação

- Por exemplo: em marketing direcionado podemos querer atribuir a cada consumidor uma classe de resposta provável versus resposta não provável para a compra de determinado produto, então poderemos alvejar os prováveis compradores.  
Infelizmente, para o marketing direcionado a probabilidade de resposta para qualquer consumidor individual é muito baixa, talvez um ou dois por cento, portanto, nenhum consumidor pode parecer um provável respondedor.
- Se escolhermos um limite de "senso comum" de 50% para decidir qual é o provável respondedor, provavelmente não direcionaremos a ninguém.
- No quadro de valor esperado, podemos ver o cerne do problema.
  - Suponha que temos um modelo que dá uma probabilidade estimada de resposta para qualquer consumidor cuja descrição do vetor de característica  $x$  seja dada como entrada. Agora, gostaríamos de decidir quando segmentar um determinado consumidor descrito pelo vetor de características  $x$ .



# 1. VE para estruturar o uso da classificação

- O valor esperado fornece uma estrutura para a realização da análise. Especificamente, vamos calcular o benefício esperado (ou custo) de segmentação do consumidor  $x$
- Benefício esperado =  $p(x) * v(x) + (1 - p(x)) * vn(x)$   
 $v(x)$  e  $vn(x)$ : determinado pelo negócio  
 $v(x) = 99$  e  $vn(x) = -1$
- **$VE = P(x) * \$99 - (1 - p(x)) * \$1 > 0$ , logo**  
 **$p(x) * \$99 > (1 - p(x)) * \$1$ , ou seja,**  
 **$p(x) > 0.01$**
- Com esses valores de exemplo, devemos segmentar o consumidor enquanto a probabilidade estimada de resposta for maior que 1%.
- **Isso mostra como o cálculo de valor esperado pode expressar como usaremos o modelo.**





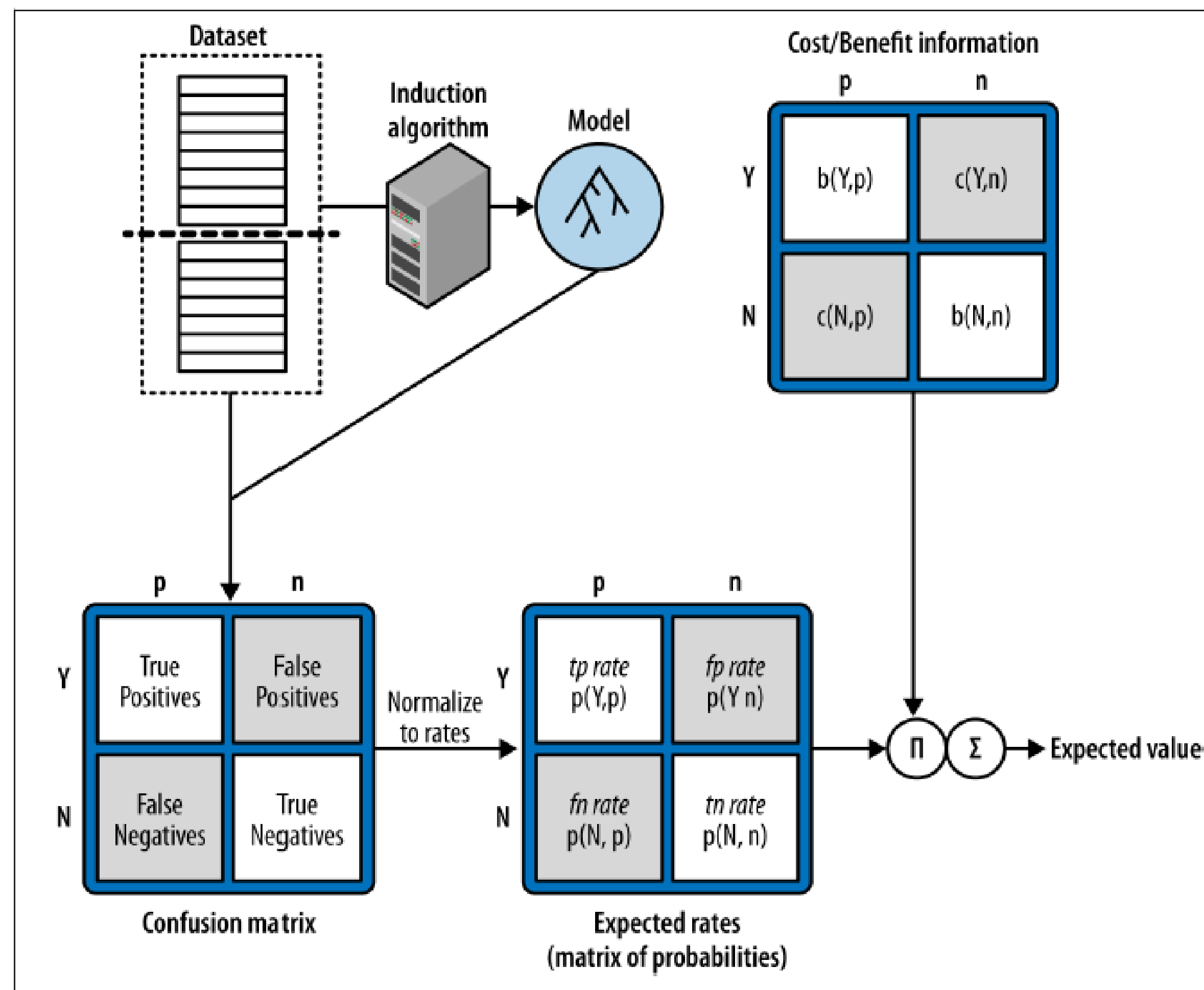
## 2. VE para estruturar o valor da classificação

- Queremos avaliar os modelos quanto as decisões tomadas no geral, para comparar modelos
  - O **modelo baseado em dados** funciona melhor do que o modelo feito à mão sugerido pelo grupo de marketing?
  - Uma árvore de classificação funciona melhor do que um modelo de regressão logística?
  - O modelo desenvolvido é **melhor que a escolha aleatória** dos consumidores a serem abordados?
- Podemos usar o “ponto de corte” que acabamos de descrever e, em seguida, usar o valor esperado de forma diferente para comparar os modelos.
- Vamos calcular o valor esperado para um modelo agregado.



## 2. VE para estruturar o valor da classificação

- Qual é a probabilidade associada à combinação particular de um consumidor que está previsto para comprar o produto mas na verdade não compra?



$$TP \text{ rate} = p(Y, p) = TP / T$$



## 2. VE para estruturar o valor da classificação

- Dada a matriz de confusão, obtida do modelo ajustado, vamos calcular as taxas esperadas, ou seja, a matriz de probabilidades

Matriz de Confusão

		Predicted Class	
		0	1
Actual Class	0	56	7
	1	5	42

		p	n
Y	Y	True Positives	False Positives
	N	False Negatives	True Negatives

Lembrando que  $TP\ rate = p(Y, p) = TP / T$

$$T = 110$$

$$p(Y, p) = 56/110 = 0.51 \quad p(Y, n) = 7/110 = 0.06$$

$$p(N, p) = 5/110 = 0.05 \quad p(N, n) = 42/110 = 0.38$$



Taxas Esperadas

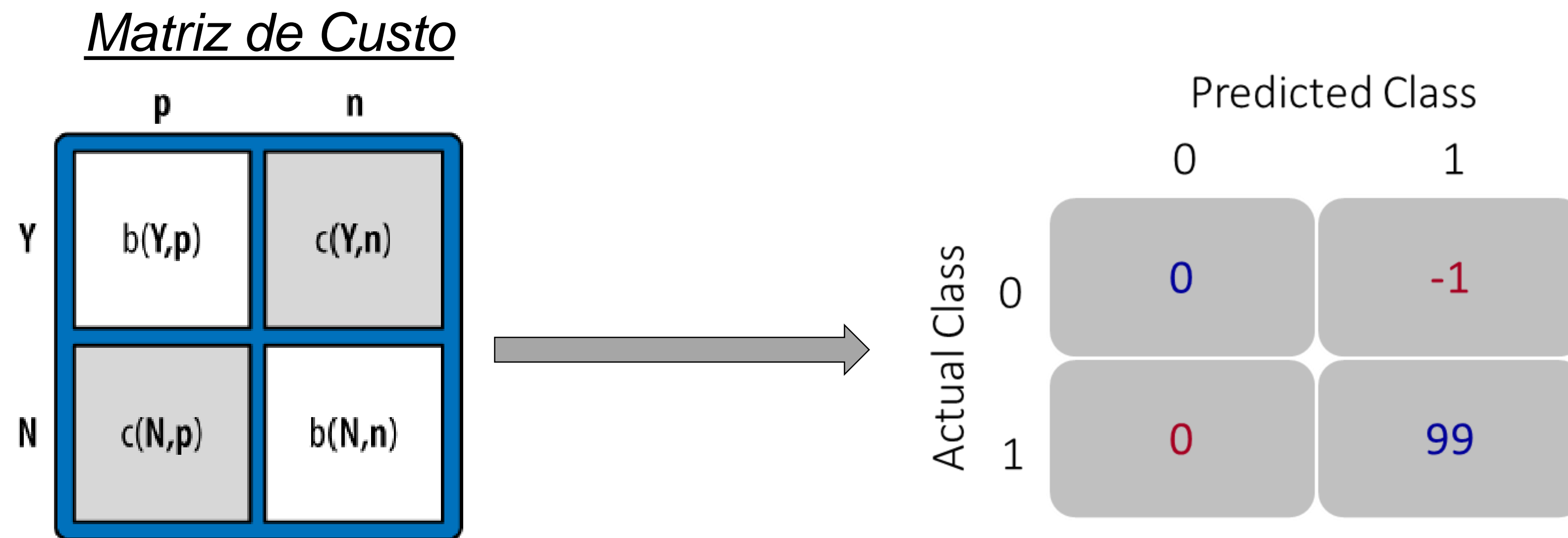
		Predicted Class	
		0	1
Actual Class	0	0,51	0,06
	1	0,05	0,38

		p	n
Y	Y	tp rate $p(Y, p)$	fp rate $p(Y, n)$
	N	fn rate $p(N, p)$	tn rate $p(N, n)$



## 2. VE para estruturar o valor da classificação

- Agora precisamos obter a matriz de custo, a partir de informação externa, ou seja, com alguém com conhecimento do negócio



# Matriz de Custo

Duas **armadilhas** que são comuns na formulação de matrizes de custo

- Os **sinais** de quantidades na matriz de custo devem ser consistentes. Aqui consideramos benefícios positivos e custos negativos, pois estamos maximizando o lucro.
  - Em muitos estudos de mineração de dados, o foco é minimizar o custo de modo que os sinais são invertidos. Matematicamente, não há diferença. No entanto, é importante escolher uma visão e ser consistente.
- Um erro fácil na formulação de matrizes custo-benefício é "**contar duas vezes**" colocando um benefício em uma célula e um custo negativo para a mesma coisa em outra célula (ou vice-versa).
  - Por exemplo: Um caso de fraude custa R\$1.000 em média
  - O benefício da captura de fraude é + R\$1.000 e o custo da falta é – R\$1.000
  - Se pensássemos assim a melhoria em benefício seria  $1000 - (-1000) = 2000$  – isso indica contagem dupla
  - O correto é colocar o benefício da captura igual a R\$1.000 e o custo da falta igual a 0 ou colocar o custo da falta como –R\$1.000 e o benefício da captura como 0.



# Estudo de Caso

## Matrix de Custo e Curva de Lucro no Python

**Parte 2: Definir a matriz de custo/benefício – a partir de informações de negócio**





## 2. VE para estruturar o valor da classificação

- Com as matrizes de taxas esperadas e de custo podemos calcular o lucro esperado

Taxas Esperadas

	p	n
Y	tp rate $p(Y,p)$	fp rate $p(Y,n)$
N	fn rate $p(N,p)$	tn rate $p(N,n)$

Matriz de Custo

	p	n
Y	$b(Y,p)$	$c(Y,n)$
N	$c(N,p)$	$b(N,n)$

$$\text{Lucro Esperado} = p(Y,p) * b(Y,p) + p(Y,n) * c(Y,n) + p(N,p) * c(N,p) + p(N,n) * b(N,n)$$

	Predicted Class	
	0	1
Actual Class 0	0,51	0,06
Actual Class 1	0,05	0,38

	Predicted Class	
	0	1
Actual Class 0	0	-1
Actual Class 1	0	99

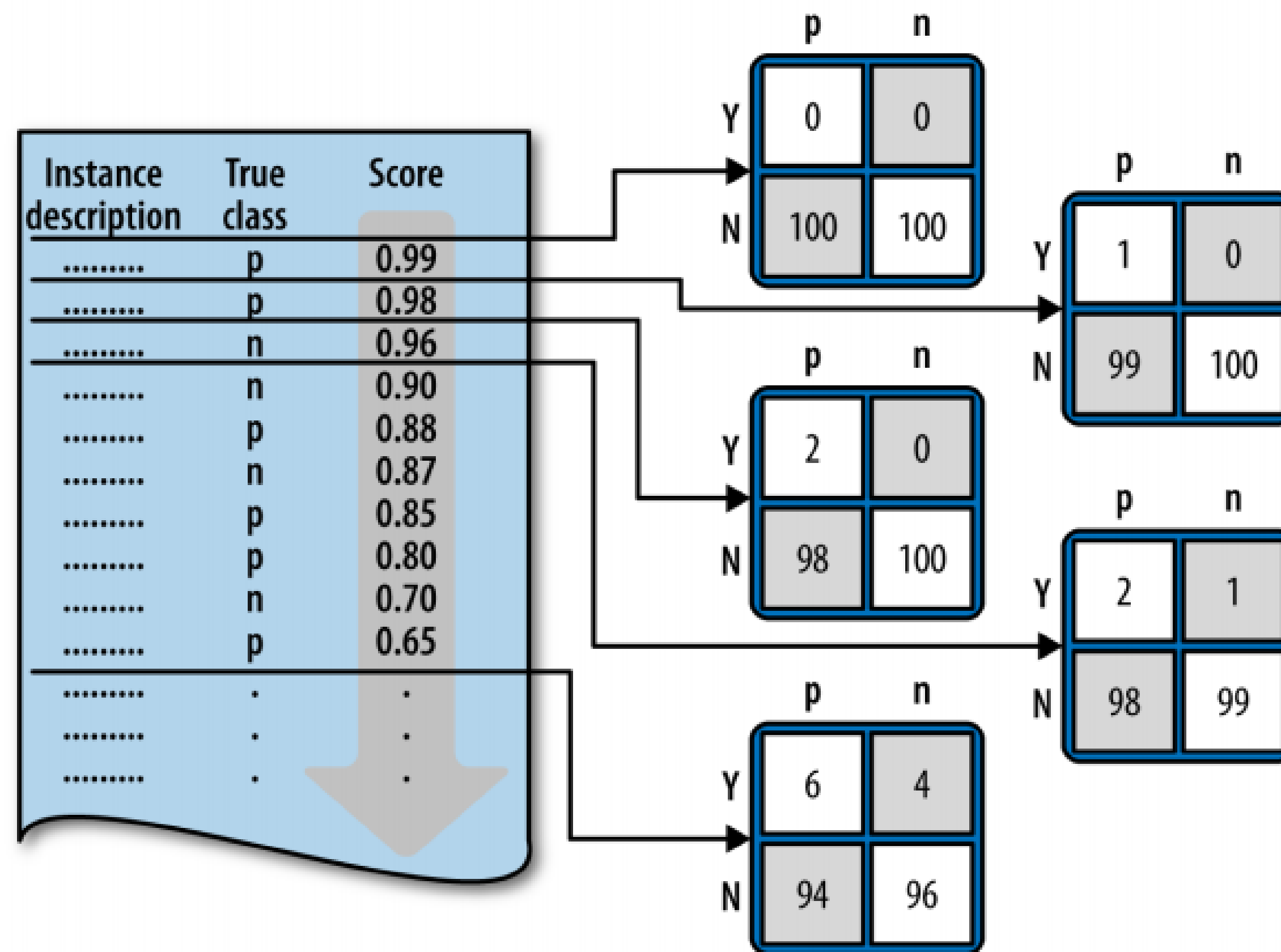
$$\begin{aligned} \text{Lucro Esperado} &= 0,51 * 0 + 0,06 * (-1) + \\ &\quad 0,05 * 0 + 0,38 * 99 \\ &= 37,56 \end{aligned}$$



# Lucro Esperado

Para cada matriz de confusão, ou seja, para cada ponto de corte, dado uma matriz de custo, temos um lucro esperado.

Como podemos, a partir do lucro esperado, determinar qual o melhor ponto de corte?





# Lucro Esperado

## Matrizes de Confusão

		<u>Matriz de Custo</u>			
		Predicted			
		0	1		
Actual	0	\$0	-\$1	70	5
	1	\$0	\$99	9	16
				66	9
				4	21
				57	18
				1	24

## Lucro Esperado



# Lucro Esperado

		<u>Matrizes de Confusão</u>		<u>Lucro Esperado</u>	
		<u>Matriz de Custo</u>			
		Predicted			
		0	1		
Actual	0	\$0	-\$1	70	5
	1	\$0	\$99	9	16
		66	9		
		4	21		
		57	18		
		1	24		

$$(16/100)*99 + (5/100)*(-1) = \$15,79$$

Beneficio

Custo

Lucro



# Lucro Esperado

# Matrizes de Confusão

## Lucro Esperado

### Matriz de Custo

Predicted

0

1

Actual

0

1

\$0

-\$1

\$0

**\$99**

70

5

9

16

66

9

4

21

57

18

1

24

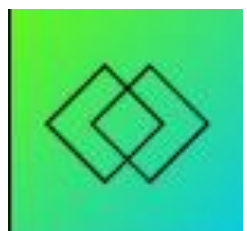
$$(16/100)*99 + (5/100)*(-1) = \$15,79$$

## Beneficio

# Custo

# Lucro

$$(21/100)*99 + (9/100)*(-1) = \$20,70$$



# Lucro Esperado

		<u>Matrizes de Confusão</u>		<u>Lucro Esperado</u>		
		<u>Matriz de Custo</u>				
		Predicted				
		0	1			
Actual	0	\$0	-\$1	70	5	$\underbrace{(16/100)*99}_{\text{Beneficio}} + \underbrace{(5/100)*(-1)}_{\text{Custo}} = \underbrace{\$15,79}_{\text{Lucro}}$
	1	\$0	\$99	9	16	
	0	66	9			$(21/100)*99 + (9/100)*(-1) = \$20,70$
	1	4	21			
	0	57	18			$(24/100)*99 + (18/100)*(-1) = \$23,58$
	1	1	24			

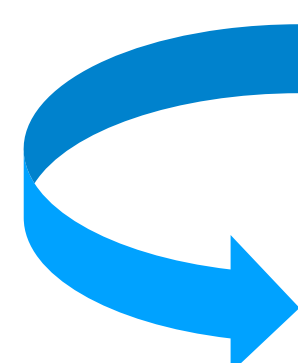


# Lucro Esperado

		<u>Matrizes de Confusão</u>		<u>Lucro Esperado</u>		
		<u>Matriz de Custo</u>				
		Predicted				
		0	1			
Actual	0	\$0	-\$1	70	5	$\underbrace{(16/100)*99}_{\text{Beneficio}} + \underbrace{(5/100)*(-1)}_{\text{Custo}} = \underbrace{\$15,79}_{\text{Lucro}}$
	1	\$0	\$99	9	16	
	0	66	9			$(21/100)*99 + (9/100)*(-1) = \$20,70$
	1	4	21			

57	18
1	24

$$(24/100)*99 + (18/100)*(-1) = \$23,58$$



O ponto de corte ótimo é aquele que gera o maior lucro



# Estudo de Caso

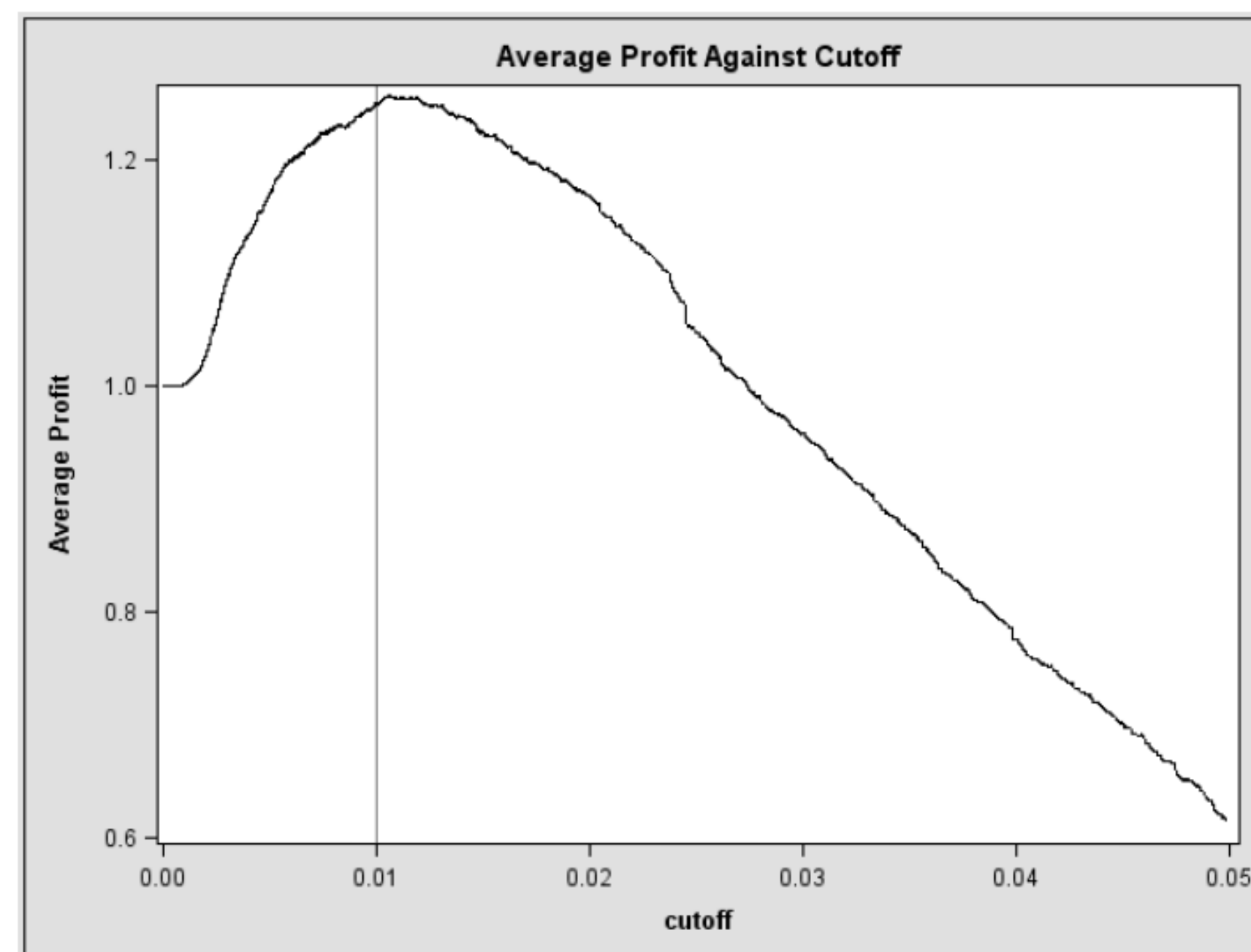
## Matrix de Custo e Curva de Lucro no Python

**Parte 3: Calcular o lucro para um ponto de corte específico**  
**Calcular o lucro para vários pontos de corte**



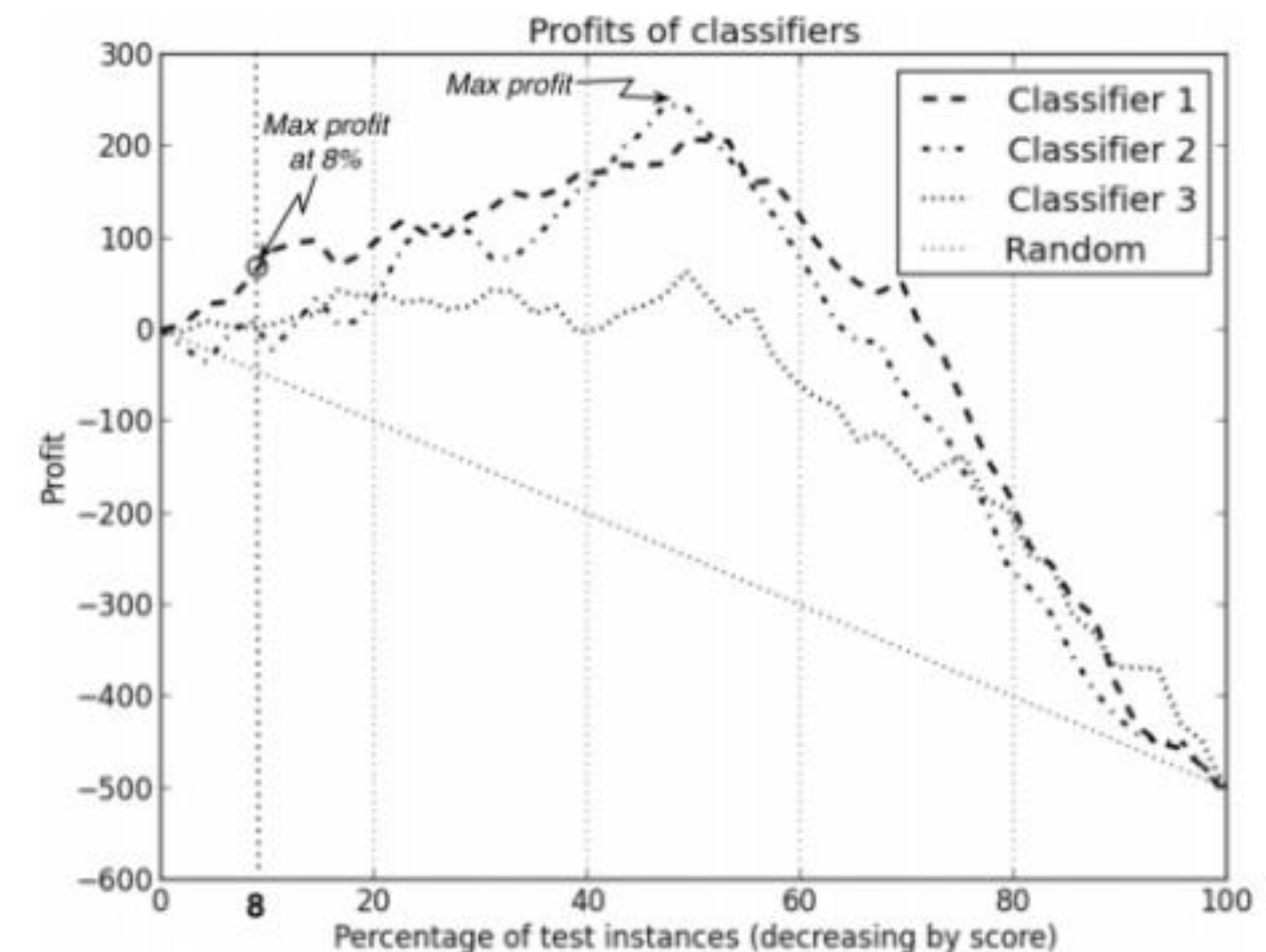
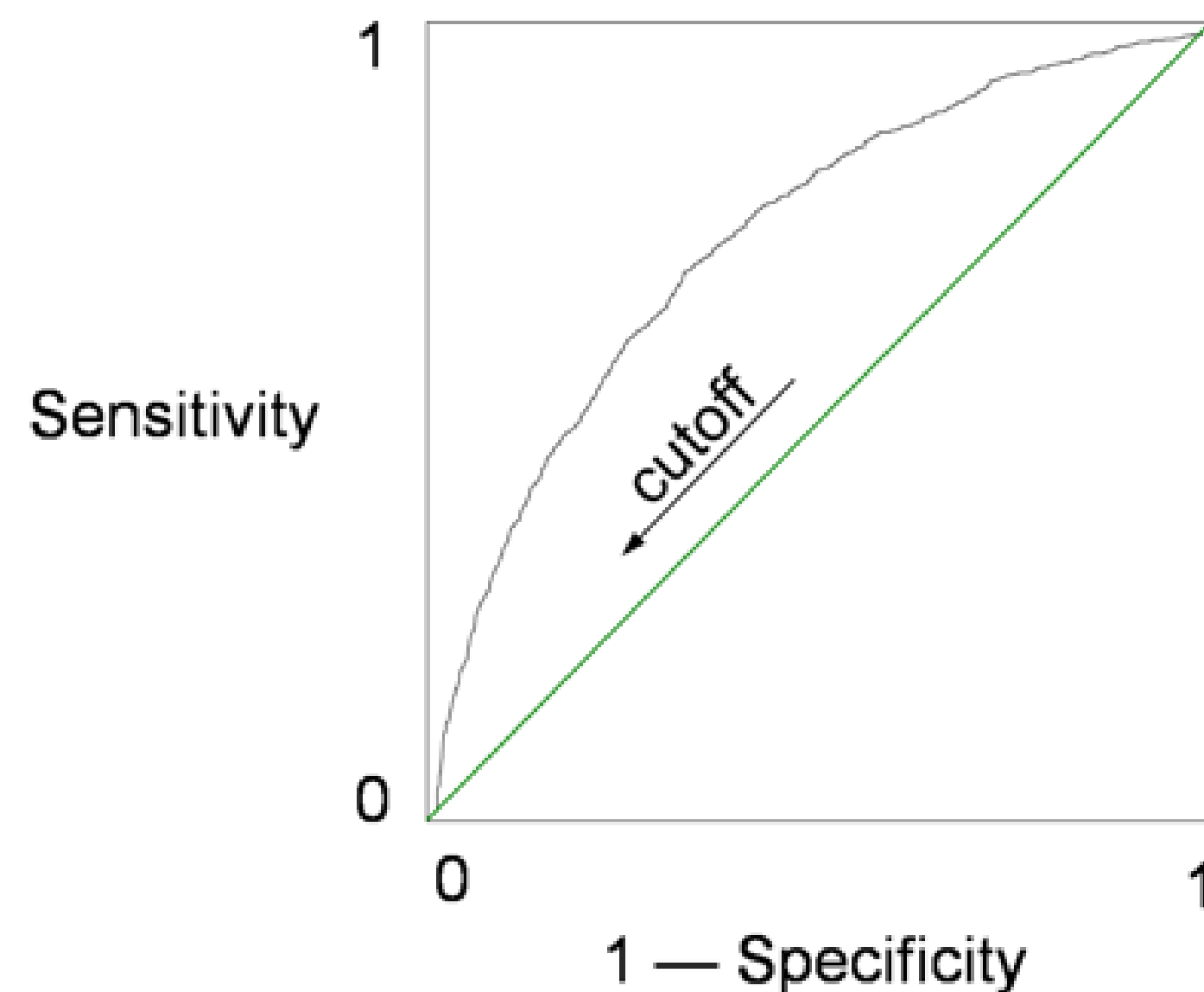
# Curva Lucro

- A **Curva de lucro** mostra a rentabilidade estimada associada ao uso de um modelo de mineração.
- Por exemplo: para um modelo que prediz quais clientes uma empresa deve entrar em contato em um cenário de negócios. Nesse caso, ao adicionar à tabela de lucros informações sobre o custo da condução da campanha é possível ver o lucro estimado se os clientes estiverem corretamente segmentados, em comparação com os clientes que são contatados aleatoriamente.



# Comparação: ROC X Curva Lucro

- A Curva ROC plota a sensibilidade versus (1-especificidade) para “todos” os pontos de corte.
  - O ponto de corte ótimo pode ser obtido tomando o máximo valor da curva ROC
- A Curva de lucro utiliza “todos” os pontos de corte para calcular os possíveis lucros
  - O ponto de corte ótimo pode ser obtido tomando o máximo valor da curva de Lucro





# Estudo de Caso

## Matrix de Custo e Curva de Lucro no Python

**Parte 4: Plotar a curva de lucro versus a probabilidade**



# Estudo de Caso

## Matrix de Custo e Curva de Lucro no Python

**Parte 5: Automatizar o processo de geração das curvas de lucro para o ajuste de vários tipos de modelos**



# Desafio

## Matrix de Custo e Curva de Lucro no Python

1. Ajustar um modelo para classificação
2. Definir a matriz de custo
3. Calcular o lucro para varios pontos de corte
4. Plotar a curva de lucro
4. Ajustar outros modelos
5. Plotar as curvas de lucro de cada modelo e selecionar o que fornece maior lucro



# DÚVIDAS?!



# Obrigada

***Cristiane Rodrigues***

*[crisrodrigues\\_27@hotmail.com](mailto:crisrodrigues_27@hotmail.com)*

