

FIAP

NBA

MBA em DATA SCIENCE & ARTIFICIAL INTELLIGENCE

APPLIED STATISTICS



Dra. Regina Tomie Ivata Bernal Cientista de Dados na área da Saúde

Formação Acadêmica:

Estatístico - UFSCar

Mestre em Saúde Pública – FSP/USP

Doutor em Ciências – Epidemiologia - FSP/USP

Atividades Profissionais:

Professora de pós-graduação na FIAP

Consultora externa da SVS/MS

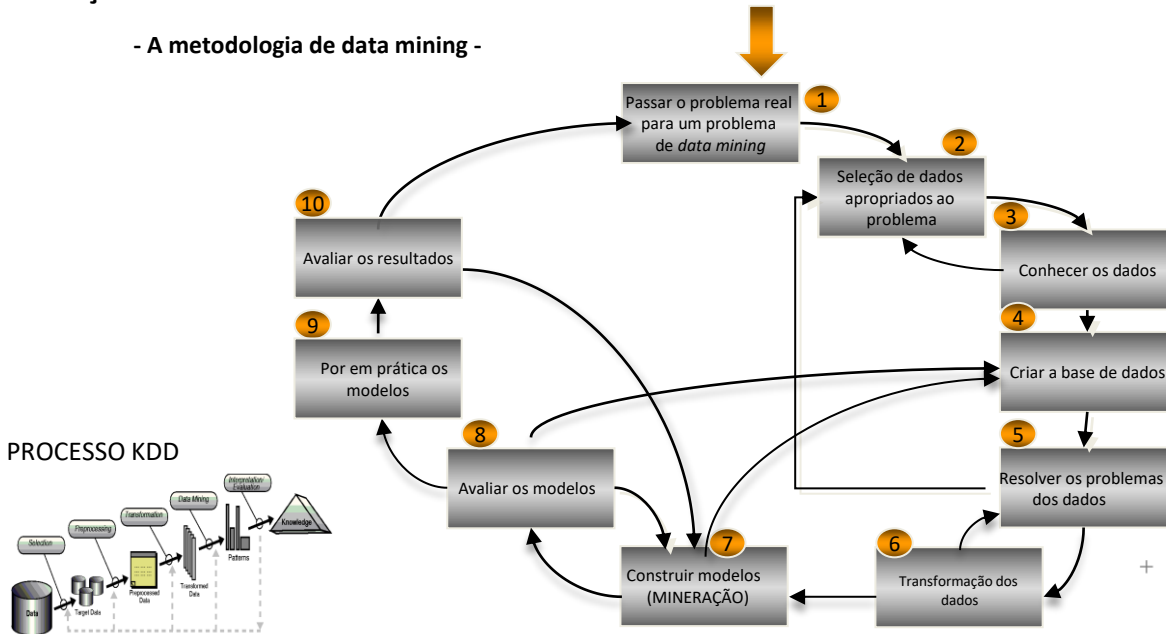
Cientista de Dados em Saúde

profregina.bernal@fiap.com.br
reginabernal@terra.com.br

Extração de Conhecimento

Extração de conhecimento em bases de dados

- A metodologia de data mining -



Fonte: Berry & Linoff, 2004.

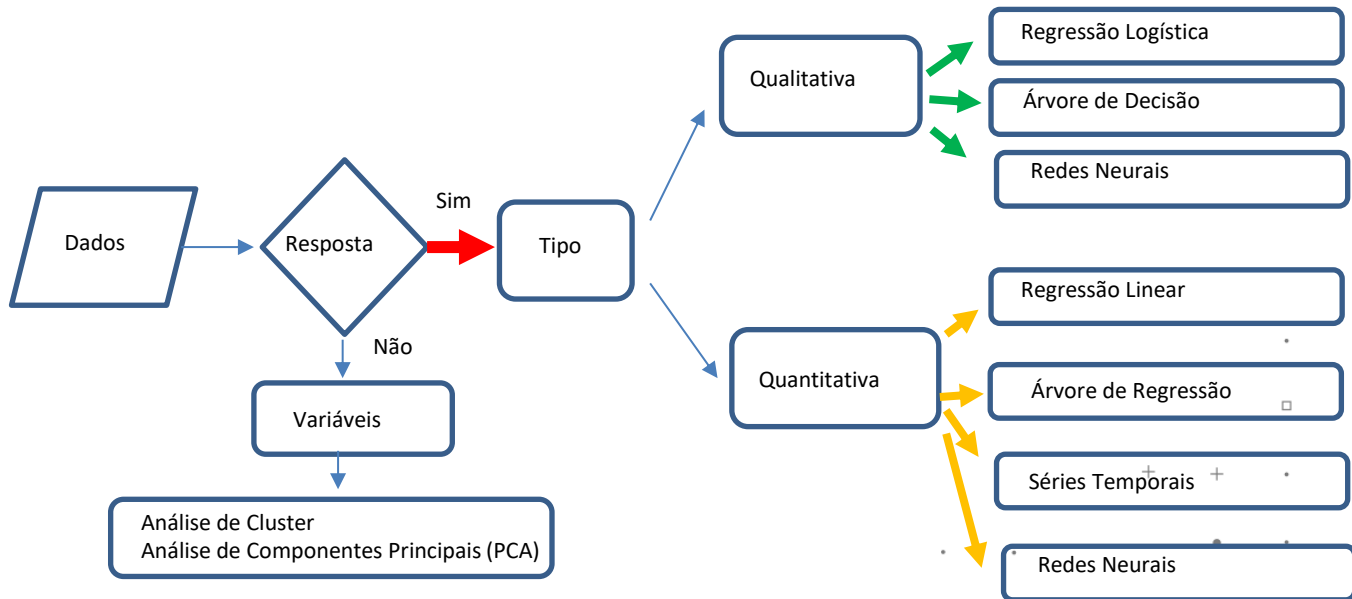
Técnicas Estatísticas

Construir modelos
(MINERAÇÃO)

Extração de conhecimento em bases de dados

- A metodologia de data mining -

Técnicas Supervisionadas



Técnicas Não Supervisionadas

Uso dos Modelos no Ciclo do Cliente



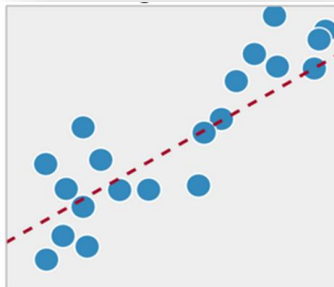
Modelos Preditivos

Data Mining: Mineração - Construção de Modelos

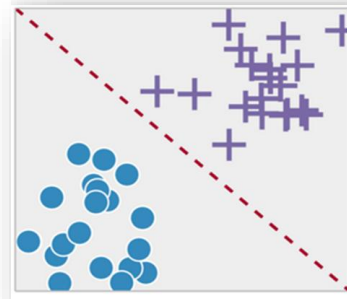


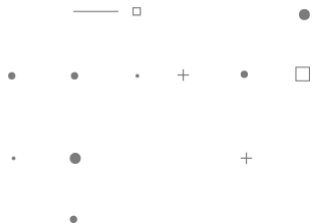
aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

- **Regressão:** Compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores numéricos reais. Esta tarefa é similar à tarefa de Classificação, com a diferença de que o atributo alvo assume valores numéricos.



- **Classificação.** A tarefa de Classificação consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram.





TÉCNICAS SUPERVISIONADAS



Análise de Discriminação de Estrutura

• • • + • □

• • +

REGRESSÃO LOGÍSTICA

Encontrar uma **função logística**, formada através de ponderações das variáveis (atributos), cuja resposta permita estabelecer a **probabilidade de ocorrência** de determinado evento e a **importância das variáveis** (peso) para esta ocorrência.

□

+ + •

• • • •

□ • • • •

Análise de Discriminação de Estrutura

• REGRESSÃO LOGÍSTICA

Cientes



Cientes sem o evento

Cientes com o evento(*)

(*) Evento (exemplos)

- Aquisição
- Cancelamento
- Pagamento

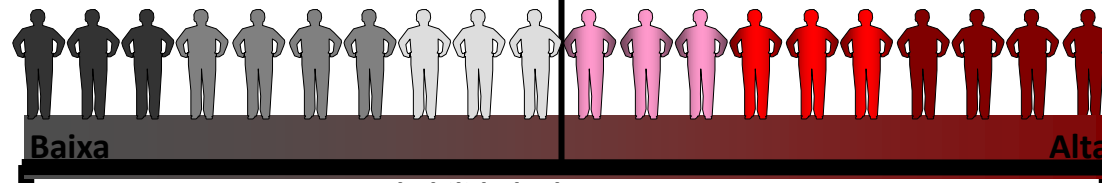
Modelo

Pontuação dos Clientes com Probabilidade



Regra de Decisão

Corte



Ranking dos +Cientes

0%

Probabilidade do evento ocorrer

100%

ANÁLISE DE REGRESSÃO LOGÍSTICA

Probabilidade

Sendo Y: a resposta à preferência por um evento (sim ou não),

→ a probabilidade de:

- Preferência (*ou sucesso*) será **p**
- Não-preferência (*de fracasso*) será **(1-p)**

“Chance de Ocorrência de um Evento”

➤ **Chance** = (probabilidade de sucesso) / (probabilidade de fracasso)

Exemplo, se a probabilidade de sucesso é 0,65:

a chance é igual a: $p / (1-p) = p / q = 0,65 / 0,35 = 1,86$

ANÁLISE DE REGRESSÃO LOGÍSTICA

Exemplo: Preferência por canal de futebol

Sexo	Prefere	Não Prefere	Total
Masculino	146	120	266
Feminino	110	124	234
Total	256	244	500

➤ **Chance** de preferir o canal de futebol entre homens:

$$➤ p_1 / (1 - p_1) = (146/266) / (120/266) = 0,55 / 0,45 = 1,22$$

➤ **Chance** de preferir o canal de futebol entre mulheres:

$$➤ p_2 / (1 - p_2) = (110/234) / (124/234) = 0,47 / 0,53 = 0,89$$

➤ **Razão de chances** de preferir canal de futebol entre homens, em relação às mulheres:

$$➤ [p_1 / (1 - p_1)] / [p_2 / (1 - p_2)] = 1,22 / 0,89 = 1,37$$

ANÁLISE DE REGRESSÃO LOGÍSTICA

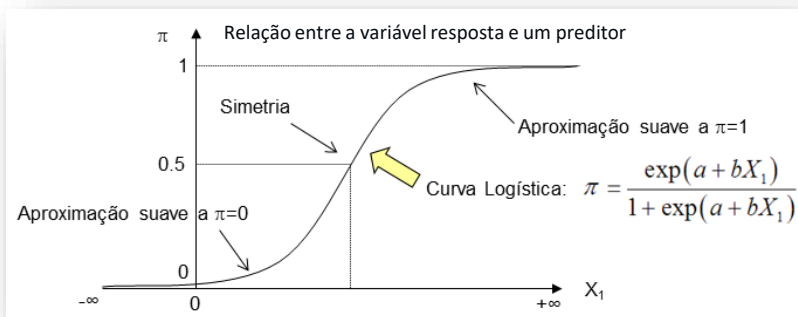
Modelo de Regressão Logística

$$Y = a + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$$

Y: logit da resposta de preferência (sim)

a : intersecção

B1, B2, ..., Bn : coeficientes logísticos



- A função logística é dada pelo logito-inverso (anti-logit) que nos permite transformar o logito em probabilidade:

$$Probabilidade = \frac{\exp(Y)}{1 + \exp(Y)}$$

Método de Estimação dos Coeficientes

➤ Regressão Linear: Método dos Mínimos Quadrados

⇒ É o método que determina a linha reta mais apropriada, minimizando a soma dos quadrados das diferenças entre os valores estimados de Y por meio da reta de regressão e os valores observados de Y.

➤ Logística: Método da Máxima Verossimilhança

(algoritmo iterativo)

⇒ Consiste em determinar uma função, denominada função de verossimilhança $[L(y, \theta)]$, que é a função de probabilidade de ocorrência de um específico conjunto de dados e estimar os parâmetros que maximizam a mesma.

ANÁLISE DE REGRESSÃO LOGÍSTICA

Seleção Conjuntos de Atributos (Variáveis)

- Variáveis Discriminantes
- Variáveis Não-Discriminantes

Instrumento para selecionar variáveis(atributos) significativos

BACKWARD
FORWARD
STEPWISE

- Backward Selection : Procedimento constrói **adicionando todas as variáveis** e vai eliminando iterativamente uma a uma até que não haja mais variáveis .
- Forward Selection: Procedimento constrói iterativamente **adicionando variáveis uma a uma** até que não haja mais variáveis preditoras .
- Stepwise: Combinação de Forward Selection e Backward elimination. Procedimento constrói iterativamente uma seqüência de modelos pela adição ou remoção de variáveis em cada etapa. + + .

ANÁLISE DE REGRESSÃO LOGÍSTICA

Qualificação do Ajuste do Modelo

Matriz de Classificação

Estatística de Ajuste

Verossimilhança : $-2 \log$ Verossimilhança

Significância do Modelo : Qui-quadrado (similar ao F regressão)

Ganho no Modelo (significância)

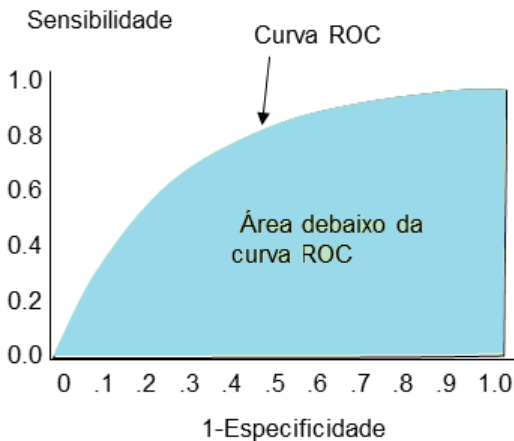
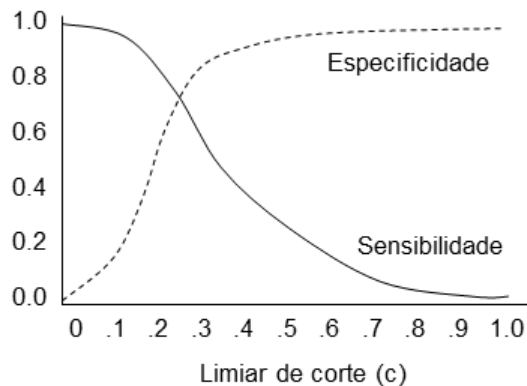
ANÁLISE DE REGRESSÃO LOGÍSTICA

Qualificação do Ajuste do Modelo

		Previsão do modelo		Total
		y=1	y=0	
Obs.	y=1	n1	n2	n1+n2
	y=0	n3	n4	n3+n4

$$\text{Sensibilidade} = n1 / (n1+n2)$$

$$\text{Especificidade} = n4 / (n3+n4)$$



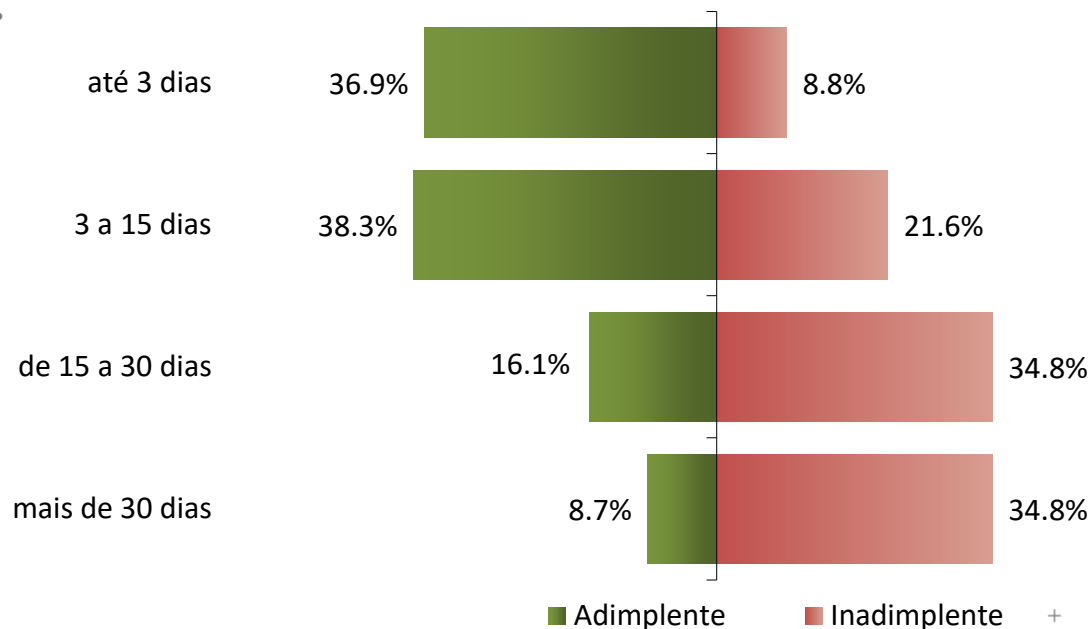
Exemplo – Modelo de Inadimplência

Segmento: Cartões de Crédito

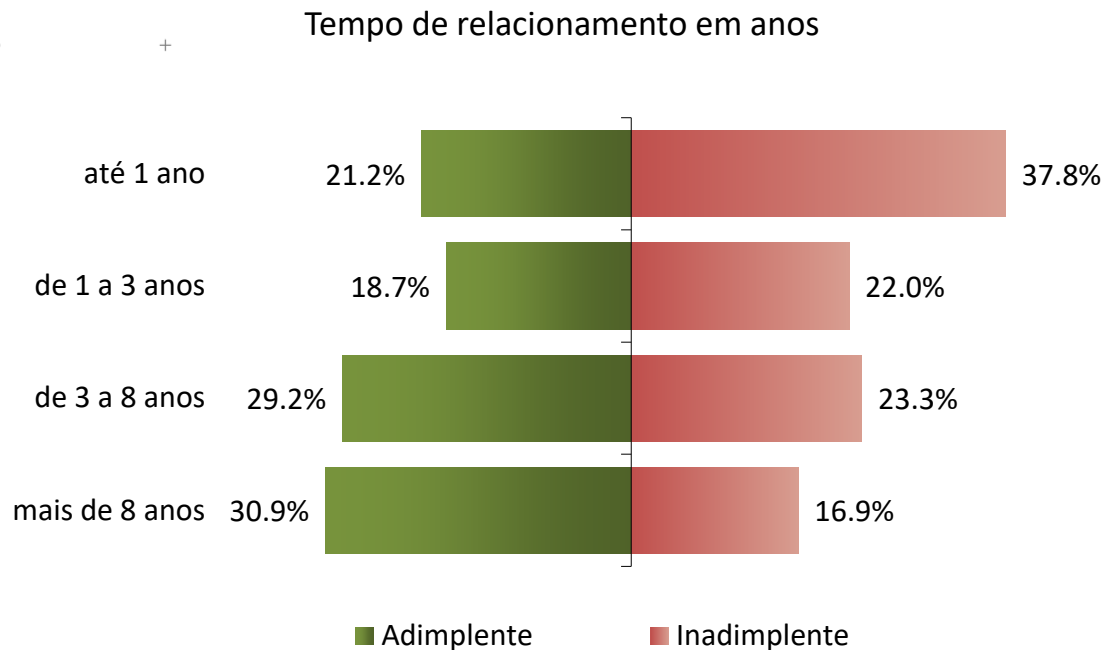
A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência

Exemplo – Modelo de Inadimplência

Média de dias com pagamentos em atraso nos últimos 6 meses

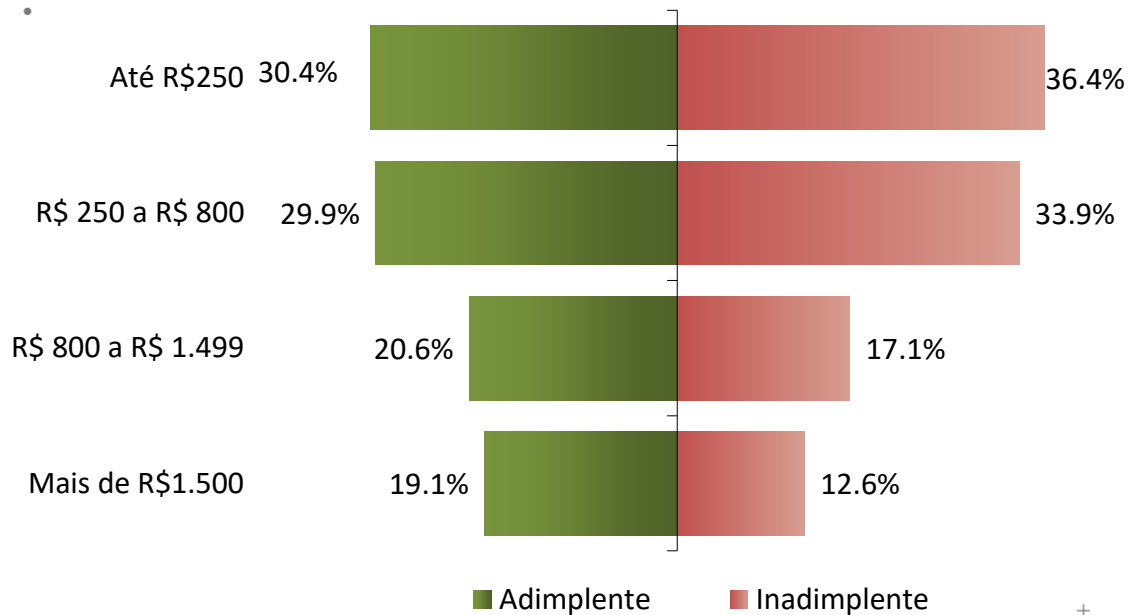


Exemplo – Modelo de Inadimplência

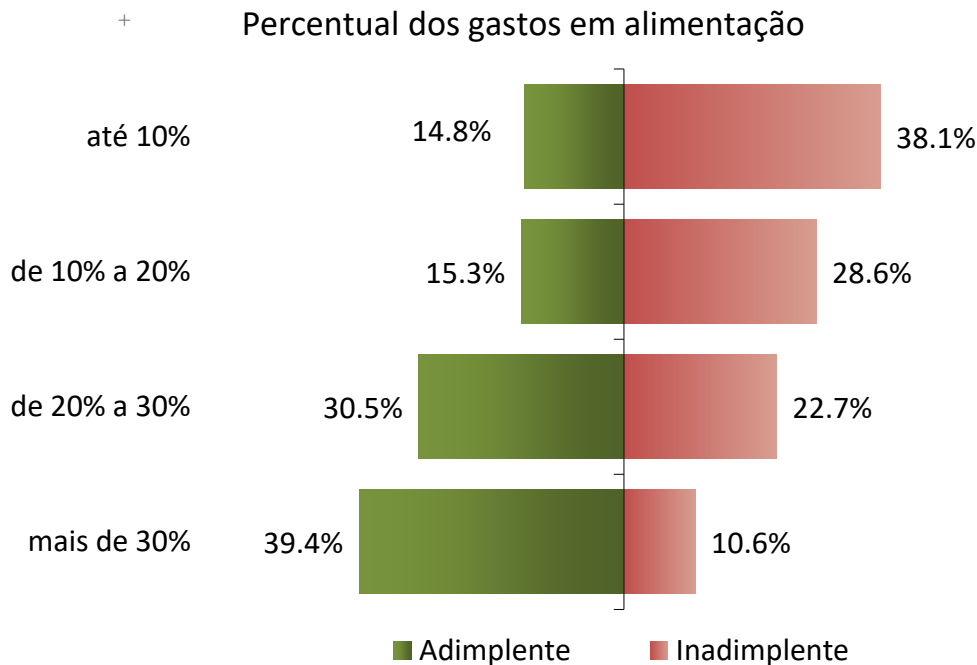


Exemplo – Modelo de Inadimplência

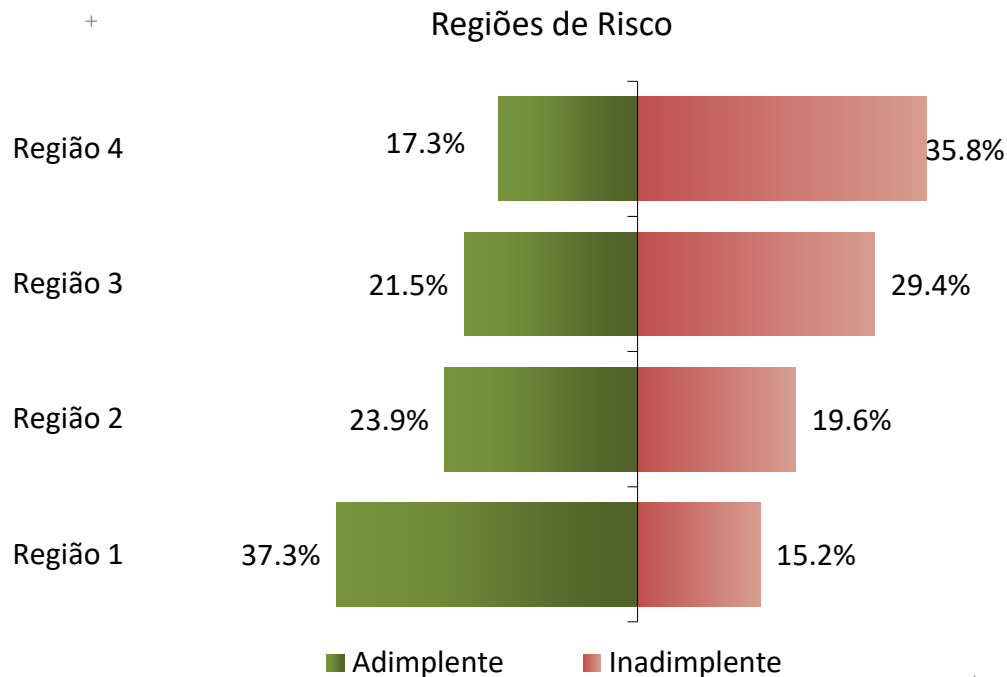
Valor Médio da Fatura Mensal



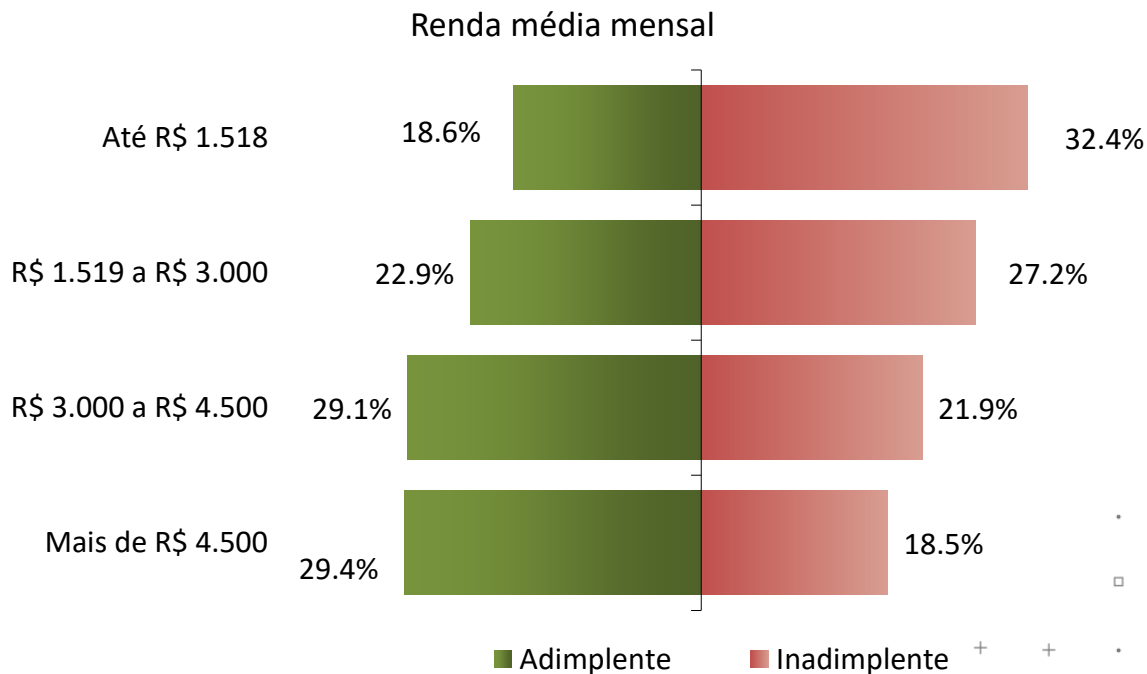
Exemplo – Modelo de Inadimplência



Exemplo – Modelo de Inadimplência



Exemplo – Modelo de Inadimplência



Exemplo – Modelo de Inadimplência

Tabela de Coeficientes do Modelo

variável	categoria	Coeficientes
fatura em atraso	até 3 dias	-1,276
	3 a 15 dias	-0,611
	de 15 a 30 dias	0,580
	mais de 30 dias	1,308
Tempo de cliente	até 1 ano	0,580
	de 1 a 3 anos	0,401
	de 3 a 8 anos	-0,264
	mais de 8 anos	-0,718
valor da fatura	Até R\$250	0,262
	R\$ 250 a R\$ 800	0,103
	R\$ 800 a R\$ 1.499	-0,105
	Mais de R\$1.500	-0,261
% de gasto com alimentação	até 10%	0,581
	de 10% a 20%	0,401
	de 20% a 30%	-0,264
	mais de 30%	-0,718
Região de Risco	Região 4	1,067
	Região 3	0,371
	Região 2	-0,368
	Região 1	-1,069
renda mensal	Até R\$ 1.518	0,455
	R\$ 1.519 a R\$ 3.000	0,080
	R\$ 3.000 a R\$ 4.500	-0,122
	Mais de R\$ 4.500	-0,413
Constante		0,099

Exemplo – Modelo de Inadimplência

Modelo Logístico

Pesos definidos na modelagem

-1,276	Até 3 dias	Fatura em atraso	Mais de 30 dias	1,308
-0,718	Mais de 8 anos	Tempo de Relacionamento	Até 1 ano	0,580
-0,261	Mais de R\$1.500	Valor da Fatura	Até R\$250	0,262
-0,718	Mais de 30%	% de gasto com alimentação	Até 10%	0,580
-1,069	Região 1	Região de Risco	Região 4	1,067
-0,413	Mais de R\$4.500	Renda Mensal	Até R\$1.518	0,455
0,099		Constante		0,099
4%	Propensão			98%

$$y = 0.099 + 1.308 + 0.580 + 0.262 + 0.580 + 1.067 + 0.455 = 4.351$$

$$\text{Propensão} = \text{probabilidade} = \frac{\exp(4.351)}{1 + \exp(4.351)} = 0.98 \text{ ou } 98\%$$

ANÁLISE DE REGRESSÃO LOGÍSTICA

Exemplo: Modelo Cross-Selling Propensão à Compra de um Produto

Objetivo

Estabelecer público-alvo para a venda qualificada de um determinado Produto X, com uso dos mailing's internos do cliente, através do desenvolvimento de modelos preditivos.

Modelos Cross Selling

• • • + • □

➔ Propensão de compra do Produto X



Sem Produto X

Com Produto X

+ + •

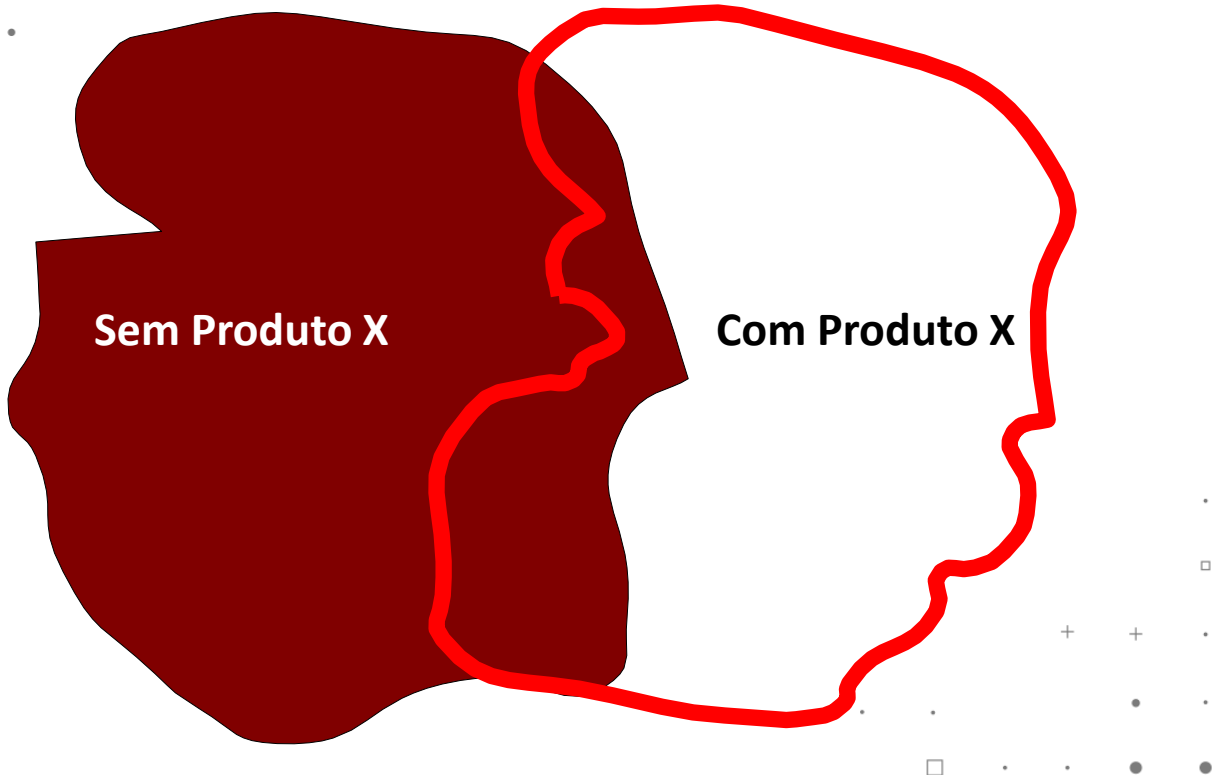
• • • •

□ • • • •

Modelos Cross Selling

• • • + • □

➔ Propensão de compra do Produto X

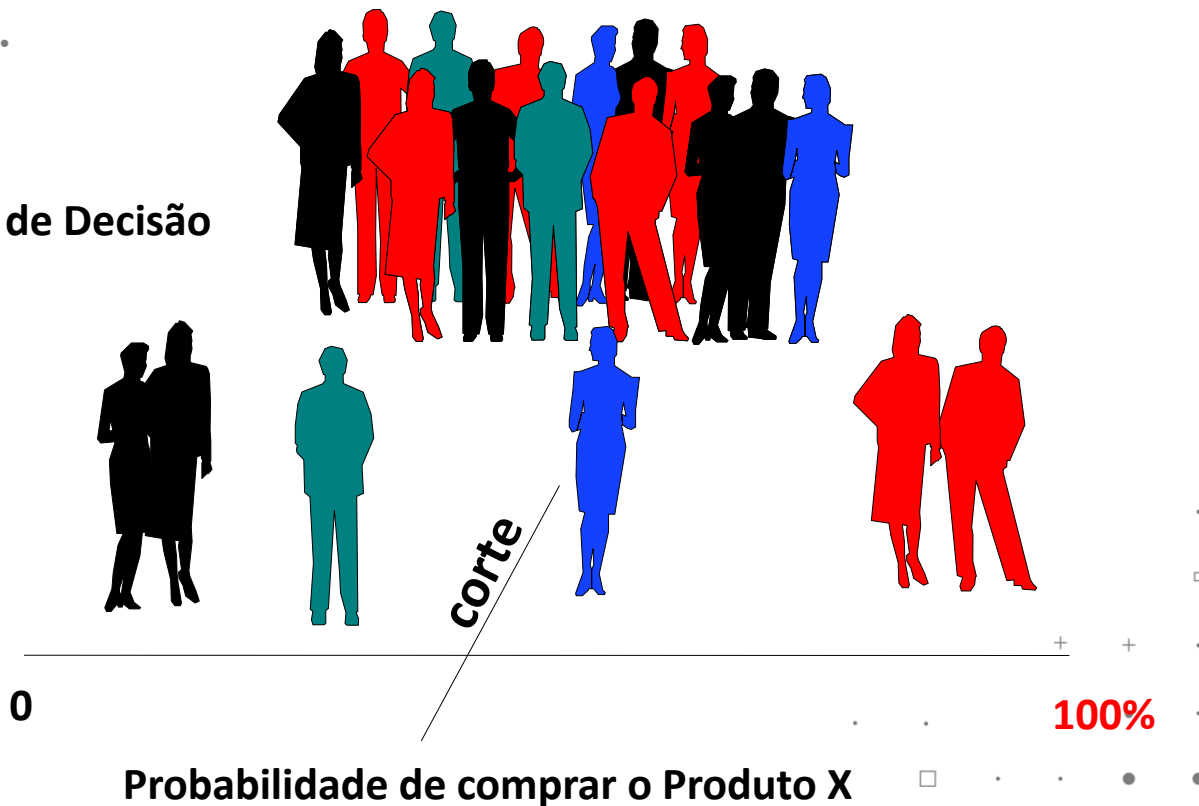


•
□
+ + •
• •
□ • • • •

Modelos Cross Selling

→ Propensão de compra do Produto X

Regra de Decisão



Modelos Cross Selling - Implementação

➔ Propensão de compra do Produto X

Algoritmo Matemático

Para associar uma probabilidade de compra de um produto X à cada cliente, os seguintes passos devem ser tomados:

1. Identificar as variáveis, associando os respectivos coeficientes;
2. Somar os coeficientes encontrados no item 1, juntamente com a constante do modelo determinando o valor de Y;
3. Efetuar a operação matemática que se segue, para determinação final do score.

$$Probabilidade = \frac{\exp(Y)}{1 + \exp(Y)}$$

Modelos Cross Selling

➔ Propensão de compra do Produto X

Regra de Decisão Estatística

Após associar à cada indivíduo sua probabilidade de compra do produto, deve-se submetê-la à Regra de Decisão, ou seja, se a probabilidade obtida for menor ou igual ao valor de corte* o assinante pertencerá ao grupo que não irá adquirir o produto, caso contrário, se esta probabilidade for maior que o valor de corte, ele pertencerá ao grupo que irá adquirir.

* valor de corte é o valor de probabilidade que define os grupos, segundo análise de acertos do modelo.

Data Mining



aplicações práticas de Data Mining se podem ser categorizadas de acordo com a tarefa que se pretende resolver

Exercitando!!!!



Base Inadimplência



SAÍDAS DO PYTHON

REGRESSÃO LOGÍSTICA

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência

Variável	Descrição	Valor	Tipo de variável	Regra de data mining
atrasos	Atrasos	1-Ate 3 dias	Qualitativa ordinal	Preditora
		2-3 a 15 dias		
		3-15 a 30 dias		
		4-Mais de 30 dias		
temporel	Tempo de relacionamento	1-ate 1 ano	Qualitativa ordinal	Preditora
		2-1 a 3 anos		
		3-3 a 8 anos		
		4-mais 8 anos		
valorfatura	Valor da fatura	1-Ate R\$250	Qualitativa ordinal	Preditora
		2-R\$250 a R\$800		
		3-R\$800 a R\$1499		
		4-R\$1500 e mais		
p_gastoalim	% de gastos com alimentação	1-Ate 10%	Qualitativa ordinal	Preditora
		2-10% a 20%		
		3-20% a 30%		
		4-30% e mais		
regiao_risco	Região de Risco	I	Qualitativa nominal	Preditora
		II		
		III		
		IV		
rendamensal	Renda mensal	1-Ate R\$18	Qualitativa ordinal	Preditora
		2-R\$18 a R\$3000		
		3-R\$3000 a R\$4500		
		4-mais de R\$4500		
Resposta	Resposta	0=Não	Qualitativa nominal	Target
		1=Sim		

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25984 entries, 0 to 25983
Data columns (total 19 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Resposta             25984 non-null  int64
1   atrasos_2            25984 non-null  uint8
2   atrasos_3            25984 non-null  uint8
3   atrasos_4            25984 non-null  uint8
4   temporel_2           25984 non-null  uint8
5   temporel_3           25984 non-null  uint8
6   temporel_4           25984 non-null  uint8
7   valorfatura_2        25984 non-null  uint8
8   valorfatura_3        25984 non-null  uint8
9   valorfatura_4        25984 non-null  uint8
10  p_gastoalim_2         25984 non-null  uint8
11  p_gastoalim_3         25984 non-null  uint8
12  p_gastoalim_4         25984 non-null  uint8
13  regiaorisco_2         25984 non-null  uint8
14  regiaorisco_3         25984 non-null  uint8
15  regiaorisco_4         25984 non-null  uint8
16  rendamensal_2         25984 non-null  uint8
17  rendamensal_3         25984 non-null  uint8
18  rendamensal_4         25984 non-null  uint8
dtypes: int64(1), uint8(18)
memory usage: 659.9 KB
```

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência

Amostra proporcional de categorias 0 e 1

	cliente	%
Resposta		
0	13056	50.25
1	12928	49.75

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência

```
import statsmodels.api as sm
```

```
X_ = sm.add_constant(X_train)
logit_model=sm.Logit(y_train,X_)
result=logit_model.fit()
print(result.summary2())
```

Optimization terminated successfully.
Current function value: 0.483152
Iterations 6

(1)

Results: Logit

```
=====
Model:                               Logit                Pseudo R-squared:    0.303
Dependent Variable:                   Resposta           AIC:                17607.1474
Date:                                2023-03-23 15:16      BIC:                17732.0836
No. Observations:                     18188             Log-Likelihood:     -8787.6
Df Model:                             15                LL-Null:            -12607.
Df Residuals:                         18172             LLR p-value:        0.0000
Converged:                            1.0000             Scale:              1.0000
No. Iterations: (2)                   6.0000             (3)
=====
```

	Coef.	Std.Err.	z	P> z	[0.025	0.975]
const	0.5846	0.0708	8.2602	0.0000	0.4459	0.7233
atrasos_2	0.5013	5691891.6909	0.0000	1.0000	-11155902.2168	11155903.2193
atrasos_3	1.2169	nan	nan	nan	nan	nan
atrasos_4	1.4690	3695323.3497	0.0000	1.0000	-7242699.2076	7242702.1457
temporel_2	-0.1041	0.0523	-1.9910	0.0465	-0.2065	-0.0016
temporel_3	-1.0233	0.0561	-18.2560	0.0000	-1.1332	-0.9135
temporel_4	-1.0490	0.0501	-20.9280	0.0000	-1.1472	-0.9507
valorfatura_2	-0.0217	0.0344	-0.6312	0.5279	-0.0891	0.0457
valorfatura_3	-0.3817	0.0438	-8.7191	0.0000	-0.4675	-0.2959
valorfatura_4	-0.5699	0.0345	-16.5249	0.0000	-0.6375	-0.5023
p_gastoalim_2	-0.2694	0.0523	-5.1494	0.0000	-0.3720	-0.1669
p_gastoalim_3	-1.1579	0.0459	-25.2020	0.0000	-1.2480	-1.0679
p_gastoalim_4	-2.3206	0.0483	-48.0367	0.0000	-2.4153	-2.2259



(1) Valor do ponto de corte

(2) Coeficientes do modelo.

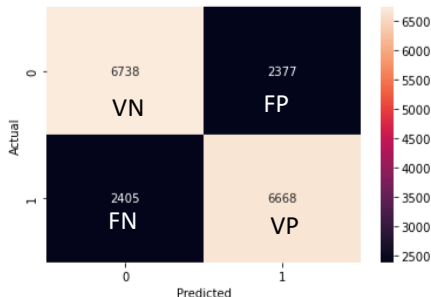
(3) Resultado do p-value do teste de hipótese para seleção da variável.

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência



Matriz de confusão



```
predictions = logmodel.predict(X_train)
print(classification_report(y_train['Resposta'], predictions))
```

	precision	recall	f1-score	support
0	0.74	0.74	0.74	9115
1	0.74	0.73	0.74	9073
accuracy			0.74	18188
macro avg	0.74	0.74	0.74	18188
weighted avg	0.74	0.74	0.74	18188

VN: Verdadeiro Negativo

FP: Falso Positivo

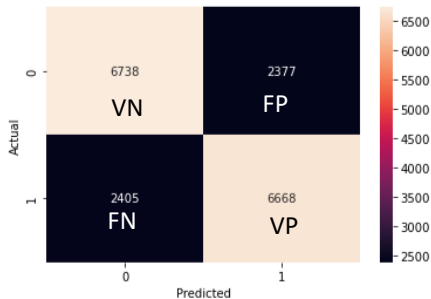
FN: Falso Negativo

VP: Verdadeiro Positivo

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência

Matriz de confusão



```
predictions = logmodel.predict(X_train)
print(classification_report(y_train['Resposta'], predictions))
```

	precision	recall	f1-score	support
0	0.74	0.74	0.74	9115
1	0.74	0.73	0.74	9073
accuracy			0.74	18188
macro avg	0.74	0.74	0.74	18188
weighted avg	0.74	0.74	0.74	18188

VN: Verdadeiro Negativo

FP: Falso Positivo

FN: Falso Negativo

VP: Verdadeiro Positivo

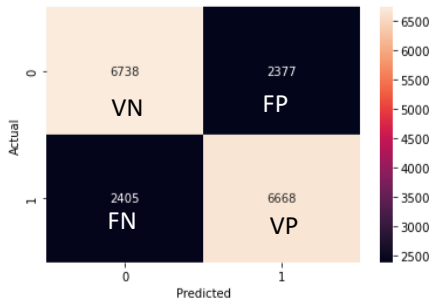
$$\text{Precision}(0) = \frac{6738}{6738 + 2405} = 0.74$$

$$\text{Precision}(1) = \frac{6668}{6668 + 2377} = 0.74$$

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência

Matriz de confusão



```
predictions = logmodel.predict(X_train)
print(classification_report(y_train['Resposta'], predictions))
```

	precision	recall	f1-score	support
0	0.74	0.74	0.74	9115
1	0.74	0.73	0.74	9073
accuracy			0.74	18188
macro avg	0.74	0.74	0.74	18188
weighted avg	0.74	0.74	0.74	18188

VN: Verdadeiro Negativo

FP: Falso Positivo

FN: Falso Negativo

VP: Verdadeiro Positivo

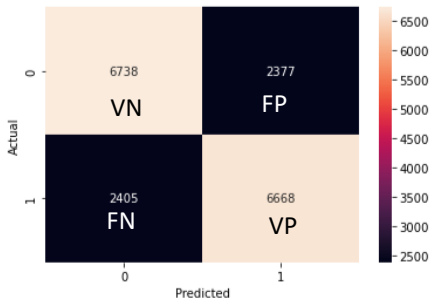
$$\text{Recall (0)} = \frac{6738}{6738 + 2377} = 0.74$$

$$\text{Recall (1)} = \frac{6668}{6668 + 2405} = 0.73$$

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência

Matriz de confusão



```
predictions = logmodel.predict(X_train)
print(classification_report(y_train['Resposta'], predictions))
```

	precision	recall	f1-score	support	
0	0.74	0.74	0.74	9115	50%
1	0.74	0.73	0.74	9073	50%
accuracy			0.74	18188	
macro avg	0.74	0.74	0.74	18188	
weighted avg	0.74	0.74	0.74	18188	

VN: Verdadeiro Negativo

FP: Falso Positivo

FN: Falso Negativo

VP: Verdadeiro Positivo

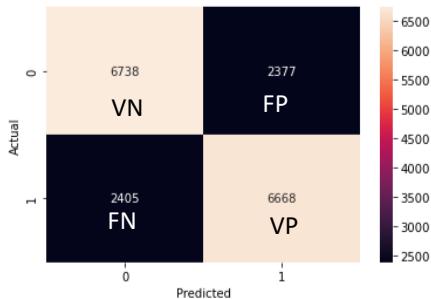
$$\begin{aligned} \text{F1-score}(0) &= 2 * (\text{precision}(0) + \text{recall}(0)) / (\text{precision}(0) + \text{recall}(0)) \\ &= (2 * (0.74 + 0.74)) / ((0.74 + 0.74)) = 0.74 \end{aligned}$$

$$\begin{aligned} \text{F1-score}(1) &= 2 * (\text{precision}(1) + \text{recall}(1)) / (\text{precision}(1) + \text{recall}(1)) \\ &= (2 * (0.74 + 0.74)) / ((0.74 + 0.74)) = 0.74 \end{aligned}$$

Segmento: Cartões de Crédito

A área de crédito deseja avaliar a propensão ao risco de seus clientes e implementar políticas de redução da inadimplência

Matriz de confusão



```
predictions = logmodel.predict(X_train)
print(classification_report(y_train['Resposta'], predictions))
```

	precision	recall	f1-score	support
0	0.74	0.74	0.74	9115
1	0.74	0.73	0.74	9073
accuracy			0.74	18188
macro avg	0.74	0.74	0.74	18188
weighted avg	0.74	0.74	0.74	18188

VN: Verdadeiro Negativo

FP: Falso Positivo

FN: Falso Negativo

VP: Verdadeiro Positivo

$$\text{Accuracy} = \frac{6016738 + 6668}{18188} = 0.66$$

$$\begin{aligned} \text{macro average} &= (\text{precision of class 0} + \text{precision of class 1}) / 2 \\ &= (0.74 + 0.74) / 2 = 0.74 \end{aligned}$$

Compute the F1 score, also known as balanced F-score or F-measure.

The F1 score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

$$F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

In the multi-class and multi-label case, this is the average of the F1 score of each class with weighting depending on the average parameter.

Média Harmônica

$$M_h = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

M_h: média harmônica

n: quantidade de elementos

$$M_h = \frac{2}{\frac{1}{80} + \frac{1}{120}} = \frac{2}{\frac{3+2}{240}} = \frac{2}{\frac{5}{240}} = 2 \cdot \frac{240}{5}$$

$$M_h = 2 \cdot \frac{240}{5} = \frac{480}{5} = 96 \text{ km/h}$$

Um carro realiza um percurso duas vezes. Na ida, ele faz o percurso com uma velocidade $v_1 = 80$ km/h. Na volta, ele realiza o mesmo percurso com velocidade de $v_2 = 120$ km/h. Qual foi a velocidade média ao juntar-se ida e volta?

Note que a distância é a mesma, para a ida e para a volta, o que muda é a velocidade e, consequentemente, o tempo.

Se eu aumento a velocidade, o tempo que eu levo para percorrer uma mesma distância diminuirá, logo, essas grandezas são inversamente proporcionais.

Interpretação da estatística weighted average :

O modelo está adequado?

Classification Report :				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	28432
1	0.02	0.02	0.02	49
accuracy			1.00	28481
macro avg	0.51	0.51	0.51	28481
weighted avg	1.00	1.00	1.00	28481

In the case of weighted average the performance metrics are weighted accordingly:

$$score_{weighted-avg} = 0.998 \cdot score_{class\ 0} + 0.002 \cdot score_{class\ 1}$$

However, macro avg is not weighted and therefore

$$score_{macro-avg} = 0.5 \cdot score_{class\ 0} + 0.5 \cdot score_{class\ 1}$$

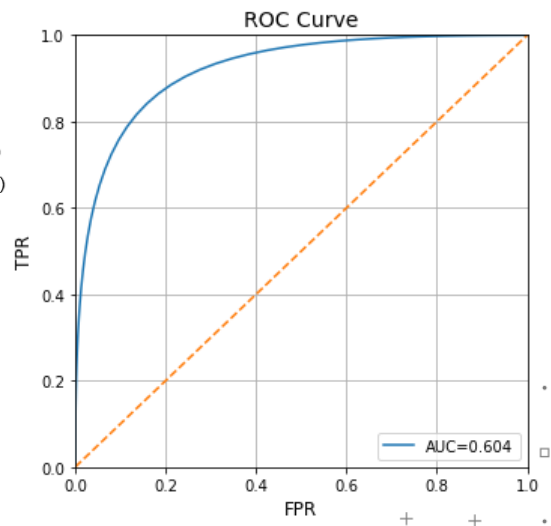
ANÁLISE DE REGRESSÃO LOGÍSTICA

Qualificação do Ajuste do Modelo

		Previsão do modelo		Total
		y=1	y=0	
Obs.	y=1	n1	n2	n1+n2
	y=0	n3	n4	n3+n4

Sensibilidade = $n1 / (n1+n2)$

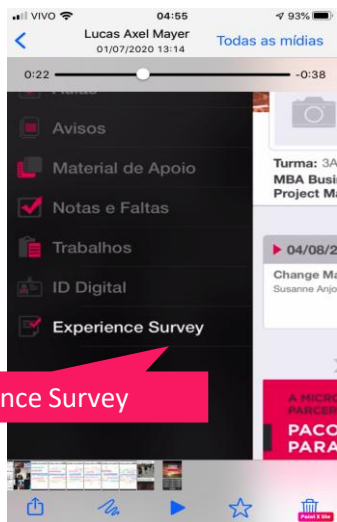
Especificidade = $n4 / (n3+n4)$



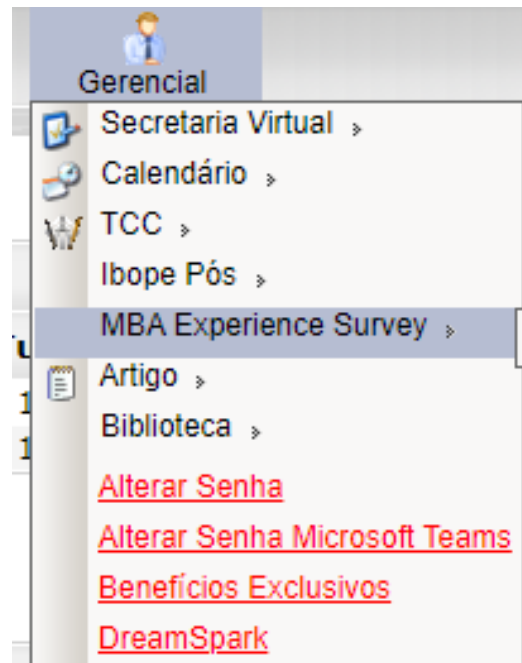
O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)



Experience Survey



A grande finalidade do
conhecimento não é conhecer,
mas agir.

T. Huxley

OBRIGADO



/ Regina T. I. Bernal

FIAP

Copyright © 2023 | Professora Dra. Regina Tomie Ivata Bernal
Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente
proibido sem consentimento formal, por escrito, do professor/autor.

FIAP