

FIAP

NBA

ADELAIDE ALVES DE OLIVEIRA

PROFESSORA



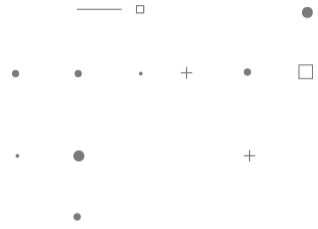
profadelaide.alves@fiap.com.br

Formação Acadêmica

- Bacharel em Estatística – UNICAMP
- Mestre em Ciências – FSP/USP

Atividades Profissionais

- Diretora Técnica Estatística da empresa **SD&W** - www.sdw.com.br
- Professora de Fundamentos Estatísticos, DataMining, Análise Preditiva e Machine Learning na FIAP dos cursos MBA: Big Data, Data Science, Business Intelligence & Analytics, Digital Data Marketing, IA & ML e Engenharia de Dados e nos Shift: People Analytics e Python Journey



Conceitos Estatísticos para IA



Introdução...(Voltando)

→ ... enfim, seus dados não servem para nada até que você saiba como tirar informações deles

DESCRITIVO O que aconteceu?	DIAGNÓSTICO Por que isto aconteceu?	PREDITIVO O que acontecerá?	PRESCRITIVO O que posso fazer?
Quantos clientes temos cancelados voluntariamente? Quais os tipos de produtos? Qual a região que mora? Qual o tempo é cliente da empresa?	Qual a relação entre o cancelamento e região? Quais os motivos de cancelamentos. Qual a taxa de cancelamento por safra?	Qual a probabilidade de um cliente cancelar ? Quais são os clientes que queremos reter? Qual é a segmentação de valor dos clientes de maior propensão ao cancelamento?	Lista de ações para reter o cliente que está ativo e que tem a propensão ao cancelamento a um determinado tempo. Qual o canal que vamos utilizar para cada cliente? Que ações?

ANÁLISE MULTIVARIADA

Análise Exploratória dos Dados

Análise de Discriminação de Estrutura

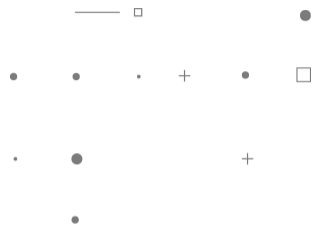
- Técnicas de dependência.
- Técnicas Multivariadas aplicáveis quando uma das variáveis **pode ser identificada como dependente (variável target)**, e as restantes como variáveis independentes.

Análise Supervisionada

Análise Estrutural

- Técnicas de Interdependência.
- Técnicas Multivariadas que procuram agrupar dados com base em semelhança, permitindo assim a interpretação das estruturas dos dados. **Não há distinção entre variáveis dependentes e independentes.**

Análise Não Supervisionada



TÉCNICAS DE DISCRIMINAÇÃO PREVISÃO E ESTIMAÇÃO

Descobertas Supervisionadas de Relações

Quando a variável target assume valores numéricos.



PREVISÃO

A escolha da técnica adequada depende:

- Horizonte de previsão.
 - Curto, médio e longo prazo.
- Acuracidade desejada.
- Relevância e disponibilidade de dados.
- Custo/benefício da previsão.
- Tempo disponível para modelagem.

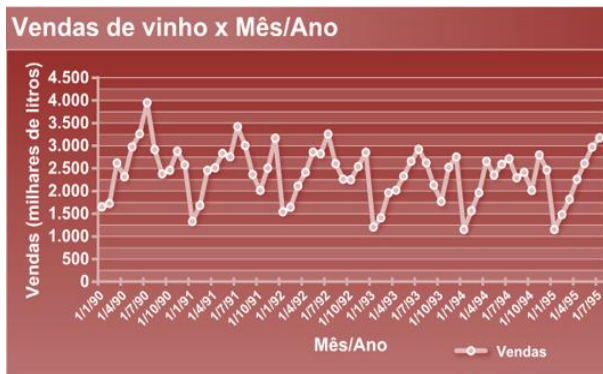
TÉCNICAS DE PREVISÃO:

TÉCNICAS QUANTITATIVAS

Quando a variável target assume valores numéricos.

Essas técnicas podem ser agrupadas em:

- Modelos de séries temporais: Enfoca os **padrões e suas mudanças**, desenvolvido por meio de sua **série histórica**.



Utilização: As técnicas quantitativas são aplicadas nas condições:

- Informações do passado disponíveis;
- Informações quantificáveis em forma numérica;
- Assumir a hipótese de que algo dos padrões do passado irá se repetir no futuro (hipótese de continuidade).

- Modelos causais: Utiliza informações refinadas e específicas **sobre relações entre elementos do sistema**.

$$\text{Qualidade do Vinho} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

- Variáveis preditoras como: tipo do vinho, acidez, ph, açúcar, ...

TÉCNICAS DE PREVISÃO REGRESSÃO

• O Modelo Causal permite:

- Expressar as relações de Causa-Efeito entre variáveis;
- Entender melhor os mecanismos geradores do fato em estudo;
- Simular situações de forma a se avaliar o seu impacto na previsão;
- Analisar situações independentes do tempo.

MODELO DE REGRESSÃO:

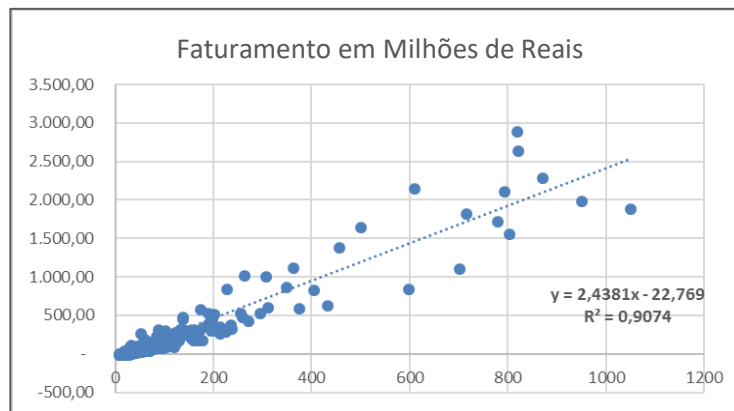
Esse modelo relaciona, funcionalmente, uma variável dependente às suas possíveis variáveis explicativas.

- Eficácia de propaganda sobre as vendas
- Número de acidentes pela velocidade desenvolvida
- Prever o tempo gasto no caixa de um supermercado em função do valor de compra
- Satisfação do Cliente em função do tempo de relacionamento e intensidade de uso

ANÁLISE DE REGRESSÃO

Técnica Estatística que relaciona funcionalmente uma variável dependente às suas possíveis variáveis explicativas. Em outras palavras, consiste na obtenção de uma equação que tenta explicar variação da variável dependente pela variação do(s) nível(is) da(s) variável(is) independente(s).

- Modelo Linear a Duas Variáveis.
- Modelo Linear Múltiplo.



Fonte:Abras

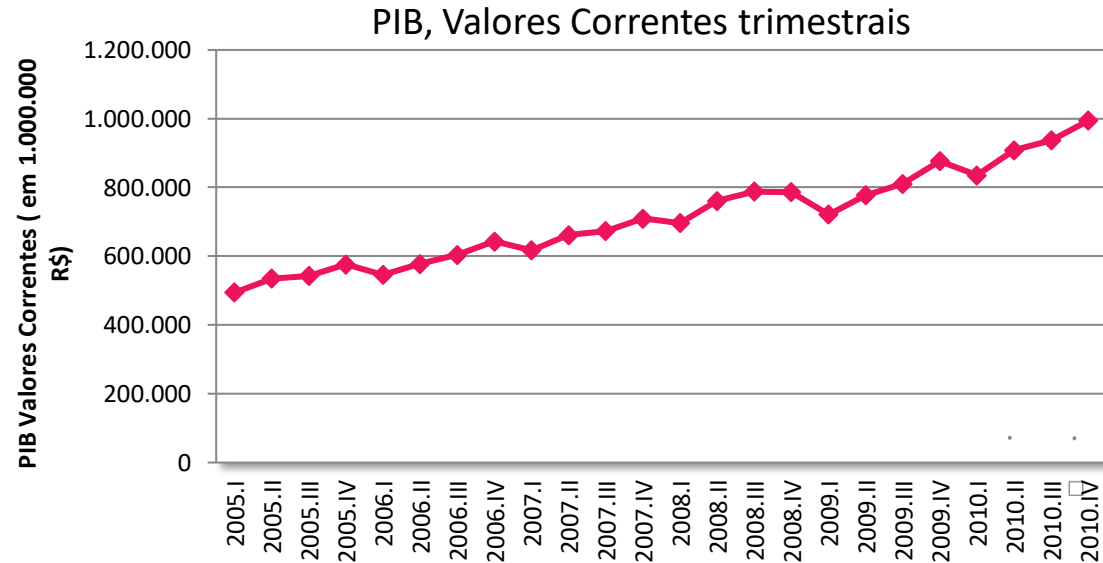
TÉCNICAS DE PREVISÃO SÉRIES TEMPORAIS

- Investigar o mecanismo gerador da série:

- Descrever o comportamento da série;

- Procurar periodicidades relevantes;

- Fazer previsões.

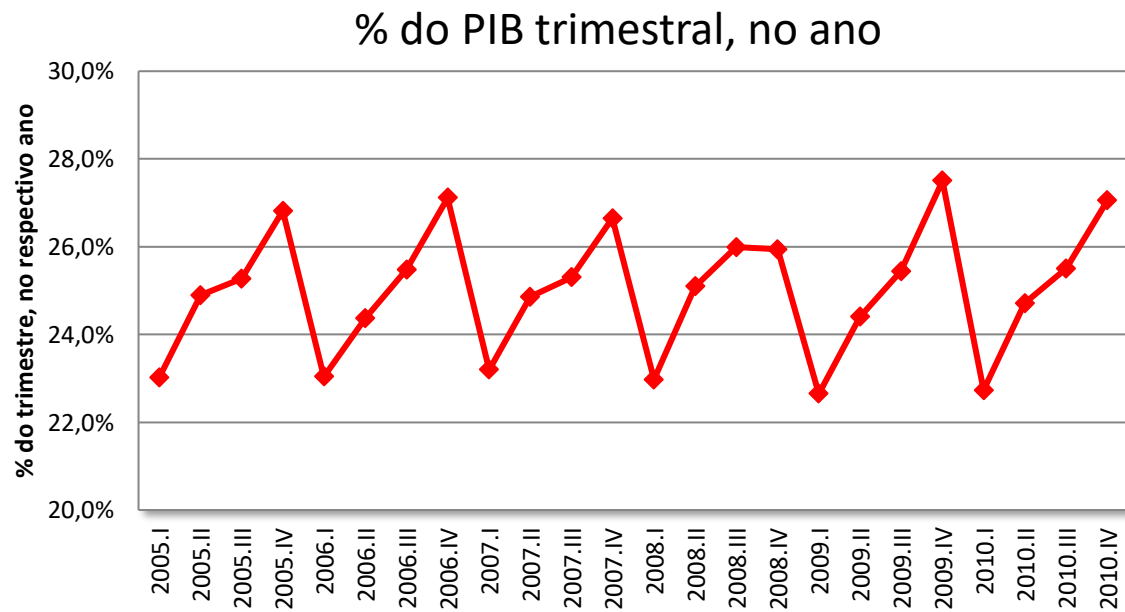


TÉCNICAS DE PREVISÃO SÉRIES TEMPORAIS

Conjunto de observações ordenadas no tempo a intervalos regulares

- vendas mensais de telefones celulares no Brasil;
- estimativas trimestrais do PIB;
- preços mensais do petróleo no mercado internacional;
- índices diários da bolsa de valores de SP;
- valores diários de temperatura na cidade de SP.

TÉCNICAS DE PREVISÃO SÉRIES TEMPORAIS



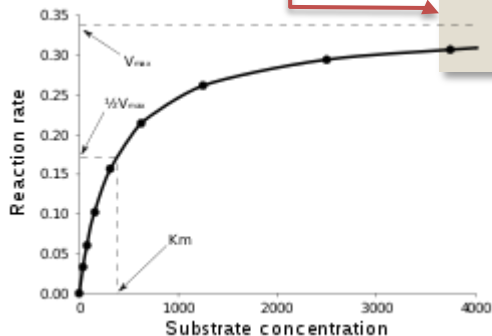
TÉCNICAS DE PREVISÃO MODELO DE REGRESSÃO

REGRESSÃO

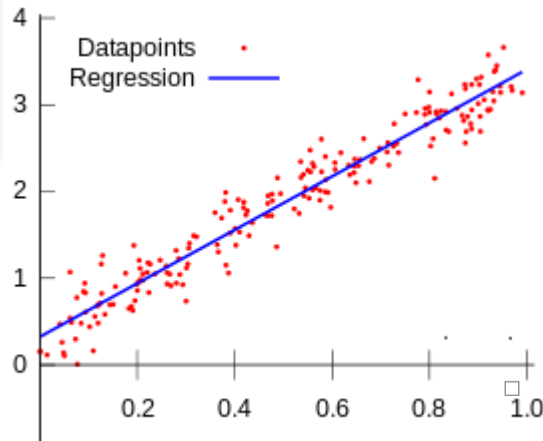
Técnica estatística que relaciona, funcionalmente, uma variável dependente às suas possíveis variáveis explicativas

Não-Linear

Linear



- SIMPLES: uma variável explicativa
- MÚLTIPLA: duas ou mais variáveis explicativas

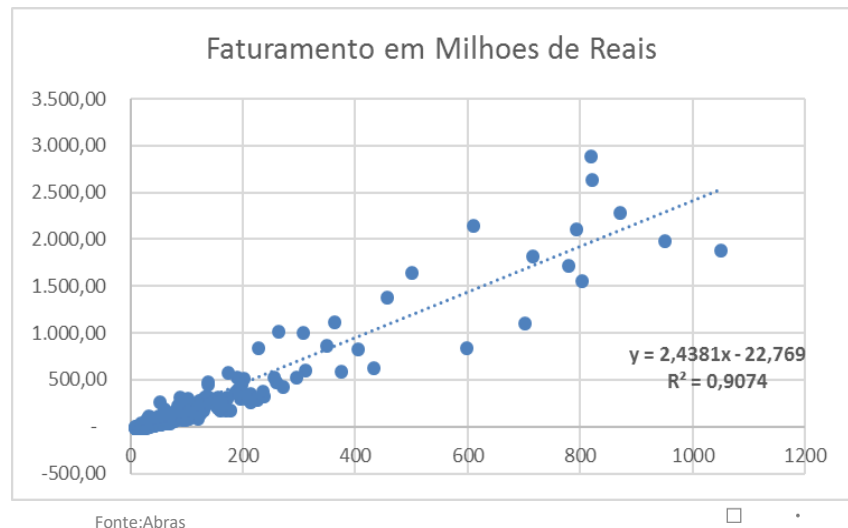


ANÁLISE DE REGRESSÃO

- Técnica Estatística que relaciona funcionalmente, uma variável dependente as suas possíveis variáveis explicativas. Em outras palavras consiste na obtenção de uma equação que tenta explicar variação da variável dependente pela variação do(s) nível(is) da(s) variável(is) independente(s).

❑ Modelo Linear à Duas Variáveis

❑ Modelo Linear Múltiplo



MODELO DE REGRESSÃO

O Modelo que relaciona Y com várias variáveis independentes

- Modelo Linear Simples: $Y = B_0 + B_1X + e$

X = variáveis independentes

Y = variável dependente

B_0 = constante

B_1 = coeficientes de regressão

- Modelo Linear Múltiplo: $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \dots + B_nX_n + e$

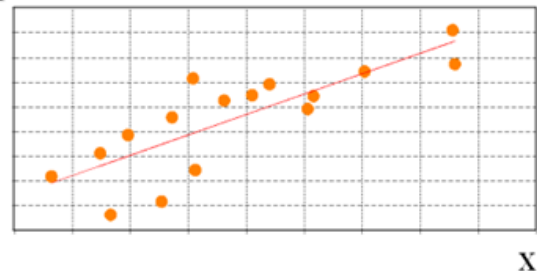
$X_1, X_2, X_3, \dots, X_n$ = variáveis independentes

Y = variável dependente

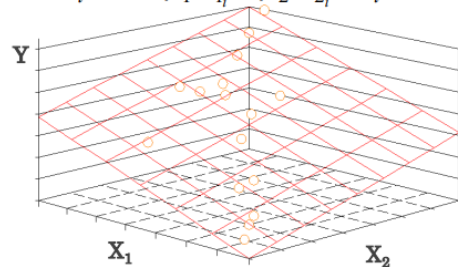
B_0 = constante

$B_1, B_2, B_3, \dots, B_n$ = coeficientes de regressão associados às n variáveis

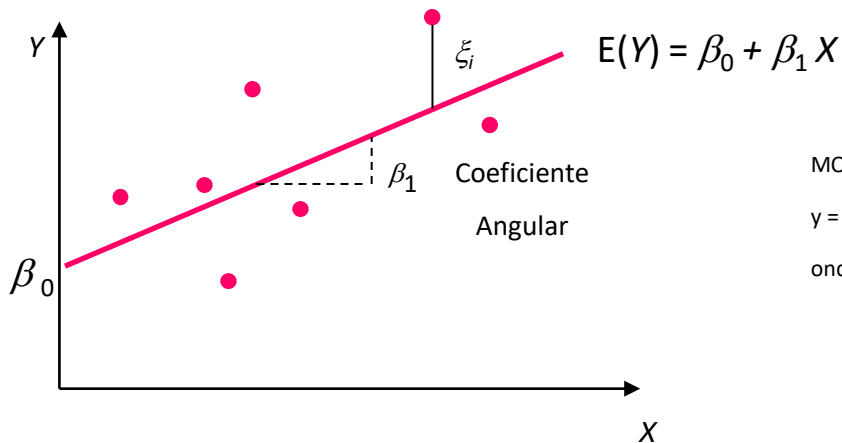
$$Y_i = \alpha + \beta X_i + e_i$$



$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i$$



MODELO DE REGRESSÃO LINEAR SIMPLES



MODELO PROBABILÍSTICO:

y = Componente Determinístico + Erro Aleatório

onde y é a variável dependente

Escrever a equação linear
envolve dois parâmetros:

O Intercepto de y

A inclinação da reta

➔ Reta Ajustada

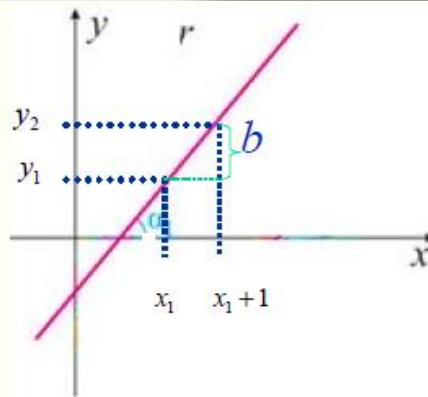
$$\hat{Y} = a + bX$$

MODELO DE REGRESSÃO LINEAR SIMPLES

→ Reta Ajustada

$$\hat{Y} = a + bX$$

$$\begin{aligned} \text{tag}(\alpha) &= \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_2 - y_1}{x_1 + 1 - x_1} \\ &= y_2 - y_1 = b \end{aligned}$$



Interpretação de b :

Para cada aumento de uma unidade em X , tem se um aumento, em média, de b unidades em Y

MODELO DE REGRESSÃO LINEAR SIMPLES

→ Reta Ajustada

$$\hat{Y} = a + bX$$

- Método de Mínimos Quadrados -

Cálculos dos
coeficientes
 a e b :

$$b = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{(n-1) S_X^2}$$

$$a = \bar{Y} - b \bar{X}$$

MODELO DE REGRESSÃO LINEAR

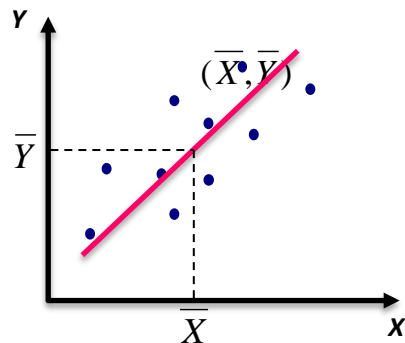
• Propriedades da equação de regressão

$$1) \sum_{i=1}^n e_i = 0$$

$$2) \sum_{i=1}^n e_i^2 \text{ é mínimo}$$

$$3) \sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$$

4) A reta de regressão passa sempre pelo ponto



MODELO DE REGRESSÃO LINEAR

- O critério de Mínimos Quadrados é um bom ajuste:
 - Escolhe a reta que minimiza a soma dos quadrados dos desvios;
 - As distribuições amostrais são conhecidas;
 - Sob certas condições, as distribuições amostrais dos estimadores de mínimos quadrados de B_0 e B_1 tem menores desvios padrões do que qualquer outro tipo de estimadores.

MODELO DE REGRESSÃO LINEAR

- Suposições necessárias:

- $E(e) = 0$
- $Var(e) = \sigma^2$
- Distribuição de Probabilidade é Normal
- O Erro associado a qualquer observação é independente
- ➔ Há técnicas baseadas na análise dos resíduos para detectar quando uma ou mais suposições foram violadas

MODELO DE REGRESSÃO LINEAR

- Indicadores de Análise de Adequacidade do Modelo:

- Análise de Variância(ANOVA) - Instrumento de Análise Estatística
- R^2 - Coeficiente de Determinação da Explicação
- Teste F - Estatística da existência do Ajuste
- Teste T - Estatística da existência do Modelo
- Resíduos - Checagem das Premissas Adotadas
- Teste DW - Checagem das Premissas Adotadas

Estatística Durbin-Watson testa a presença de autocorrelação nos erros de um modelo de regressão. A autocorrelação significa que os erros de observações adjacentes são correlacionados. Se os erros estiverem correlacionados, a regressão de mínimos quadrados pode subestimar o erro padrão dos coeficientes. Os erros padrão subestimados podem fazer com que seus preditores pareçam significativos quando eles não são. Por exemplo, os erros de um modelo de regressão dos dados de preços diários de ações podem depender da observação anterior porque o preço das ações em um dia afeta o preço do dia seguinte

- Multicolinearidade (forte correlação entre as variáveis independentes) - Checagem das Premissas Adotadas

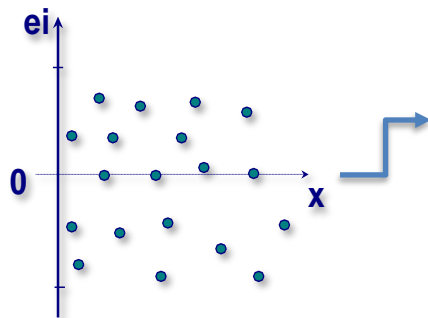
MODELO DE REGRESSÃO LINEAR

Adequacidade do Modelo: Análise de Resíduos

Forma de avaliar se as suposições colocadas no desenvolvimento do modelo não foram violadas

$$\hat{e}_i = y_i - \hat{y}_i$$

Resíduos: diferença entre o valor observado e o valor ajustado pelo modelo



Pelo gráfico de dispersão, visualizamos o comportamento dos resíduos

Formas "padronizadas"

Resíduos PADRONIZADOS:

$$\frac{e_i}{se}$$

Resíduos STUDENTIZADOS:

$$\frac{e_i}{se\sqrt{1-v_{ii}}}$$

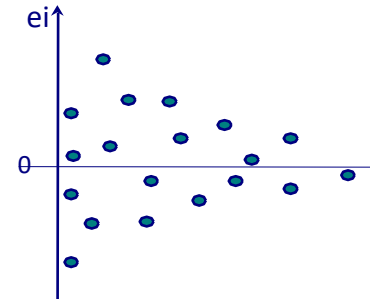
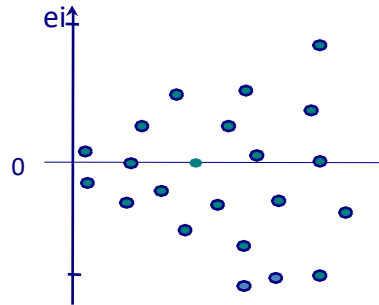
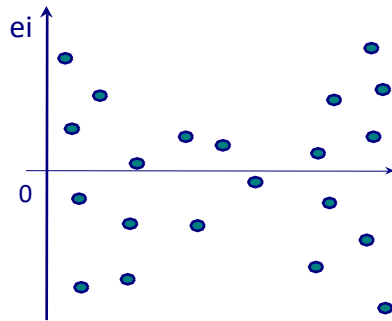
onde $v_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$

REGRESSÃO LINEAR ANÁLISE DE RESÍDUOS

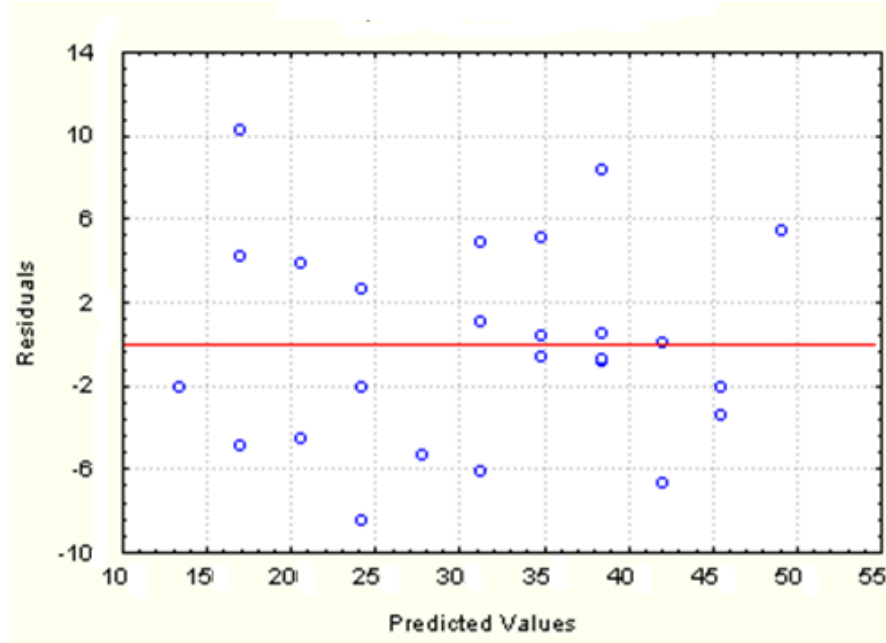
- IGUALDADE DE VARIÂNCIA

Quando o gráfico de dispersão dos Resíduos Studentizados, contra o **valor predito**, indica que a extensão dos resíduos aumentam com a magnitude dos valores preditos:

Então a suposição de igualdade da variância está violada



REGRESSÃO LINEAR ANÁLISE DE RESÍDUOS



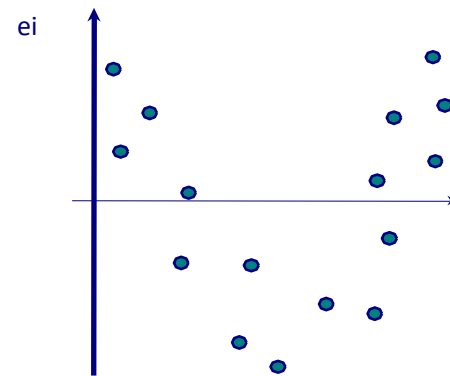
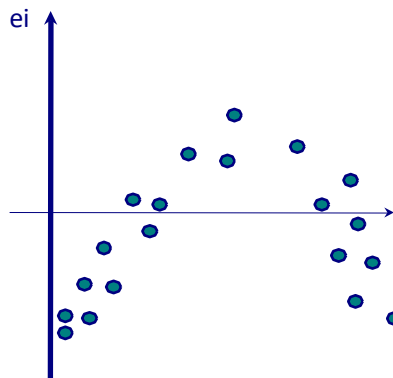
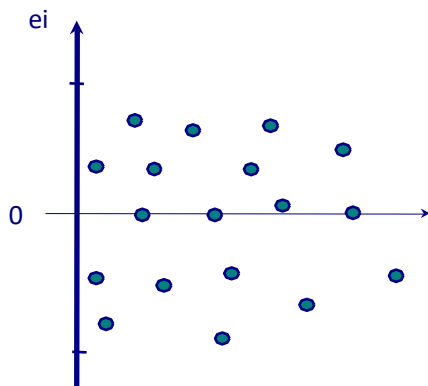
Resíduos se distribuem aleatoriamente em torno da média zero

➔ Modelo de Regressão Adequado

REGRESSÃO LINEAR ANÁLISE DE RESÍDUOS

- LINEARIDADE : Gráfico de dispersão dos valores preditos (\hat{y}) e o resíduo

⇒ os resíduos devem estar distribuídos aleatoriamente.

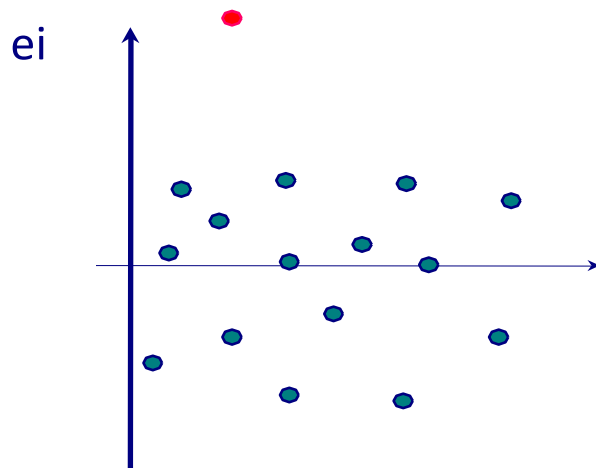


O comportamento não aleatório dos resíduos podem ser eliminados ajustando no modelo um termo quadrático.

REGRESSÃO LINEAR ANÁLISE DE RESÍDUOS

- LOCALIZANDO OS OUTLIERS:

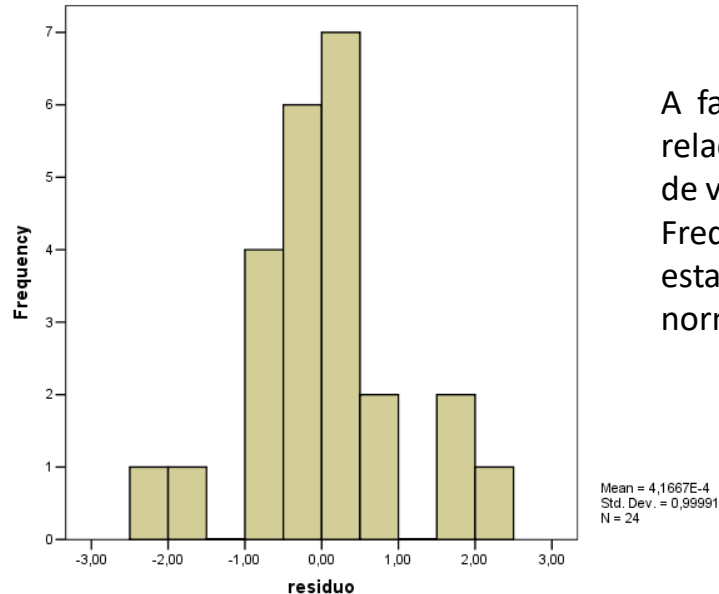
Em geral, resíduos padronizados com valores maiores que 3 são considerados outliers



REGRESSÃO LINEAR ANÁLISE DE RESÍDUOS

- NORMALIDADE

Pelo histograma dos resíduos padronizados pode-se analisar a suposição de normalidade



A falta de normalidade muitas vezes está relacionado com a falta de homogeneidade de variâncias.

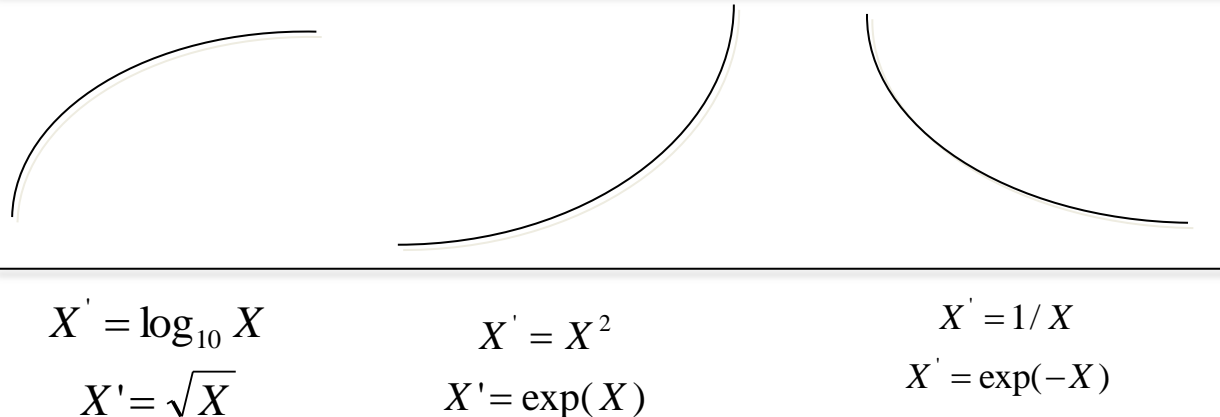
Frequentemente, a mesma transformação estabiliza a variância e aproxima para a normalidade.

REGRESSÃO LINEAR ANÁLISE DE RESÍDUOS

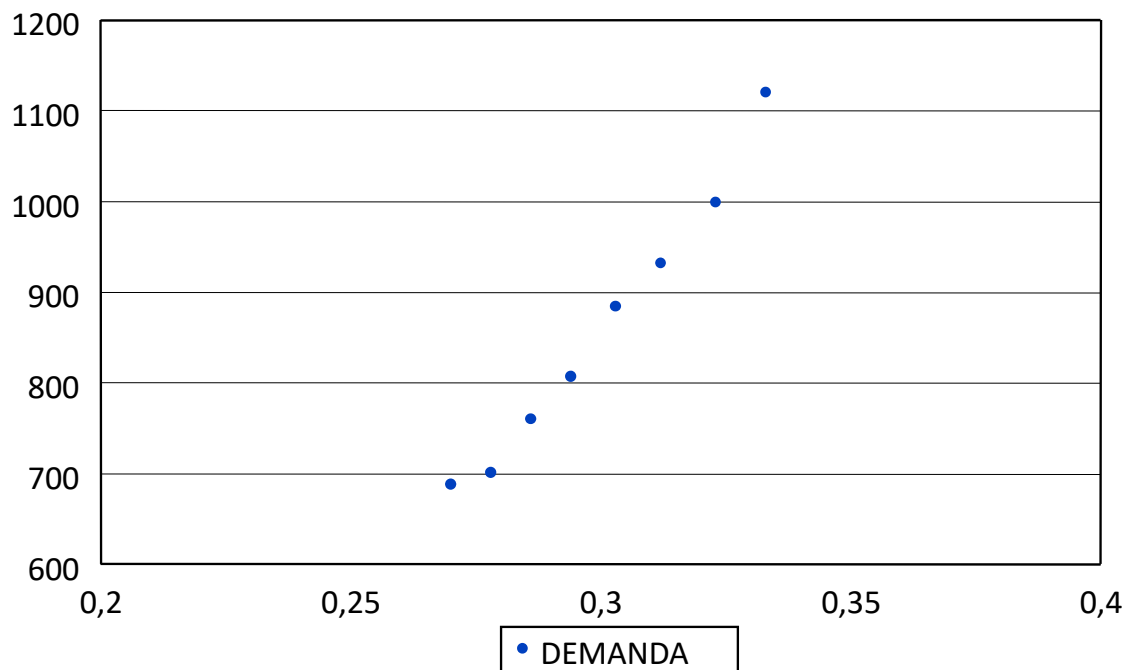
TRANSFORMAÇÃO DE VARIÁVEIS

Quando o modelo não é conhecido, pode-se escolher a transformação examinando o gráfico x e y .

Exemplos: Padrões de relação entre X e Y



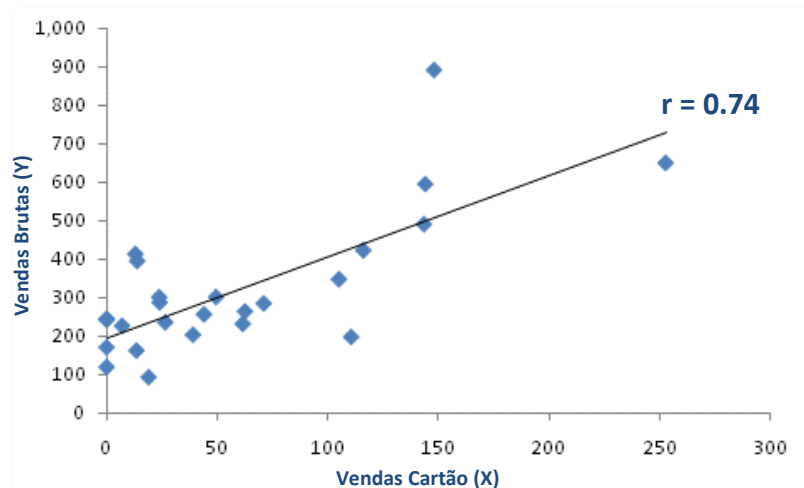
REGRESSÃO LINEAR ANÁLISE DE RESÍDUOS

TRANSFORMAÇÃO DE VARIÁVEIS ($X=1/P$)

MODELO DE REGRESSÃO LINEAR SIMPLES

- Exemplo: Prevendo as vendas diárias na loja XYZ

Dia	VendasBrutas (Y)	VendasCartão (X)
1	890.50	148.00
2	197.00	110.50
3	231.00	61.50
4	170.00	0.00
5	202.50	39.00
6	225.50	7.00
7	489.70	143.40
8	234.80	26.50
9	161.50	13.50
10	284.00	71.00
11	422.00	116.00
12	300.70	49.50
13	412.40	13.00
14	346.80	105.00
15	92.30	19.00
16	255.80	44.00
17	118.50	0.00
18	286.50	24.00
19	594.00	144.00
20	263.29	62.55
21	244.08	0.00
22	394.28	13.80
23	241.31	0.00
24	299.97	23.75
25	649.04	252.60



MODELO DE REGRESSÃO LINEAR SIMPLES

Exemplo

Resultados do Excel

<i>Estatística de regressão</i>	
R múltiplo	0.74
R-Quadrado	0.55
R-quadrado ajustado	0.53
Erro padrão	123.27
Observações	25

ANOVA

	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
Regressão	1	428912.3522	428912.3522	28.22781	0.00
Resíduo	23	349477.4932	15194.67362		
Total	24	778389.8455			

	<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
Interseção	194.93	34.13	5.71	0.00	124.33	265.52
VendasCartão (X)	2.11	0.40	5.31	0.00	1.29	2.93

Exemplo

MODELO DE REGRESSÃO LINEAR SIMPLES

Resultados do Excel

Estatística de regressão	
R múltiplo	0.74
R-Quadrado	0.55
R-quadrado ajustado	0.53
Erro padrão	123.27
Observações	25

A relação linear entre as duas variáveis é medida pelo coeficiente de correlação

R-quadrado da regressão, que mede a proporção da variabilidade em Y que é explicada por X. É uma função direta da correlação entre as variáveis

é uma medida semelhante ao R-quadrado mas que, ao contrário deste, não aumenta com a inclusão de variáveis independentes não significativas

Erro padrão: mede a dispersão dos valores observados em relação a equação da reta

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Soma de Quadrados Total
Soma de Quadrados Residual
Soma de Quadrados da Regressão

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1	428912.3522	428912.3522	28.22781	0.00
Resíduo	23	349477.4932	15194.67362		
Total	24	778389.8455			

A estatística F serve para testar quanto o modelo de regressão ajusta os dados. Se a probabilidade associada com F é pequena, a hipótese que $R^2_{pop} = 0$ é rejeitada.

	Coeficientes	Erro padrão	Stat t (3)	valor-P (4)	95% inferiores (5)	95% superiores
(1) Interseção	194.93	34.13	5.71	0.00	124.33	265.52
(2) VendasCartão (X)	2.11	0.40	5.31	0.00	1.29	2.93

(1) Parâmetro B_0 (intercepto)

(2) Parâmetro B_1 (inclinação da reta)

(3) Teste de hipóteses dos parâmetros B_0 e B_1

(4) Nível descritivo do teste de hipóteses (3)

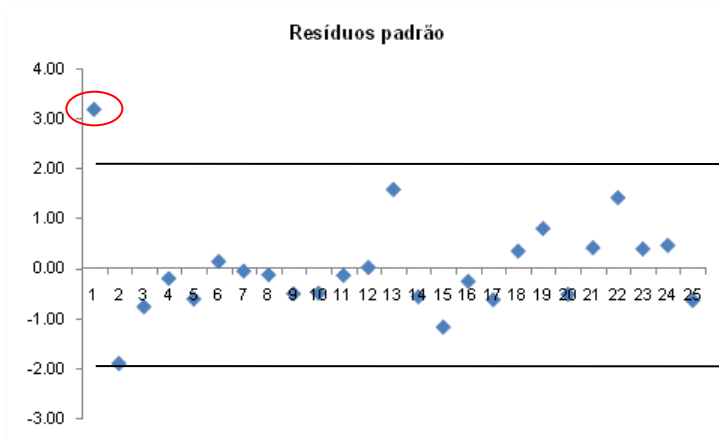
(5) Intervalo de confiança da estimativa do parâmetro

MODELO DE REGRESSÃO LINEAR SIMPLES

Análise de resíduos

Observação	Previsto(a) Vendas Brutas (Y)	Resíduos	Resíduos padrão
1	506.75	383.75	3.18
2	427.74	-230.74	-1.91
3	324.50	-93.50	-0.77
4	194.93	-24.93	-0.21
5	277.10	-74.60	-0.62
6	209.68	15.82	0.13
7	497.06	-7.36	-0.06
8	250.76	-15.96	-0.13
9	223.37	-61.87	-0.51
10	344.52	-60.52	-0.50
11	439.33	-17.33	-0.14
12	299.22	1.48	0.01
13	222.32	190.08	1.58
14	416.16	-69.36	-0.57
15	234.96	-142.66	-1.18
16	287.63	-31.83	-0.26
17	194.93	-76.43	-0.63
18	245.49	41.01	0.34
19	498.33	95.67	0.79
20	326.72	-63.43	-0.53
21	194.93	49.15	0.41
22	224.00	170.28	1.41
23	194.93	46.38	0.38
24	244.97	55.00	0.46
25	727.14	-78.10	-0.65

Exemplo

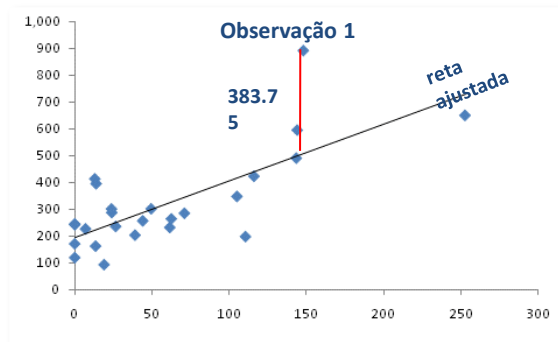


MODELO DE REGRESSÃO LINEAR SIMPLES

Análise de resíduos

Observação	Previsto(a) VendasBrutas (Y)	Resíduos	Resíduos padrão
1	506.75	383.75	3.18
2	427.74	-230.74	-1.91
3	324.50	-93.50	-0.77
4	194.93	-24.93	-0.21
5	277.10	-74.60	-0.62
6	209.68	15.82	0.13
7	497.06	-7.36	-0.06
8	250.76	-15.96	-0.13
9	223.37	-61.87	-0.51
10	344.52	-60.52	-0.50
11	439.33	-17.33	-0.14
12	299.22	1.48	0.01
13	222.32	190.08	1.58
14	416.16	-69.36	-0.57
15	234.96	-142.66	-1.18
16	287.63	-31.83	-0.26
17	194.93	-76.43	-0.63
18	245.49	41.01	0.34
19	498.33	95.67	0.79
20	326.72	-63.43	-0.53
21	194.93	49.15	0.41
22	224.00	170.28	1.41
23	194.93	46.38	0.38
24	244.97	55.00	0.46

Exemplo



MODELO DE REGRESSÃO LINEAR SIMPLES

Novo ajuste

Estatística de regressão	
R múltiplo	0.77
R-Quadrado	0.59
R-quadrado ajustado	0.57
Erro padrão	90.94
Observações	24

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1	257779.1	257779.1	31.17137	0.000
Resíduo	22	181934.3	8269.739		
Total	23	439713.4			

	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	201.26	25.22	7.98	0.00	148.97	253.55
VendasCartão (X)	1.71	0.31	5.58	0.00	1.07	2.34

Exemplo

MODELO DE REGRESSÃO LINEAR SIMPLES

Exemplo

Análise de resíduos

Observação	Previsto(a) Vendas Brutas (Y)	Resíduos	Resíduos padrão
1	389.89	-192.89	-2.17
2	306.24	-75.24	-0.85
3	201.26	-31.26	-0.35
4	267.83	-65.33	-0.73
5	213.21	12.29	0.14
6	446.05	43.65	0.49
7	246.50	-11.70	-0.13
8	224.30	-62.80	-0.71
9	322.46	-38.46	-0.43
10	399.28	22.72	0.26
11	285.76	14.94	0.17
12	223.45	188.95	2.12
13	380.50	-33.70	-0.38
14	233.69	-141.39	-1.59
15	276.37	-20.57	-0.23
16	201.26	-82.76	-0.93
17	242.23	44.27	0.50
18	447.07	146.93	1.65
19	308.03	-44.74	-0.50
20	201.26	42.82	0.48
21	224.82	169.46	1.91
22	201.26	40.05	0.45
23	241.80	58.17	0.65
24	632.46	16.58	0.19

Há necessidade de excluir mais alguma observação?

Corresponde ao Dia = 2

excluir da análise

Corresponde ao Dia = 13

MODELO DE REGRESSÃO LINEAR SIMPLES

Exemplo

Novo ajuste

Estatística de regressão	
R múltiplo	0.86
R-Quadrado	0.75
R-quadrado ajustado	0.74
Erro padrão	72.41
Observações	22

ANOVA					
	gl	SQ	MQ	F	F de significação
Regressão	1	311497.8	311497.8	59.40548	0.00
Resíduo	20	104871.735	5243.587		
Total	21	416369.535			

	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	189.12	20.73	9.12	0.00	145.87	232.37
VendasCartão (X)	1.93	0.25	7.71	0.00	1.41	2.45

Observação	Previsto(a) VendasBrutas (Y)	Resíduos	Resíduos padrão
1	307.808	-76.808	-1.087
2	189.118	-19.118	-0.271
3	264.385	-61.885	-0.876
4	202.627	22.873	0.324
5	465.869	23.831	0.337
6	240.261	-5.461	-0.077
7	215.172	-53.672	-0.759
8	326.142	-42.142	-0.596
9	412.989	9.011	0.128
10	284.649	16.051	0.227
11	391.760	-44.960	-0.636
12	225.786	-133.486	-1.889
13	274.034	-18.234	-0.258
14	189.118	-70.618	-0.999
15	235.436	51.064	0.723
16	467.027	126.973	1.797
17	309.834	-46.544	-0.659
18	189.118	54.962	0.778
19	215.751	178.529	2.526
20	189.118	52.192	0.739
21	234.953	65.017	0.920
22	676.616	-27.576	-0.390

Há necessidade de excluir mais alguma observação?

excluir

MODELO DE REGRESSÃO LINEAR SIMPLES



Modelo Final

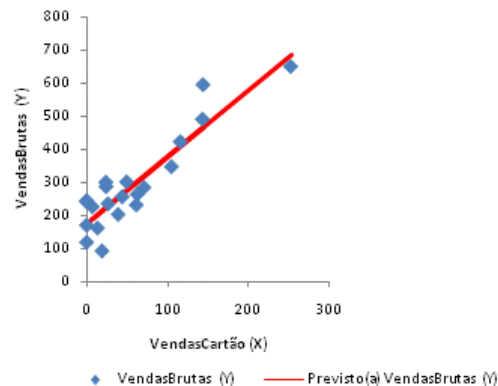
Estatística de regressão	
R múltiplo	0.91
R-Quadrado	0.83
R-quadrado ajustado	0.82
Erro padrão	61.02
Observações	21

ANOVA					
	gl	SQ	MQ	F	F de significação
Regressão	1	335463.2	335463.2	90.1	0.00
Resíduo	19	70745.9	3723.5		
Total	20	406209.1			

	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores
Interseção	175.19	18.07	9.70	0.00	137.37	213.01
VendasCartão (X)	2.02	0.21	9.49	0.00	1.58	2.47

Exemplo

$$\hat{y}_i = 175.19 + 2.02 \cdot x_i$$



MODELO DE REGRESSÃO LINEAR SIMPLES

Exemplo

Estimativas

Considerando que a venda com cartão crédito é de R\$100, qual será o valor total de vendas?

$$\hat{y}_i = 175.19 + 2.02 \cdot x$$

$$\hat{y}_i = 175.19 + 2.02 \cdot 100 = 377$$

❑ Como obter um intervalo de confiança para esta estimativa?

$$IC(95\%) = \text{estimativa} \pm t(1 - \alpha/2; n-2) * \sigma$$

O standard error da distribuição de um estimador \hat{y} de um valor médio, fixado x é raiz da:

$$\text{Var}(\hat{y}) = \sigma_e^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

Exemplo

MODELO DE REGRESSÃO LINEAR SIMPLES

Dia	VendasBrutas (Y)	VendasCartão (X)	$(X - \text{média})^2$
3	231.00	61.50	18.06
4	170.00	0.00	3277.56
5	202.50	39.00	333.06
6	225.50	7.00	2525.06
7	489.70	143.40	7421.82
8	234.80	26.50	945.56
9	161.50	13.50	1914.06
10	284.00	71.00	189.06
11	422.00	116.00	3451.56
12	300.70	49.50	60.06
14	346.80	105.00	2280.06
15	92.30	19.00	1463.06
16	255.80	44.00	175.56
17	118.50	0.00	3277.56
18	286.50	24.00	1105.56
19	594.00	144.00	7525.56
20	263.29	62.55	28.09
21	244.08	0.00	3277.56
23	241.31	0.00	3277.56
24	299.97	23.75	1122.25
25	649.04	252.60	38161.62

média =	57.25	
Total		81830.35

Exemplo

MODELO DE REGRESSÃO LINEAR SIMPLES

Considerando que a venda com cartão crédito é de R\$100, qual será o valor total de vendas?

$$\hat{y}_i = 175.19 + 2.02 \cdot x$$

$$\hat{y}_i = 175.19 + 2.02 \cdot 100 = 377$$

□ Intervalo de Confiança

$$IC(95\%) = 377 \pm 3,182 \cdot 61,02 \cdot \sqrt{\frac{1}{22} + \frac{(100 - 57,25)^2}{81.930,35}}$$

$$IC(95\%) = 377 \pm 1,9427 = (375; 379)$$

PREVISÃO E ESTIMAÇÃO

REGRESSÃO LINEAR MULTIVARIADA

MODELO REGRESSÃO MÚLTIPLA

O Modelo que relaciona Y com várias variáveis independentes

- Modelo Linear Simples: $Y = B_0 + B_1X + e$

X = variáveis independentes

Y = variável dependente

B_0 = constante

B_1 = coeficientes de regressão

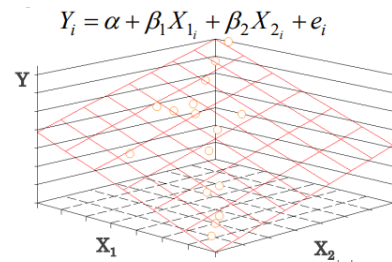
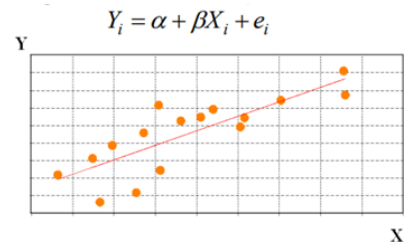
- Modelo Linear Múltiplo: $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \dots + B_nX_n + e$

$X_1, X_2, X_3, \dots, X_n$ = variáveis independentes

Y = variável dependente

B_0 = constante

$B_1, B_2, B_3, \dots, B_n$ = coeficientes de regressão associados às n variáveis



REGRESSÃO LINEAR MÚLTIPLA

Com os dados de uma amostra, podemos calcular as estimativas dos parâmetros (B) não conhecidos. Usando para isso o ajuste pelo método dos mínimos quadrados.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

Que minimiza $SSE = \sum (\bar{y} - y)^2$

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Suposições Necessárias:

$$E(e) = 0$$

$$VAR(e) = \sigma^2$$

Distribuição de probabilidade de e é normal;

O erro associado com qualquer observação é independente.

Há técnicas baseadas na análise dos resíduos para detectar quando uma ou mais suposições foram violadas.

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

COEFICIENTE DE DETERMINAÇÃO

Coeficiente de determinação Ajustado:

$$R_a^2 = 1 - \frac{n-1}{n-(K+1)}(1-R^2)$$

→ um critério para a seleção de um modelo ótimo é escolher o modelo que tem o R_a^2 máximo

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

- TESTANDO OS PARÂMETROS B'S

$$H_0: B_i = 0$$

$$H_1: B_i \neq 0$$

$$t = \frac{b_i}{Sb_i} \quad \text{com gl} = n - p$$

Quando $t > t_{\alpha/2} \Rightarrow$ região de rejeição

$$IC : \bar{b}_i \pm t_{\alpha/2} Sb_i$$

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

TESTANDO O MODELO PREDITIVO DE Y

$$H_0: B_1 = B_2 = B_3 = \dots = B_n$$

H_1 : pelo menos um B não é zero

Teste $\rightarrow F = \frac{\text{Quadrado Médio da Regressão}}{\text{Quadrado Médio dos Resíduos}}$

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

ANALISE DE VARIÂNCIA

A variabilidade total observada na variável dependente está dividido em 2 componentes

- $\hat{y}_i - y_i$ = resíduo da regressão
- $\hat{y}_i - \bar{y}$ = distância da regressão da média dos y's

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Soma de
quadrados
total

SQTot

Soma de
quadrados
residual

SQRes

Soma de
quadrados
regressão

SQReg

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

ANALISE DE VARIÂNCIA

Podemos resumir todas essas informações numa única tabela anova:

Fonte	gl	SQ	QM	F
Regressão	$p - 1$	SQReg	$QMReg = \frac{SQReg}{p - 1}$	$\frac{QMReg}{S_e^2}$
Resíduo	$n - p$	SQRes	$Se^2 = \frac{SQRes}{n - p}$	
Total	$n - 1$	SQTOT	$S^2 = \frac{SQTOT}{n - 1}$	

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- Este estudo é um caso de aplicação do método dos valores hedônicos(*), para valorar benefícios ambientais associados à proximidade a áreas verdes, existência de vista panorâmica e a localização da propriedade em rua com ou sem poluição sonora, relacionados a preços de apartamentos.

Trecho do arquivo de dados

Ordem	ValorR\$	Áream2	IA	Andar	Suítes	Vista	Dist. BM	Semruído	AV200m
1	160.000	167.81	1	5	1	1	294	1	0
2	67.000	128.80	1	6	0	0	1.505	1	0
3	190.000	217.37	1	8	1	0	251	0	1
4	110.000	180.00	12	4	1	0	245	0	0
5	70.000	120.00	15	3	1	0	956	1	0
6	75.000	160.00	18	2	0	1	85	0	1
7	95.000	155.00	5	3	1	0	1.401	1	0
8	135.000	165.00	1	2	1	1	148	0	1
9	110.000	150.00	10	4	1	0	143	0	0
10	115.000	185.00	15	5	1	0	831	0	0
11	325.669	392.40	1	4	2	0	421	1	1
12	362.400	392.40	1	8	2	0	421	1	1
13	163.798	225.60	1	2	1	0	397	1	0
14	261.250	312.82	1	3	2	0	319	1	0
15	276.870	304.35	1	5	4	0	461	1	1
16	284.626	304.35	1	7	4	0	461	1	1
17	95.000	161.00	6	3	1	0	143	0	0

(*)hedonic price models Lancaster (1966).

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

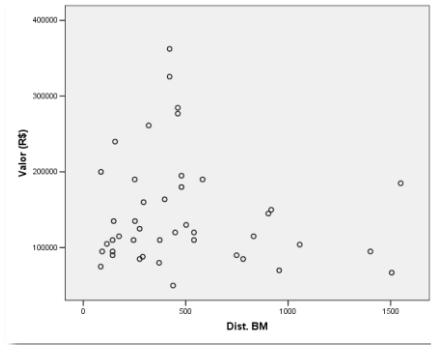
Variáveis:

- Valor do Imóvel [Valor]: Valor do imóvel
- Área [Area]: Utilizou-se a área total do apartamento em metros quadrados;
- Idade Aparente [IA]: Idade aparente em anos
- Andar [Andar]: É o número do andar do apartamento;
- Suítes [Suites]: Número de suítes;
- Vista Panorâmica [Vista]: A variável ambiental vista panorâmica é uma variável dicotômica: se o apartamento tiver vista panorâmica a variável vista assume valor igual a 1, se não tiver vista seu valor será 0;
- Sem Ruído na rua [Sem Ruído]: A variável ambiental Sem Ruído é uma variável dicotômica: se o apartamento está localizado em rua onde o nível de ruído está abaixo do que é considerado não prejudicial terá valor 1, se tiver nível de ruído acima terá valor 0;
- Distância a Avenida Beira Mar [Dist. BM]: A distância é medida em metros, pelo eixo da rua do prédio onde os apartamentos estão localizados até a Avenida Beira Mar;
- Área Verde a uma distância de 200 metros [AV 200m]: Área verde a uma distância de 200 metros assume valor igual a 1, ultrapassando 200 metros assume valor 0.

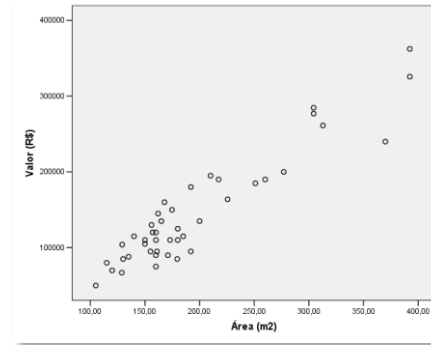
MODELO DE REGRESSÃO LINEAR MÚLTIPLA

• Análise descritiva

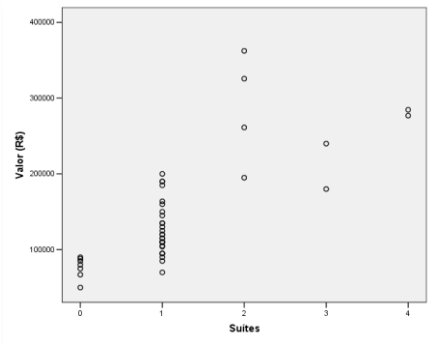
Distância a Avenida Beira Mar



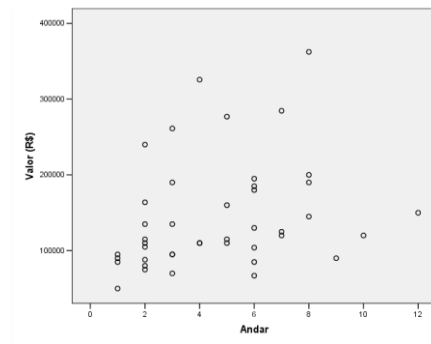
Exemplo
Área do Imóvel (em m2)



Quantidade de Suites



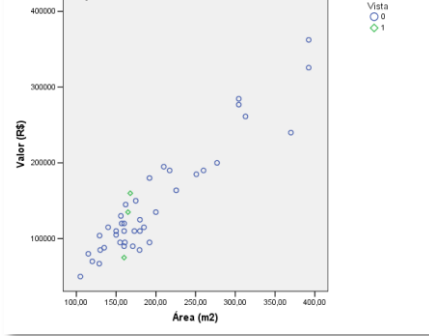
Número do Andar



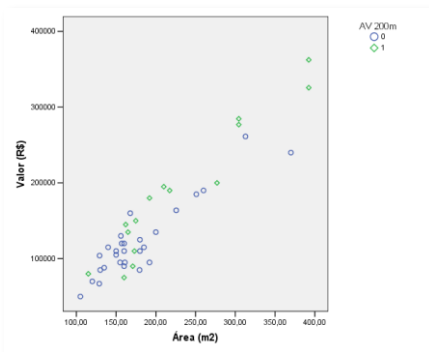
MODELO DE REGRESSÃO LINEAR MÚLTIPLA

- Análise descritiva

Vista panorâmica



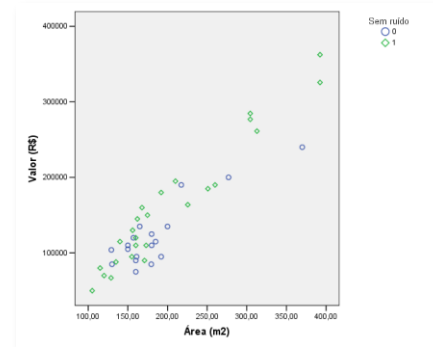
Próximo área verde



Exemplo

Área do Imóvel (em m2)
Com outras variáveis

Sem ruído



MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- Análise de correlação

Correlations						
		Valor (R\$)	Área (m2)	Andar	Suites	Dist. BM
Valor (R\$)	Pearson Correlation	1	.936**	.281	.758**	-.108
	Sig. (2-tailed)		.000	.068	.000	.490
	N	43	43	43	43	43
Área (m2)	Pearson Correlation	.936**	1	.118	.690**	-.153
	Sig. (2-tailed)	.000		.453	.000	.327
	N	43	43	43	43	43
Andar	Pearson Correlation	.281	.118	1	.197	.252
	Sig. (2-tailed)	.068	.453		.206	.103
	N	43	43	43	43	43
Suites	Pearson Correlation	.758**	.690**	.197	1	-.135
	Sig. (2-tailed)	.000	.000	.206		.387
	N	43	43	43	43	43
Dist. BM	Pearson Correlation	-.108	-.153	.252	-.135	1
	Sig. (2-tailed)	.490	.327	.103	.387	
	N	43	43	43	43	43

** . Correlation is significant at the 0.01 level (2-tailed).

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- Saída da regressão linear múltipla do SPSS

Variables Entered/Removed ^a			
Model	Variables Entered	Variables Removed	Method
1	AV 200m, Dist. BM, Vista, Área (m2), Sem ruído	.	Enter

a. All requested variables entered.

b. Dependent Variable: Valor (R\$)

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.954 ^a	.911	.899	22906.352	.911	75.787	5	37	.000

a. Predictors: (Constant), AV 200m, Dist. BM, Vista, Área (m2), Sem ruído

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- Saída da regressão linear múltipla do SPSS

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	198826510187.688	5	39765302037.538	75.787	.000 ^a
	Residual	19413935906.499	37	524700970.446		
	Total	218240446094.186	42			

a. Predictors: (Constant), AV 200m, Dist. BM, Vista, Área (m2), Sem ruído

b. Dependent Variable: Valor (R\$)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients		95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	
1	(Constant)	-43386.850	12526.681		-3.464	.001	-68768.317 -18005.383
	Área (m2)	878.181	53.553	.885	16.398	.000	769.673 986.688
	Vista	5476.052	14721.511	.020	.372	.712	-24352.562 35304.667
	Dist. BM	-3.050	10.996	-.016	-.277	.783	-25.331 19.231
	Sem ruído	20058.757	8240.080	.139	2.434	.020	3362.769 36754.745
	AV 200m	16252.429	8252.737	.109	1.969	.056	-469.206 32974.063

a. Dependent Variable: Valor (R\$)

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- Saída da regressão linear múltipla do SPSS

Excluindo as variáveis Vista, Dist BM e Av 200m

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Sem ruído Área (m2)	.	Enter

a. All requested variables entered.

b. Dependent Variable: Valor (R\$)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.948 ^a	.899	.894	23509.440

a. Predictors: (Constant), Sem ruído, Área (m2)

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- Saída da regressão linear múltipla do SPSS

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	196132696058.004	2	98066348029.002	177.433	.000 ^a
	Residual	22107750036.183	40	552693750.905		
	Total	218240446094.186	42			

a. Predictors: (Constant), Sem ruído, Área (m2)

b. Dependent Variable: Valor (R\$)

Coefficients ^a						
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	-46369.160	10861.800		-4.269	.000
	Área (m2)	911.921	50.300	.919	18.130	.000
	Sem ruído	21677.313	7317.427	.150	2.962	.005

a. Dependent Variable: Valor (R\$)

Modelo final:

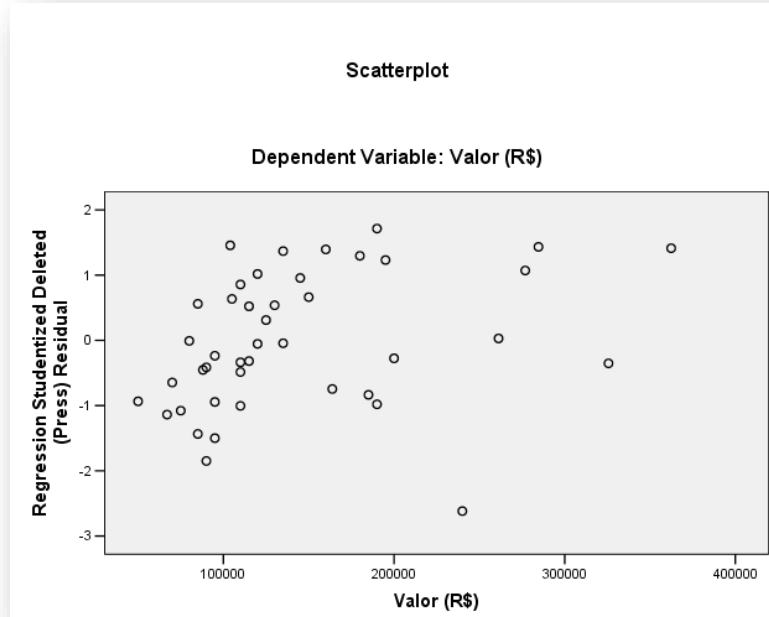
$$\hat{y}_i = -46.369,16 + 911,92 \cdot x_1 + 21.677,31 \cdot x_2$$

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- Saída da regressão linear múltipla do SPSS

Análise de resíduos



MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

ESTIMAÇÃO DOS PARÂMETROS DO MODELO NO EXEMPLO

$$Y = b_0 + b_1x_1 + b_2x_2 + e$$

Onde:

Y = Valor do imóvel (em R\$)

X_1 = Área (m²)

X_2 = Sem ruído

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

• ESTIMAÇÃO DOS PARÂMETROS DO MODELO NO EXEMPLO

As estimativas de mínimos quadrados para B's

$$b_0 = -46.369,16$$

$$b_1 = 911,92$$

$$b_2 = 21677,31$$

Então a equação que minimiza se para os dados é:

$$\hat{y}_i = -46.369,16 + 911,92 \cdot x_1 + 21.677,31 \cdot x_2$$

- Para um aumento de uma unidade na área, o valor do imóvel aumenta em média R\$ 911,92.
- Se o apartamento estiver localizado em rua onde o nível de ruído está abaixo do que é considerado não prejudicial, o valor do imóvel aumenta em média R\$ 21.677,31.

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- COEFICIENTE DE DETERMINAÇÃO

Calculando o r^2 para o modelo do valor das residências:

$$R^2 = 0.899$$

Significa que 90% da soma quadrada dos desvios de y sobre sua média é atribuído pela relação entre x_i e y .

O erro da previsão pode ser reduzido em 90% quando a equação de mínimos quadrados, ao invés de y , é usado para prever y .

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

ANALISE DE VARIÂNCIA

Testando ($\alpha = 0.05$) se o modelo com as 3 variáveis independentes é adequado para prever o preço de venda y

$$H_0: B_1 = B_2 = B_3 = 0$$

$$H_1: \text{pelo menos um } B \neq 0$$

$$F = \frac{QM_{Regressão}}{QM_{Resíduos}} = 177,43$$

Para determinar se o valor de F é significativo

$$P(F > 177,43) = 0.0001$$

Há uma forte evidência para rejeitar H_0

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

As vezes, 2 ou mais variáveis independentes usadas no modelo **poderá contribuir com informação redundante.**

Uma variável independente é correlata com uma outra variável independente.

MULTICOLINEARIDADE

☐ Altas correlações entre as variáveis independentes aumenta a probabilidade de erros no cálculo dos parâmetros B's, dos Standards erros, etc.

☐ A regressão resultante pode ser confusa

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Para detectar multicolinearidade nos modelos de regressão:

☐ Correlações significativas entre variáveis independentes do modelo;

Teste t não significativo para todos (ou quase todos) parâmetros B's, quando o teste F indica adequidade do modelo;

Sinais opostos do que é esperado para as estimativas de B's.

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

- Como resolver problemas de MULTICOLINIARIDADE:

1) Excluir variáveis independentes correlatas do modelo

2) Se a decisão for deixar essas variáveis:

- ter cuidado na previsão

3) Usar transformações e modelo de ordem maiores para reduzir o erro

4) Utilizar a técnica de sumarização, por exemplo Componentes Principais

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

- Medidas de desempenho dos modelos

Mean Error (ME):
$$ME = \frac{\sum_{i=1}^n y_i - \hat{y}_i}{n}$$

Mean Absolute Error (MAE):
$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

Root Mean Squared Error (RMSE):
$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

Mean Percentage Error (MPE):
$$MPE = \frac{\sum_{i=1}^n \frac{y_i - \hat{y}_i}{y_i} * 100}{n}$$

Mean Absolute Percentage Error (MAPE):
$$MAPE = \frac{\sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} * 100}{n}$$

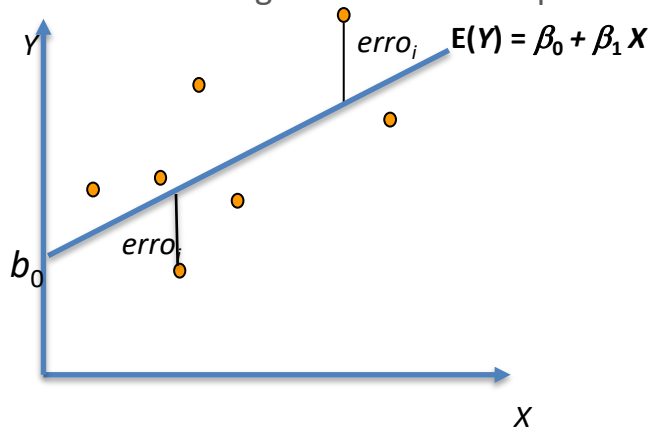
MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

- Medidas de desempenho dos modelos

→ problema de estimação ou previsão:
(variável target quantitativa):

Modelo de Regressão Linear Simples



Raiz do Erro quadrático médio(RMSE)

$$RMSE = \sqrt{\frac{5.743,3}{12}} = \sqrt{478,5} = 21,9$$

	Observado (A)	Estimado (B)	Erro (A-B)	Erro absoluto A-B	Erro^2
Id					
1	207	236	-28,7	28,7	822,8
2	289	265	24,0	24,0	576,1
3	285	272	13,5	13,5	181,9
4	292	278	14,0	14,0	195,2
5	269	285	-15,5	15,5	241,6
6	291	298	-6,6	6,6	43,2
7	331	304	26,9	26,9	724,4
8	283	307	-24,3	24,3	592,6
9	364	337	27,3	27,3	747,6
10	345	340	5,1	5,1	25,9
11	370	366	4,0	4,0	16,2
12	310	350	-39,7	39,7	1.575,1
			0,0	229,7	5.742,3

MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Exemplo

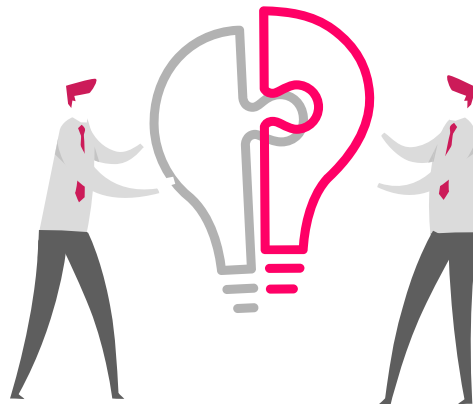
- Medidas de desempenho dos modelos

Data	Vendas (y) A	Budget (x)	Vendas estimadas (B)	Erro (A-B)	Erro absoluto A-B	Erro^2	%erro	% erro absoluto
jan/18	207	13	236	-28.7	28.7	822.8	-13.9	13.9
fev/18	289	22	265	24.0	24.0	576.1	8.3	8.3
mar/18	285	24	272	13.5	13.5	181.9	4.7	4.7
abr/18	292	26	278	14.0	14.0	195.2	4.8	4.8
mai/18	269	28	285	-15.5	15.5	241.6	-5.8	5.8
jun/18	291	32	298	-6.6	6.6	43.2	-2.3	2.3
jul/18	331	34	304	26.9	26.9	724.4	8.1	8.1
ago/18	283	35	307	-24.3	24.3	592.6	-8.6	8.6
set/18	364	44	337	27.3	27.3	747.6	7.5	7.5
out/18	345	45	340	5.1	5.1	25.9	1.5	1.5
nov/18	370	53	366	4.0	4.0	16.2	1.1	1.1
dez/18	310	48	350	-39.7	39.7	1575.1	-12.8	12.8
soma				0.0	229.7	5742.3	-7.3	79.3

MEDIDA	SOMA	n	RESULTADO
ME	0.0	12	0.0
MAE	229.7	12	19.14
RMSE	5742.3	12	21.88
MPE	-7.3	12	-1.04
MAPE	79.3	12	6.61

EXERCITANDO

Regressão Multipla



Base
Imóveis

+ + .
.
□ . . ● ●

PREVISÃO E ESTIMAÇÃO

SÉRIES TEMPORAIS

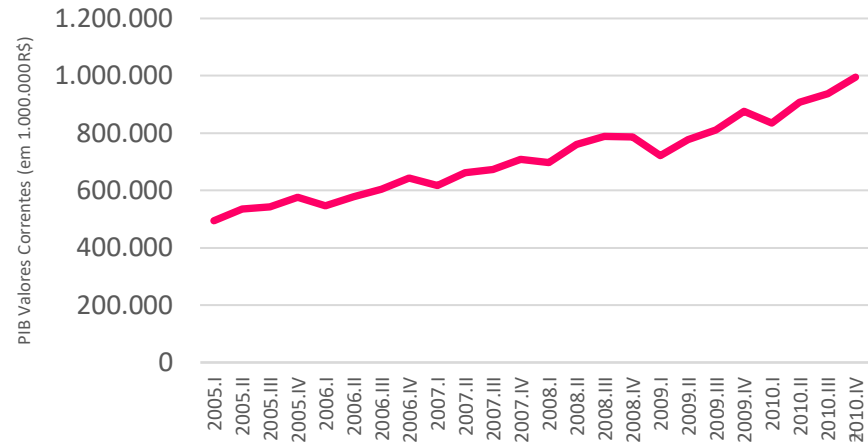
MODELOS DE SÉRIES TEMPORAIS

Considerações gerais

Uma série temporal é qualquer conjunto de observações ordenadas no tempo.

Exemplos:

- faturamento da campanha
- número de pedidos
- produção mensal
- estoque mensal

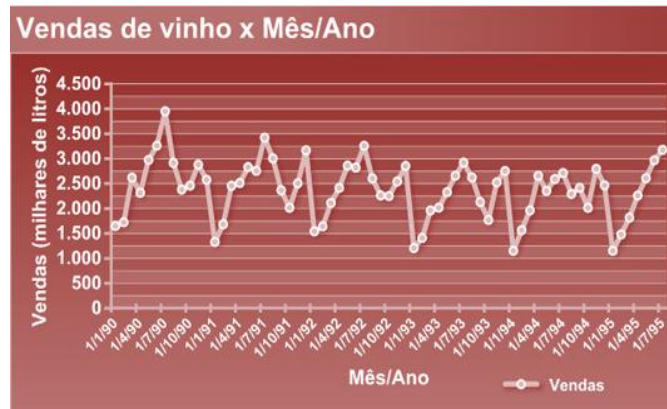


TÉCNICAS DE PREVISÃO

QUANTITATIVAS

MODELOS DE SÉRIES TEMPORAIS

O objetivo é identificar os padrões e suas mudanças, desenvolvido através de sua série histórica.

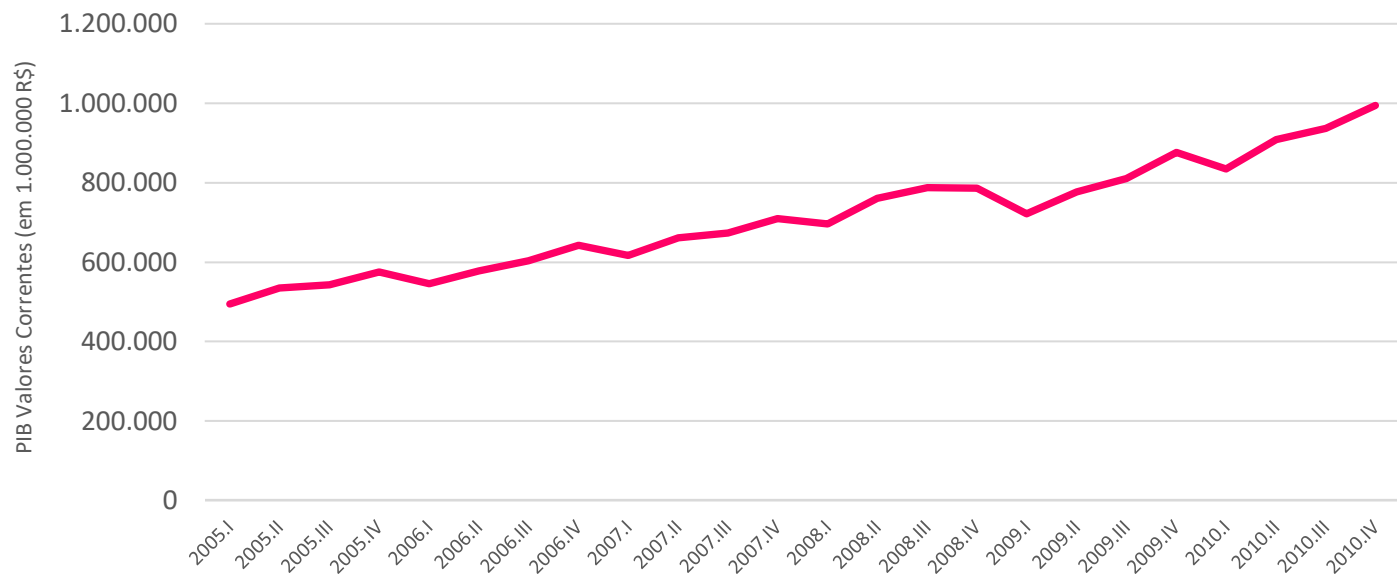


Utilização: As técnicas quantitativas são aplicadas nas condições:

- Informações históricas de pelo menos dois anos disponíveis;
- Informações quantificáveis em forma numérica;
- Assumir a hipótese de que algo dos padrões do passado irá se repetir no futuro (hipótese de continuidade).

SÉRIES TEMPORAIS

PIB, Valores Correntes Trimestrais



SÉRIES TEMPORAIS

Análise de Séries Temporais visa identificar e explicar:

$$\text{Série} = T + S + C + A$$

T: Tendência

S: Sazonalidade

C: Ciclo

A: Aleatório

Tendência – evolução do fenômeno de interesse.

Sazonalidade – regularidade ou variação sistemática na série de dados.

Padrões Cíclicos – repetição de padrão num prazo superior a 2 anos.

Aleatório – comportamento não explicável pelos três componentes anteriores (Erro Aleatório).

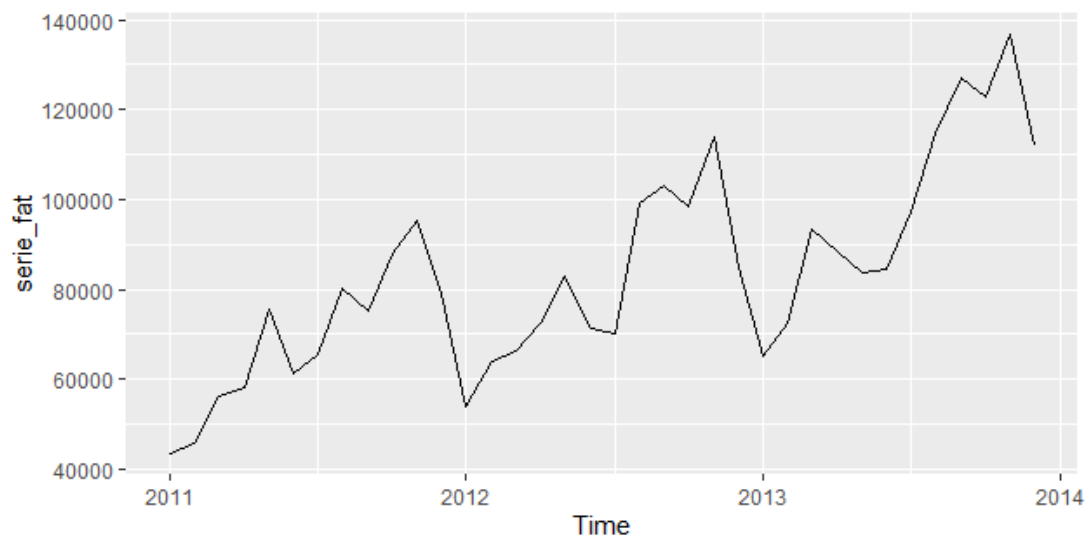
SÉRIES TEMPORAIS

Principais objetivos ao analisar uma série temporal:

- Investigar o mecanismo gerador da série temporal; por exemplo, analisando uma série de altura de ondas, queremos saber como estas ondas foram geradas;
- Fazer previsões de valores futuros (curto ou longo prazo) da série;
- Descrever apenas o comportamento da série;
- Procurar periodicidade relevante nos dados.

SÉRIES TEMPORAIS

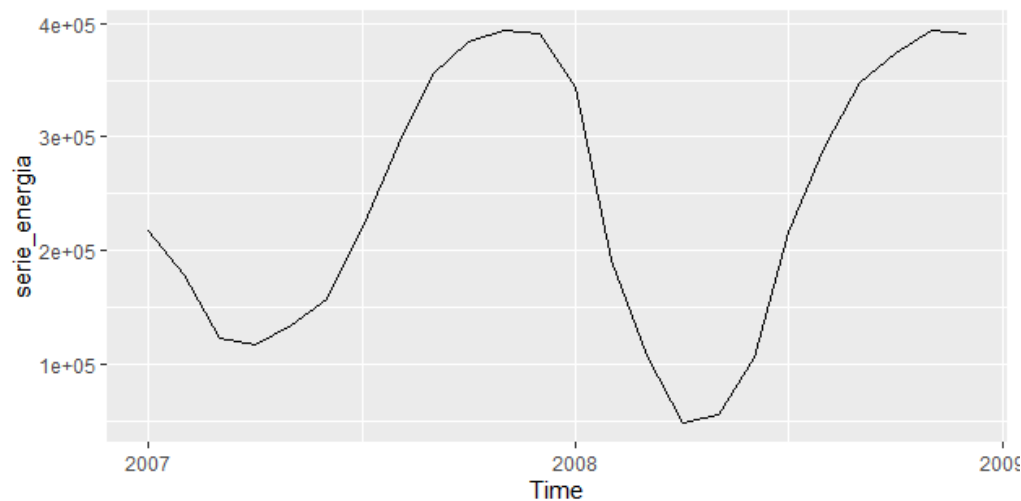
Série temporal do faturamento (R\$)



A série apresenta tendência? Sazonalidade?


SÉRIES TEMPORAIS

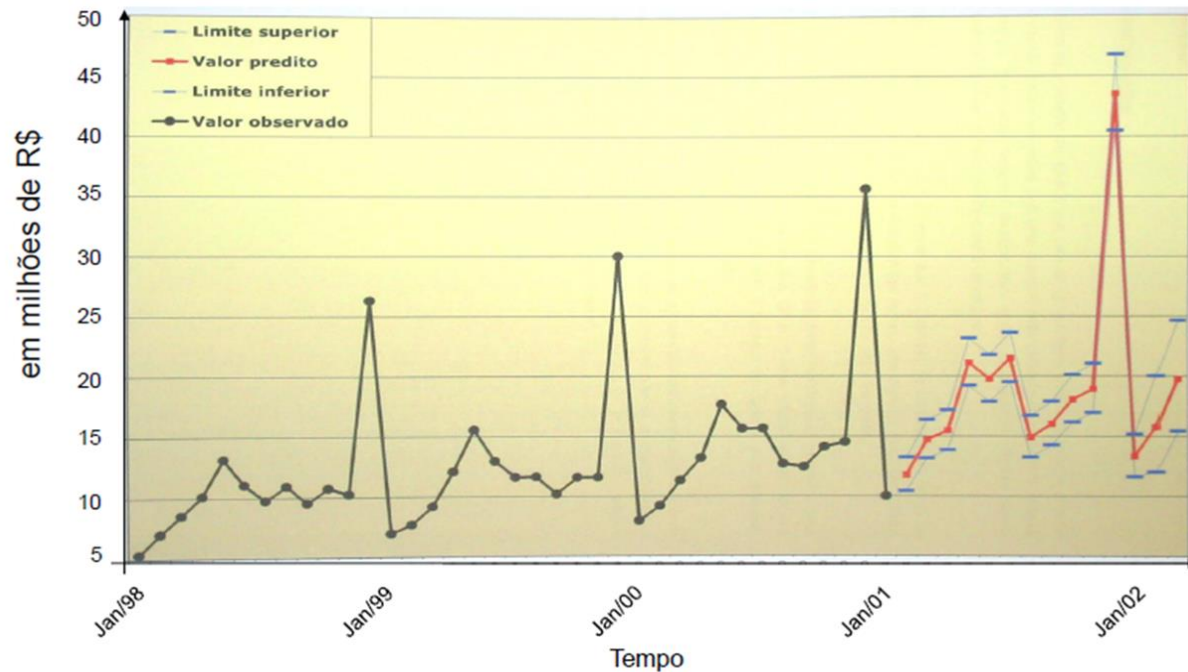
Série temporal do consumo de energia (Kw/h) de empresas do setor Agricultura



A série apresenta tendência? Sazonalidade?

- MODELOS DE SÉRIES TEMPORAIS

-  Previsão de 12 meses para o faturamento mensal - Varejo (Vestuário) -



FREQUÊNCIA DA SÉRIE

SÉRIES TEMPORAIS

FREQUÊNCIA DA SÉRIE

UNIDADE DE ANÁLISE	FREQUÊNCIA
Anual	1
Mensal	12
Diária	365
Trimestral	4
Semanal	52

SÉRIES TEMPORAIS

Exemplo 1:

Ano	Mes	Faturamento
2011	1	43484
2011	2	45859
2011	3	56254
2011	4	58224
2011	5	75403
2011	6	61255
2011	7	65601
2011	8	80099
2011	9	75017
2011	10	87932
2011	11	95266
2011	12	79175
2012	1	54085
2012	2	63808
2012	3	66330
2012	4	72442
2012	5	83072
2012	6	71321
2012	7	70095
2012	8	99071
2012	9	103100
2012	10	98380
2012	11	113751
2012	12	84933

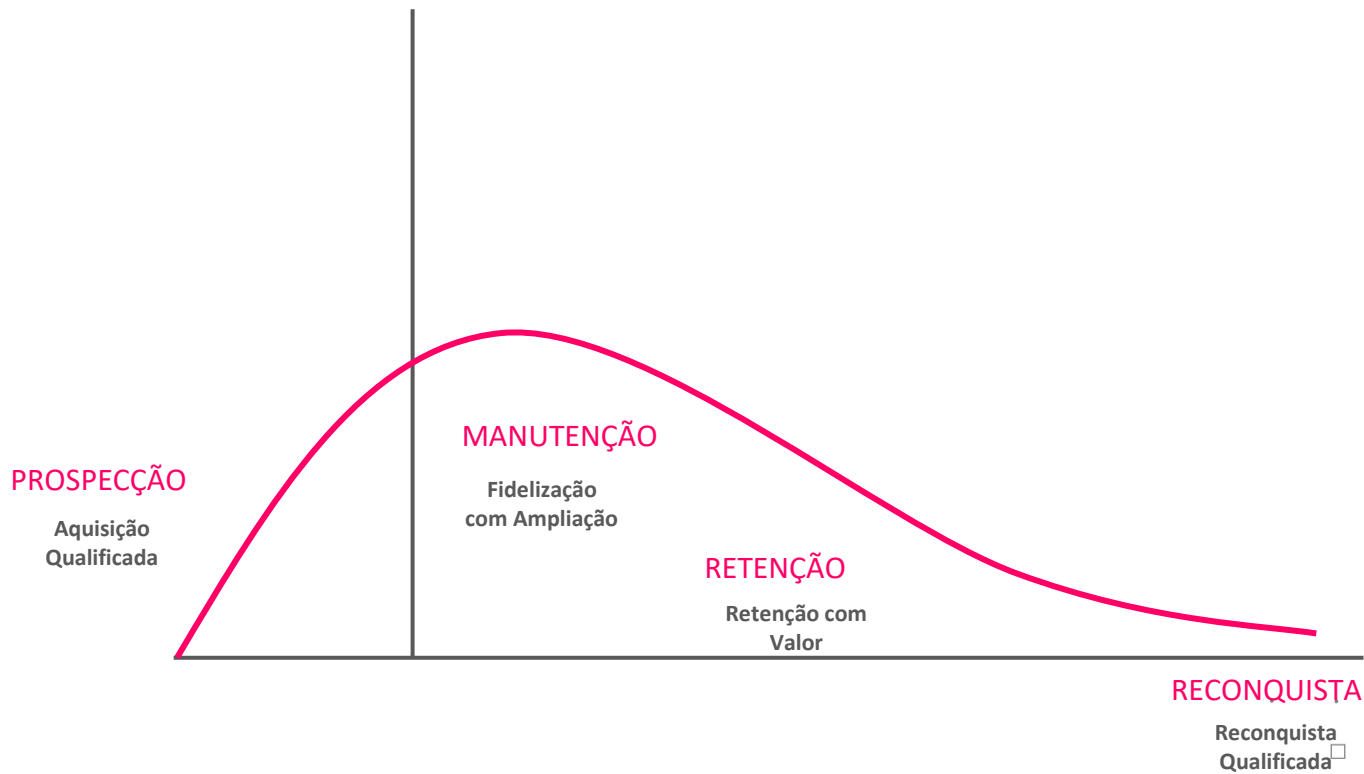
Exemplo 2:

Período	Proporção de vendas
17/01 a 23/01	34.1
24/01 a 30/01	27.9
31/01 a 06/02	26.7
07/02 a 13/02	15.4
14/02 a 20/02	37.0
21/02 a 27/02	25.0
28/02 a 06/03	46.7

Exemplo 3:

instant	dteday	Bikes alugadas
1	01/01/2011	985
2	02/01/2011	801
3	03/01/2011	1349
4	04/01/2011	1562
5	05/01/2011	1600
6	06/01/2011	1606
7	07/01/2011	1510
8	08/01/2011	959
9	09/01/2011	822
10	10/01/2011	1321
11	11/01/2011	1263
12	12/01/2011	1162
13	13/01/2011	1406
14	14/01/2011	1421
15	15/01/2011	1248
16	16/01/2011	1204
17	17/01/2011	1000

USO DOS MODELOS NO CICLO DO CLIENTE



USO DOS MODELOS NO CICLO DO CLIENTE

- Ciclo de Vida
- Segmentação Geográfica
- Propensão à Compra (1a.)
- Segmentação Atitudinal
- Potencialidade de Mercado
- Modelos Geomarketing
- ...

PROSPECÇÃO

Aquisição
Qualificada

- Ciclo de Vida
- Segmentação Comportamental
- Segmentação Atitudinal
- Segmentação Geográfica
- Score de Cross Selling (Ativação)
- Score de Risco (Pagamento)
- Valor do Cliente
- Modelos de Churn
- Detecção de Fraude
- ...

MANUTENÇÃO

Fidelização
com Ampliação

RETENÇÃO

Retenção com
Valor

- Segmentação Comportamental
- Segmentação Geográfica
- Score de Reconquista
- Propensão à Compra
- Valor do Cliente
- Collection Score
- ...

RECONQUISTA

Reconquista
Qualificada



MODELOS PREDITIVOS

Estatística Tradicional

Machine Learning

Inferência estatística



Conjunto de técnicas estatísticas baseadas em I.C. e erro padrão.

- Baseado em algoritmos.
- O objetivo é identificar o que funciona.
- Interessa em prever os resultados de amostras futuras.
- Foco na praticidade: Desenvolve em uma amostra e aplica em outra.
- Algoritmos para tomada de decisão.
- Limitações/grandes desafios:
 - Tendência ao sobre ajuste.
 - Dados influenciados por erros de medição e fatores aleatórios.
 - Ajuste perfeito para um grupo de dados e pode não funcionar bem para outro.
 - Algoritmo preconceituoso.

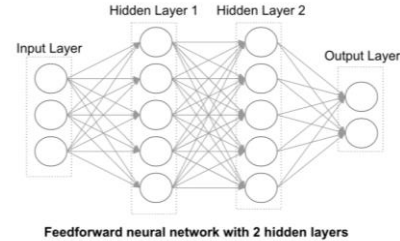
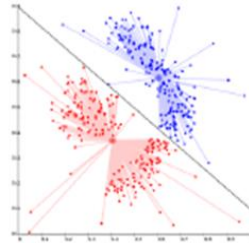
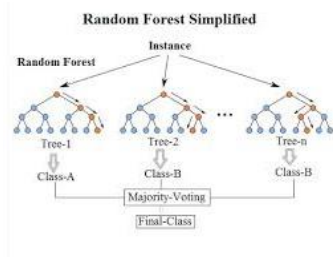
MODELOS PREDITIVOS

Machine Learning

Problemas práticos de predição (para tomada de decisão).

Pouco interesse em interpretar os modelos.

Liberdade para modelar a complexidade do mundo real.



Se machine learning não se importa muito com interpretação, então se importa de fato com o quê?



Performance preditiva (ou seja, acurácia das decisões).

ANÁLISE MULTIVARIADA

Análise Exploratória dos Dados

Análise de Discriminação de Estrutura

- Técnicas de dependência.
- Técnicas Multivariadas aplicáveis **quando uma das variáveis pode ser identificada como dependente** (variável *target*), e as restantes como variáveis independentes (ou preditoras).

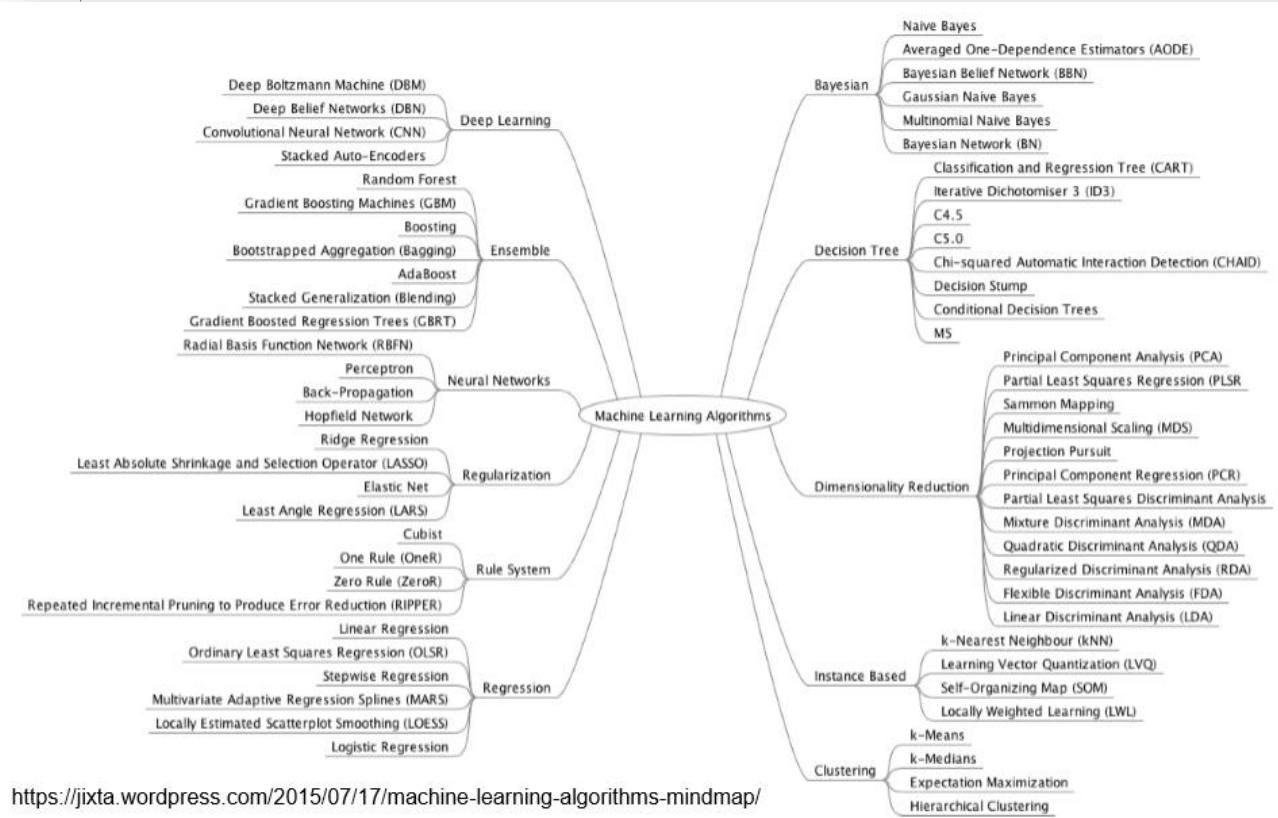
Análise Estrutural

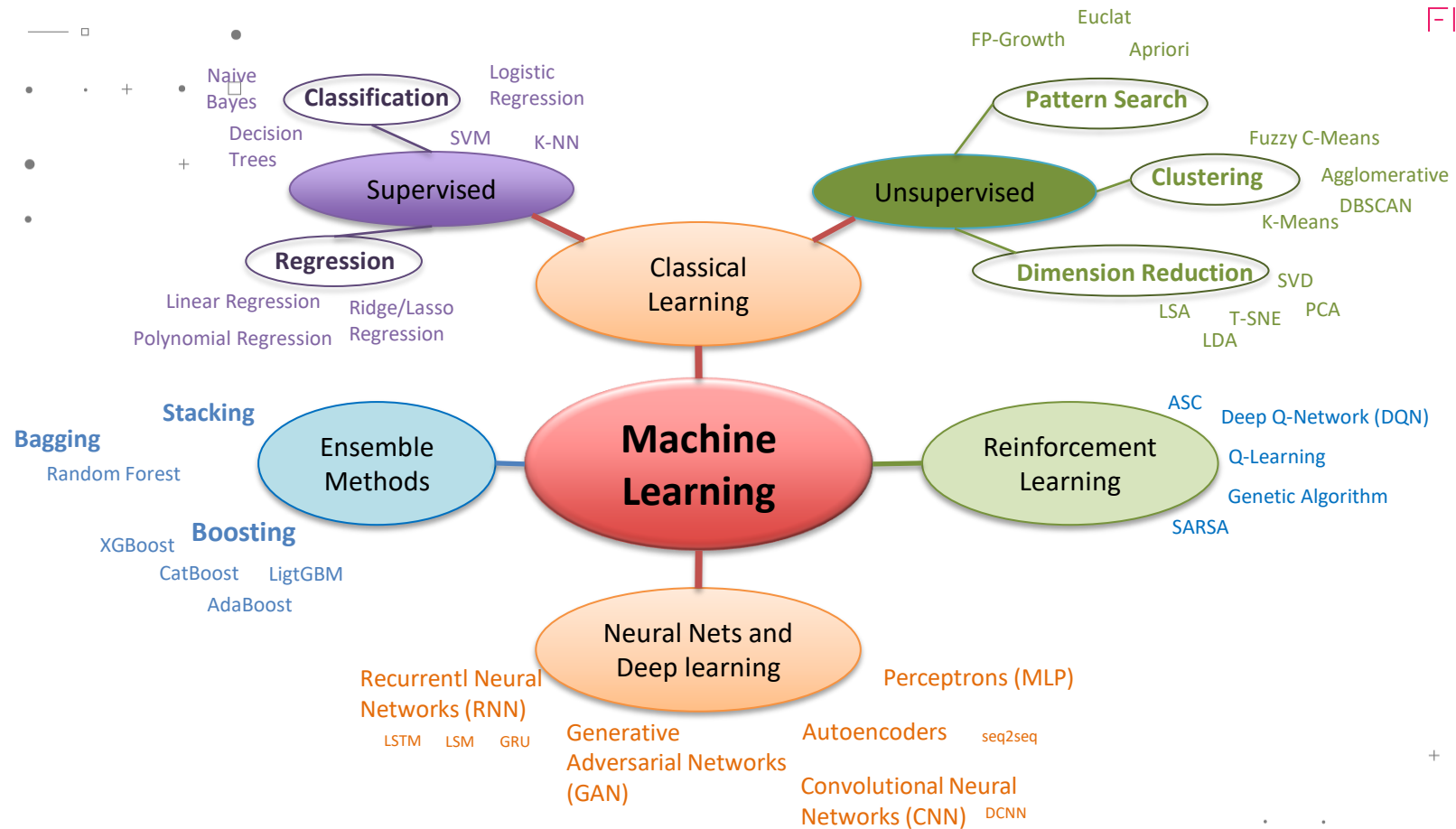
- Técnicas de Interdependência.
- Técnicas Multivariadas que procuram agrupar dados com base em semelhança, permitindo assim a interpretação das estruturas dos dados. **Não há distinção entre variáveis dependentes e independentes.**

Aprendizado Supervisionado

Aprendizado Não Supervisionado

ALGORITMOS de MACHINE LEARNING





TÉCNICAS DE DISCRIMINAÇÃO

CLASSIFICAÇÃO

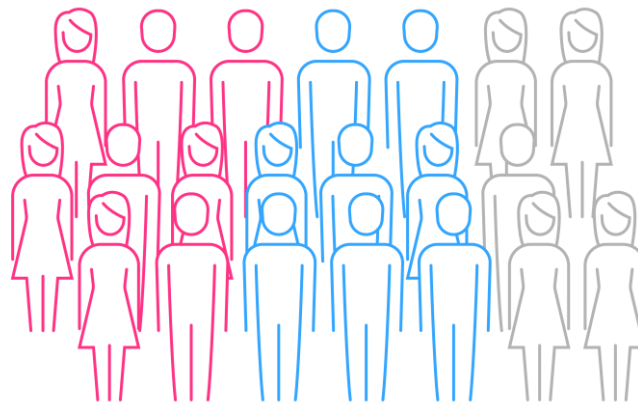
Descobertas Supervisionadas de Relações

Quando a variável target assume “classes/categorias”.

TÉCNICAS DE DISCRIMINAÇÃO

MÉTODO DE CLASSIFICAÇÃO

- Como os heavy users **se diferem** em seu perfil demográfico dos light users ?
- Quais são os clientes ativos que **se assemelham** aos clientes cancelados?
- Que **fatores ou atitudes** fazem com que os meus clientes **prefiram** o meu produto?
- Quais são as **características** que apresentam os clientes que compraram o produto de maior rentabilidade?



GRUPO A

GRUPO B

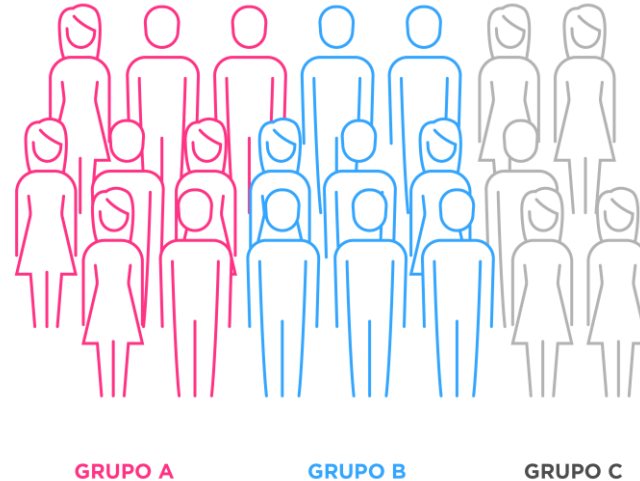
GRUPO C

Como separar grupos **previamente definidos**? Como definir critérios, funções das variáveis que discriminem os grupos?

TÉCNICAS DE DISCRIMINAÇÃO

MÉTODO DE CLASSIFICAÇÃO

- Dado um conjunto de treinamento onde cada registro contém um conjunto de atributos, e um dos atributos é a nossa variável de interesse é tipo categórica/classes.
- Encontrar um modelo para determinar o valor do atributo/classe em função dos valores de outros atributos.
- Objetivo: definir a classe de novos registros, a classe deve ser atribuída o mais corretamente possível.



MODELOS PREDITIVOS - AVALIAÇÃO

Existem diversas métricas para determinar a qualidade de um modelo.
Dois exemplos muito utilizados:

→ problema de estimação ou previsão:
(variável target quantitativa):

- erro quadrático médio (MSE). Calculado por:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 = \frac{(\hat{y}_1 - y_1)^2 + \dots + (\hat{y}_n - y_n)^2}{n},$$

n é o número de observações,

y_i é o valor real e

\hat{y}_i é a predição do modelo.

Utilizamos a Raiz quadrada desse valor RMSE

Nesse caso, um modelo bom é aquele que possui o
menor erro quadrático médio.

→ problema de classificação:
(variável target categórica)

- Percentagem de acertos do modelo

Acurácia: É a proporção de predições corretas.

É dada por:

$$\text{Acurácia} = \frac{\text{Quantidade de Acertos}}{\text{Total}}$$

Nesse caso, um modelo bom é aquele que possui **a maior acurácia.**

MODELOS PREDITIVOS - AVALIAÇÃO

Medidas de desempenho dos modelos

→ problema de classificação:
(variável target categórica/classes)

Id	Observado	Estimado
1	NÃO	SIM
2	SIM	SIM
3	NÃO	NÃO
4	SIM	SIM
5	SIM	NÃO
6	NÃO	NÃO
7	SIM	SIM
8	SIM	SIM
9	NÃO	NÃO
10	NÃO	SIM
11	SIM	SIM
12	NÃO	NÃO

		Estimado		
		NÃO	SIM	Total
Observado	NÃO	4	2	6
	SIM	1	5	6
	Total	5	7	12

$$\text{Acurácia} = (4+5) / 12 = 75,0\%$$

MODELOS DE AQUISIÇÃO

- Adquirir **prospects** com os mesmos perfis dos bons clientes da empresa;
- Campanhas sobre os clientes da concorrência;
- Estimular os clientes à aquisição de novos produtos/serviços (*cross selling*).

TIPOS DE MODELOS

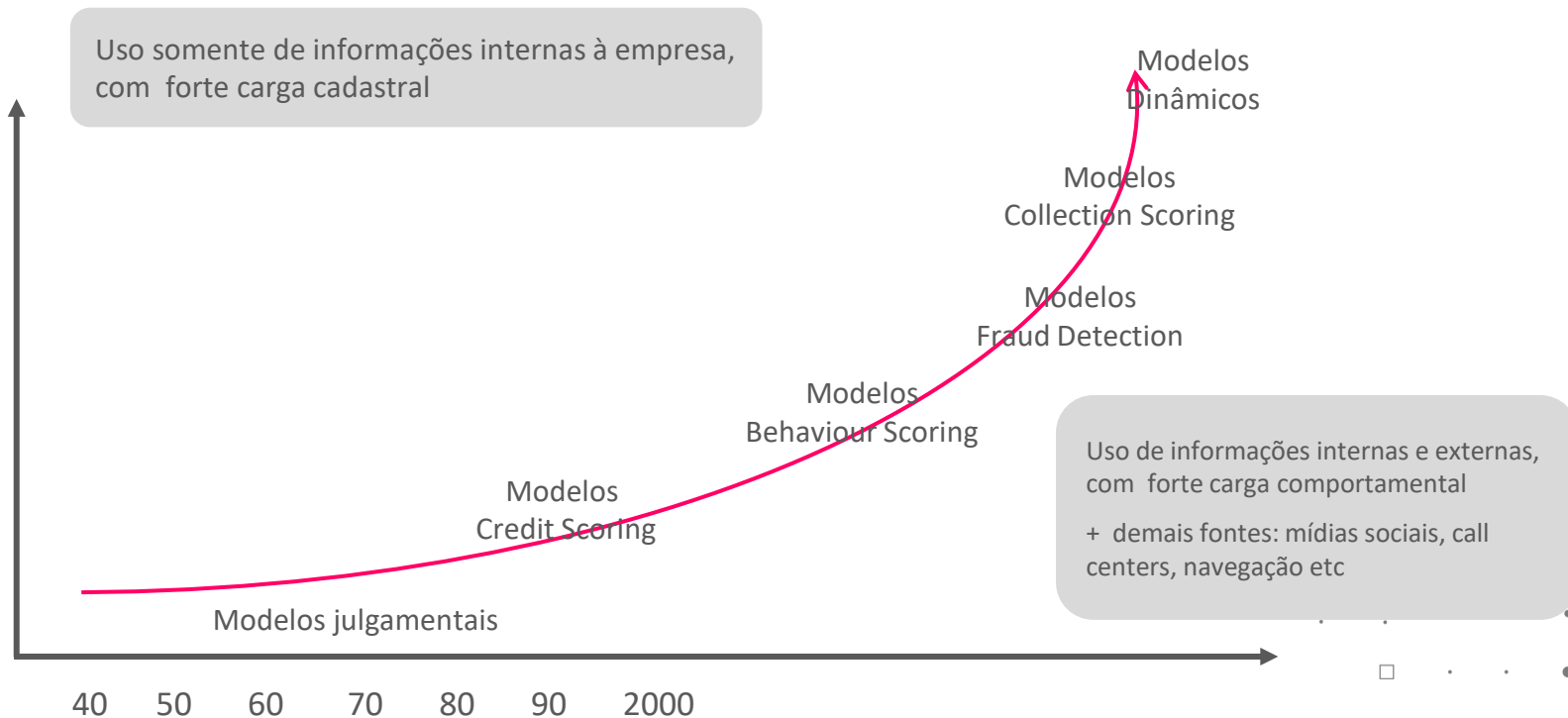
- **MODELOS DE RETENÇÃO ou MODELOS DE CHURN**

Objetivo:

- Identificar na base de dados de clientes prováveis a cancelar o relacionamento com a empresa.
- Oferecer suporte a área de relacionamento e permitir que campanhas de fidelização sejam direcionadas a clientes com risco real de interromper o relacionamento com a instituição.

- Melhores resultados nas campanhas realizadas;
- Redução de custos de abordagens indesejadas;
- Satisfação dos clientes;
- Maior credibilidade.

EVOLUÇÃO DAS FERRAMENTAS DE GESTÃO DE RISCO



TIPOS DE MODELOS

- **Modelo de *Credit Scoring***
 - Considera informações/dados do contrato (tempo de relacionamento recente);
 - Probabilidade de o novo cliente vir a ser inadimplente.
- **Modelo de Inadimplência (*Behaviour Scoring*)**
 - Considera dados de utilização/comportamento dos clientes;
 - Probabilidade de o cliente vir a ser tornar um inadimplente.
- **Modelo de Cobrança (*Collection Scoring*)**
 - Considera dados de utilização dos clientes e do mercado;
 - Probabilidade de um cliente pagar.
- **Modelo de *Churn* e fraude/anomalias/abusos**
 - Considera dados de utilização dos clientes e do mercado;
 - Probabilidade de o cliente cancelar a “conta/serviço/produto”.

MODELOS PREDITIVOS

Mercado Financeiro - Exemplos

Objetivo:

- Identificar na base de dados correntistas prováveis a cancelar/inativar o relacionamento (conta corrente) com o banco;

Dimensões:

- Utilização: diretamente relacionadas à geração de receita de cada correntista (dados transacionais).

Exemplos: produto adquirido, quantidade de cheques emitidos, saldo médio, tempo de relacionamento, conta conjunta, etc.

- Demográficas: informações descritivas do cliente.

Exemplos: sexo, idade, endereço, profissão, estado civil, renda, etc

- Definição da janela de tempo de análise

- Planejamento amostral (técnicas estatísticas aliadas às restrições do Banco)

Benefícios:

- Realizar ações fidelizadoras sobre os correntistas propensos a cancelar/inativar sua conta corrente

TÉCNICAS DE DISCRIMINAÇÃO

MÉTODO DE CLASSIFICAÇÃO

Classificadores *eager* (espertos)

A partir da amostragem inicial (conjunto de treinamento), constroem um modelo de classificação capaz de classificar novos registros.

Uma vez pronto o modelo, o conjunto de treinamento não é mais utilizado na classificação de novos objetos (registros)

- Árvores e Regras de Decisão
- Redes Neurais
- Redes Bayesianas e Naïve Bayes
- SVM-Máquinas de Vetores de Suporte

Classificadores *lazy* (preguiçosos)

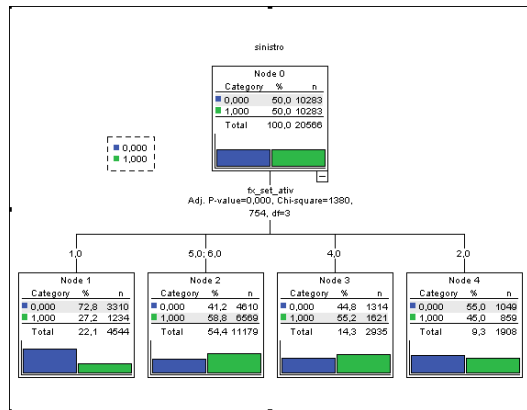
Cada novo registro é comparado com todo o conjunto de treinamento e é classificado segundo a classe do registro que é mais similar. Também conhecido como: Aprendizado baseado em exemplo (*Instance-based Learning*):

- Método kNN (k-nearest-neighbor)

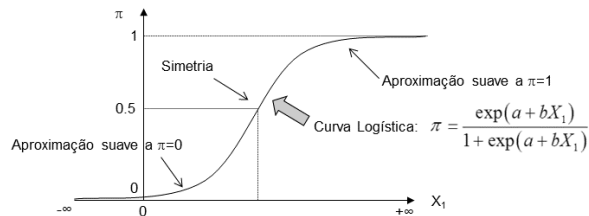
Outros Métodos

- Algoritmos Genéticos
- Conjuntos Fuzzy

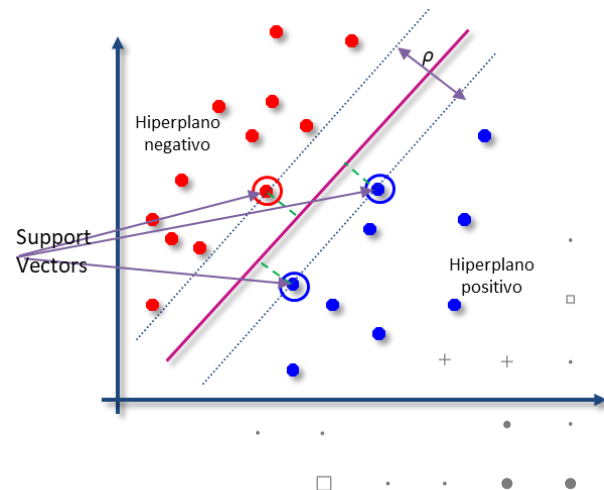
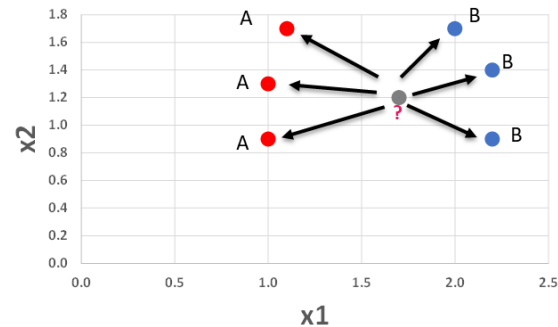
Técnicas de Classificação:



variável	categoria	Coefficientes
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



Qual a distância euclidiana entre os pontos?



TÉCNICAS DE CLASSIFICAÇÃO

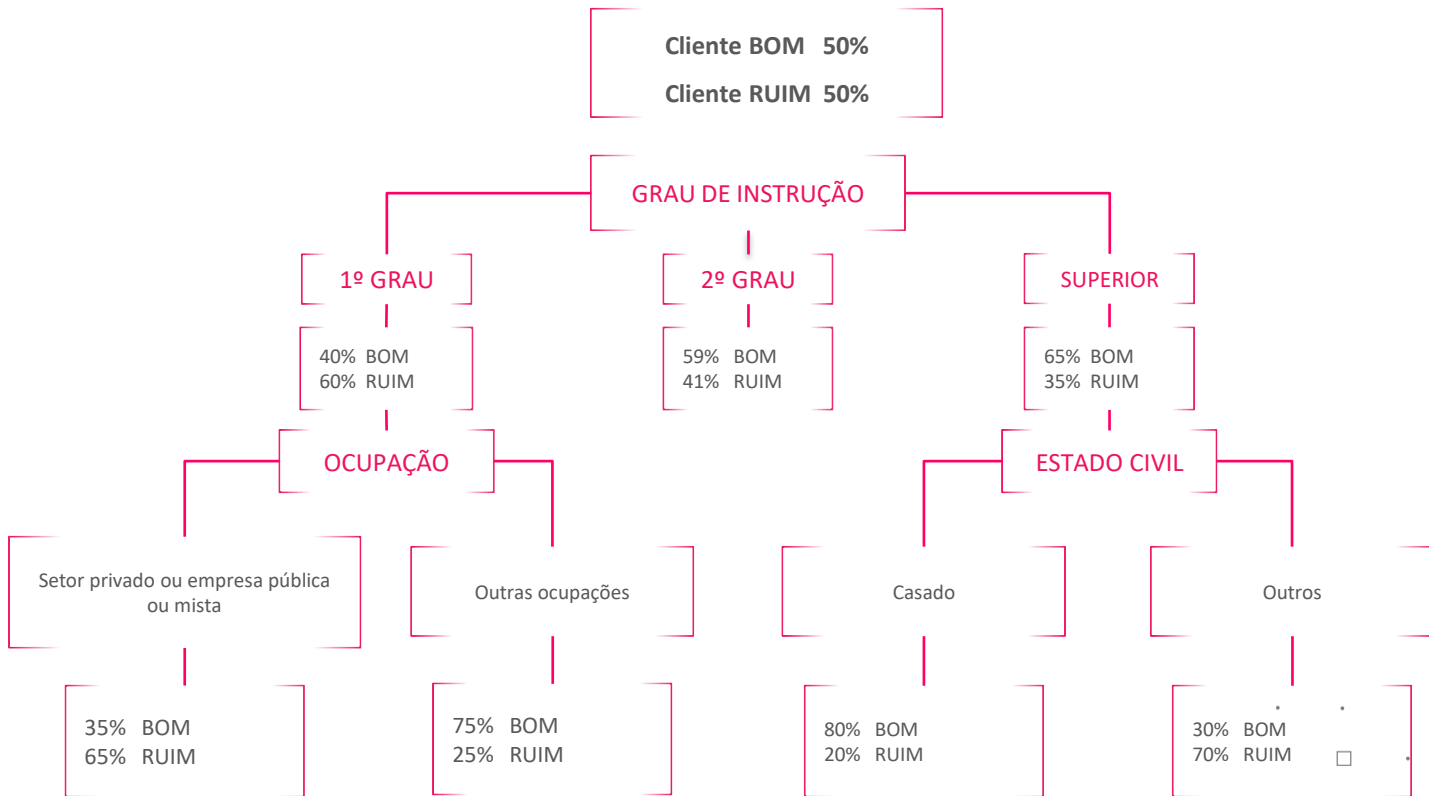
ÁRVORES DE DECISÃO

TÉCNICAS DE DISCRIMINAÇÃO

ÁRVORES DE DECISÃO

- Metodologia estatística de fácil interpretação e utilização.
 - São estruturas de dados compostas de um nó raiz e vários nós filhos, que por sua vez têm seus filhos também e se interligam por ramos, cada um representando uma regra. Os nós que não possuem filhos são chamados de nós folhas e os que têm são chamados de nós pais, ou de decisão.
- Têm como objetivo encontrar regras que discriminem dois grupos previamente conhecidos.
- Exemplo: Encontrar uma regra que trace perfil de pessoas mais propensas a aderir a um certo produto.

ÁRVORES DE DECISÃO – EXEMPLO



ÁRVORES DE DECISÃO – EXEMPLO

Decisões da árvore:

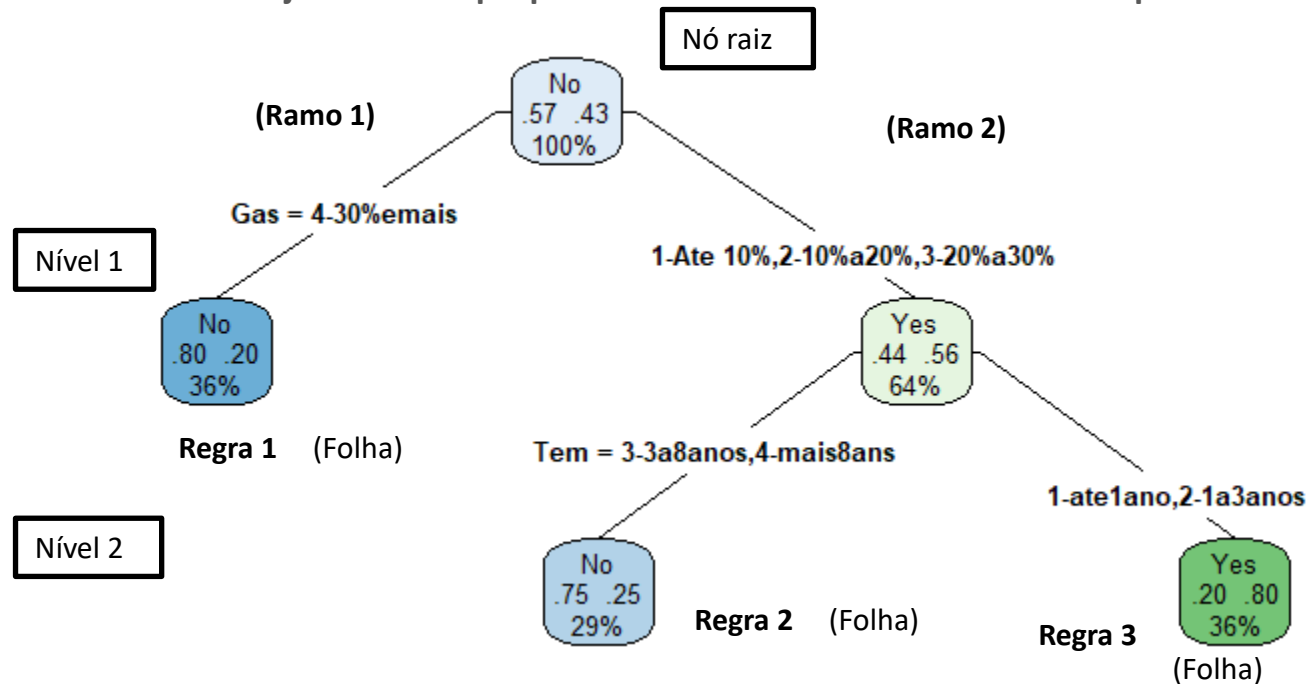
- Qual preditor e qual valor dividir os dados
- Profundidade e complexidade da árvore
- Resultado de cada folha



ÁRVORES DE DECISÃO – EXEMPLO

Segmento: Área Financeira

A área de crédito deseja avaliar a propensão de um cliente tornar-se inadimplente.



ÁRVORES DE DECISÃO

Critérios de parada (hiperparâmetros):

- ☐ Número de observações na divisão dos nós filhos
- ☐ 100% de classificação de uma categoria de resposta
- ☐ Número de níveis

EXEMPLO – MODELO DE PROPENSÃO

Segmento: Seguro Residencial

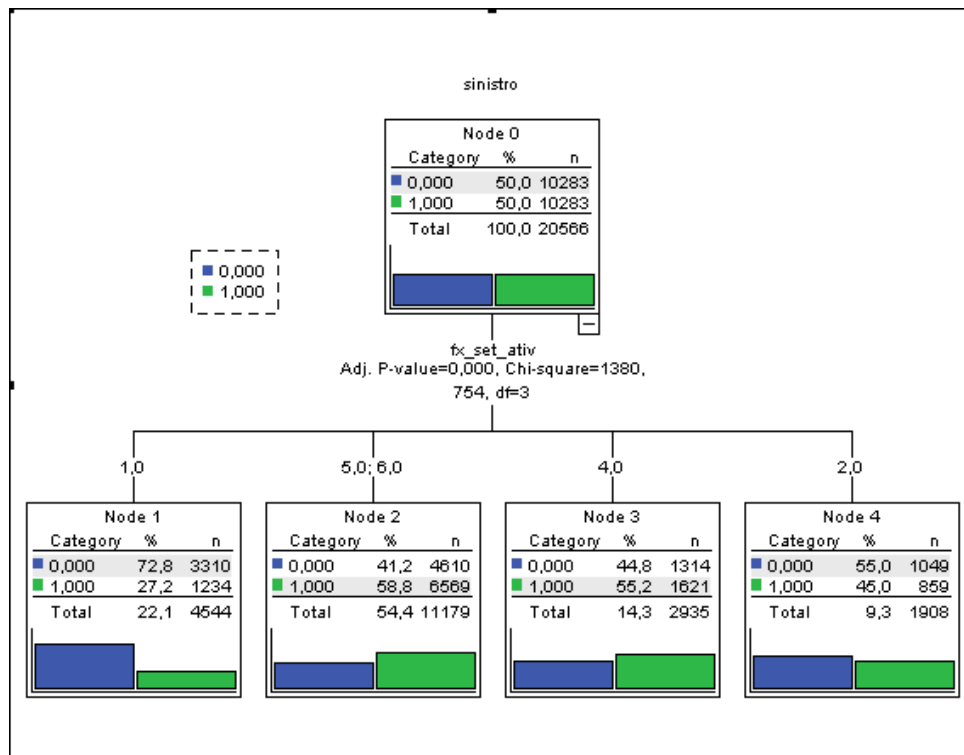
A área de Seguros deseja avaliar a propensão de um novo cliente sinistrar na apresentação de uma proposta.

EXEMPLO – MODELO DE PROPENSÃO

<u>apolice</u>	<u>parcelas</u>	<u>qtde_cob</u>	<u>tpconstr</u>	<u>tipmora</u>	<u>clasmora</u>	<u>corretor</u>	<u>corrent</u>	<u>uf</u>	<u>set_ativ</u>	<u>Impseg(R\$)</u>	<u>sinistro</u>
925578	6	9	6	casa	moradia	2	N	MS	90	100000	1
395699	1	9	6	apto	moradia	1	S	ES	26	30000	0
863771	11	9	6	casa	moradia	1	S	SP	24	200000	0
892165	11	9	6	casa	moradia	1	S	MG	27	30000	0
923092	1	9	6	casa	veraneio	2	N	SP	90	70000	0
1003098	4	9	6	casa	veraneio	1	S	SP	7	150000	1
955644	11	9	6	casa	moradia	1	S	MG	11	30000	1
987421	1	9	6	casa	moradia	2	N	SP	90	65000	1
744959	4	9	6	casa	veraneio	1	S	RS	18	70000	1
920814	11	9	6	casa	moradia	2	S	SP	90	100000	0
395550	2	9	6	casa	moradia	1	S	ES	26	20000	0
972615	6	9	6	casa	veraneio	2	N	SP	90	87500	1
958900	11	9	6	casa	moradia	1	S	MG	23	85000	1
911272	4	9	6	casa	veraneio	2	N	SP	90	150000	0
895508	11	9	6	casa	moradia	1	S	MG	33	50000	0
374234	1	9	6	apto	moradia	1	N	DF	6	30000	0
883254	11	9	6	casa	moradia	1	S	SP	24	100000	0
727885	3	9	6	casa	moradia	2	S	RS	90	180000	1
327315	11	9	6	casa	moradia	1	S	BA	21	20000	0
910241	11	9	6	apto	moradia	1	S	SP	49	50000	0
956554	10	9	6	casa	moradia	1	S	MG	27	70000	1
1000162	3	9	6	casa	moradia	2	S	MS	90	80000	1
920421	1	9	6	casa	veraneio	1	S	SP	1	40000	1

EXEMPLO – MODELO DE PROPENSÃO

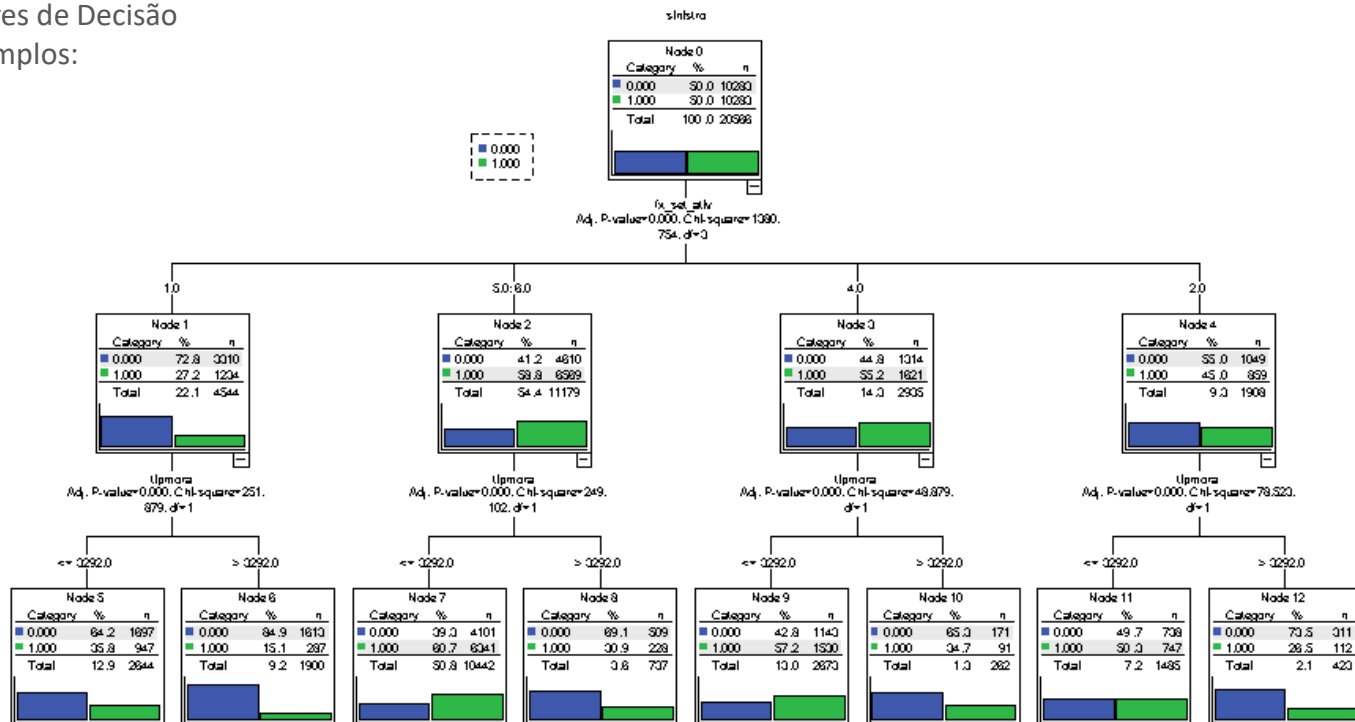
Árvores de Decisão
- exemplos:



EXEMPLO – MODELO DE PROPENSÃO

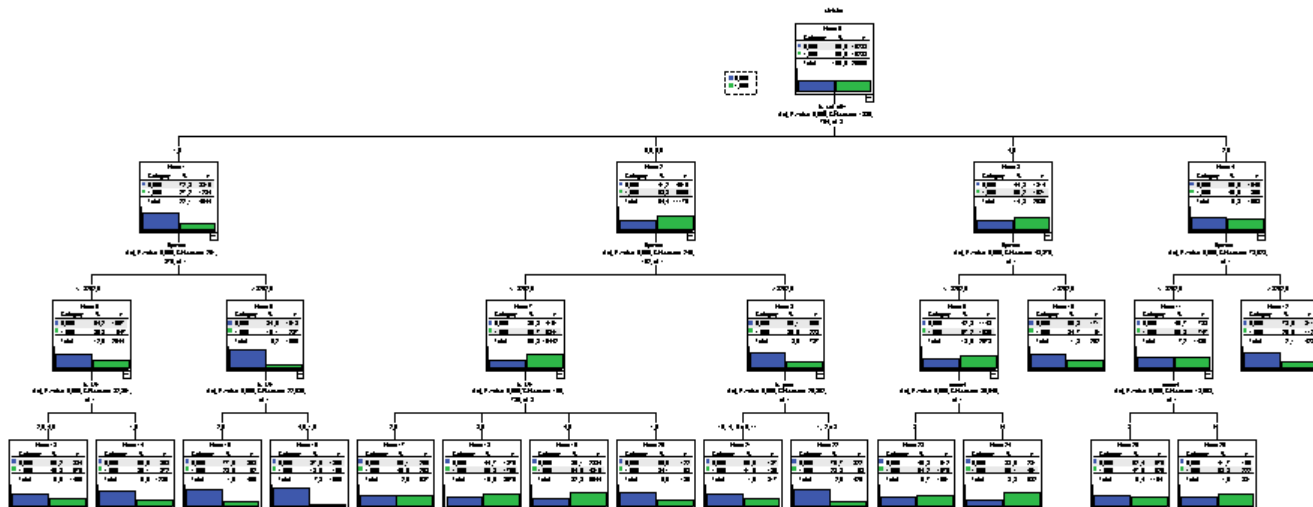
Árvores de Decisão

- exemplos:



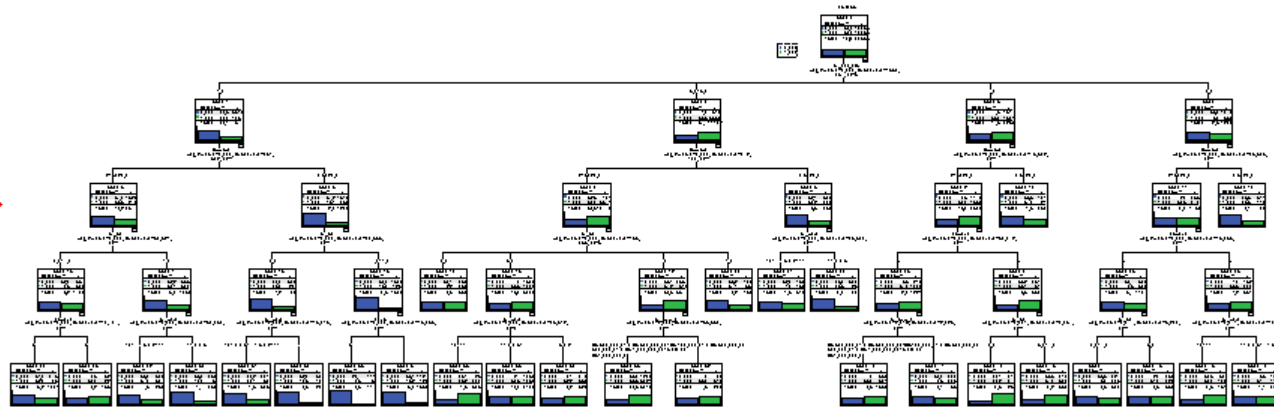
EXEMPLO – MODELO DE PROPENSÃO

Árvores de Decisão
- exemplos:



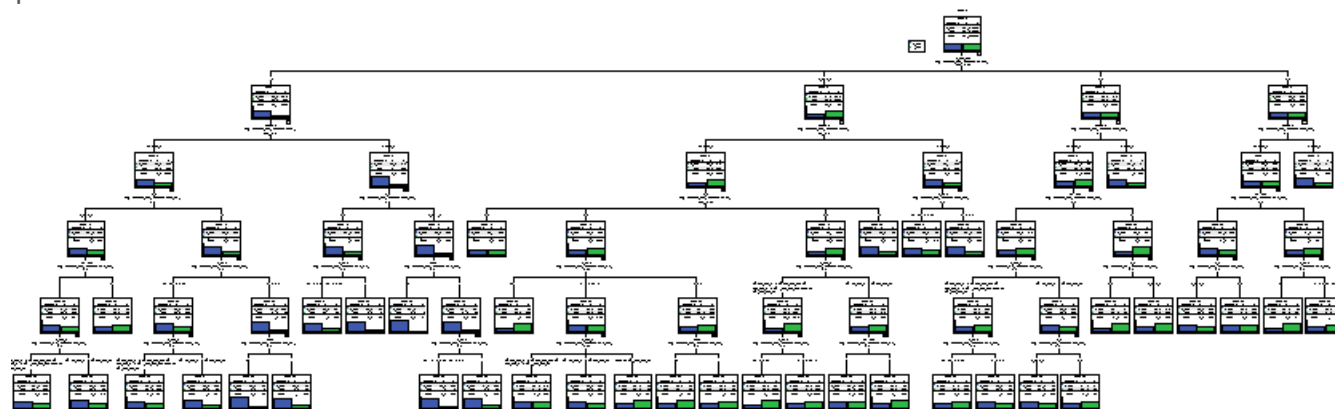
EXEMPLO – MODELO DE PROPENSÃO

Árvores de Decisão
- exemplos:



EXEMPLO – MODELO DE PROPENSÃO

Árvores de Decisão
- exemplos:



AVALIAÇÃO DO MODELO

Exemplo

Classification			
Observed	Predicted		
	0	1	Percent Correct
0	5.137	999	83,7%
1	1.208	4.412	78,5%
Overall Percentage	81,0%	81,5%	81,2%
Growing Method: EXHAUSTIVE CHAID Dependent Variable: Resposta			

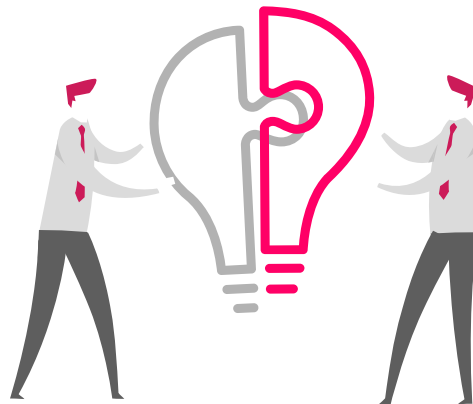
EXEMPLO – MODELO DE PROPENSÃO

Árvores de decisão

- Algoritmos utilizados:
 - CHAID: CHi-square Automatic Interaction Detector
 - CART: Classification And Regression Trees
- Tipos de Variáveis
 - Variáveis Categóricas (nominais ou ordinais)
 - Variável Frequência e Ponderada (Weight)

EXERCITANDO

ÁRVORE DE DECISÃO



Inadimplência

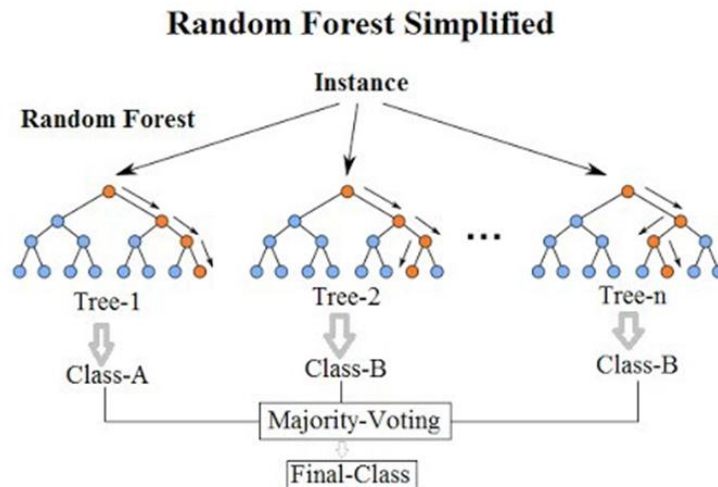
MÉTODOS DE ENSEMBLE LEARNING

BAGGING – RANDOM FOREST

Random Forest é uma técnica de bagging.

Usa **diversas árvores de decisão como modelos individuais**, além de fazer uma seleção aleatória de casos e de variáveis. As árvores são extremamente interpretáveis, entretanto costumam ter um poder preditivo muito baixo quando comparados aos demais estimadores. Uma forma de contornar isso é através **da combinação da predição fornecida por diversas árvores para se fazer predições**.

Cada árvore tenta estimar uma classificação e isso é chamado como “voto”. Idealmente, consideramos cada voto de cada árvore e escolhemos a classificação mais votada (estatística: Moda). No caso de problemas de regressão funciona similarmente, cada árvore tenta estimar a variável target e depois é considerada a média dos valores estimado em cada árvore.

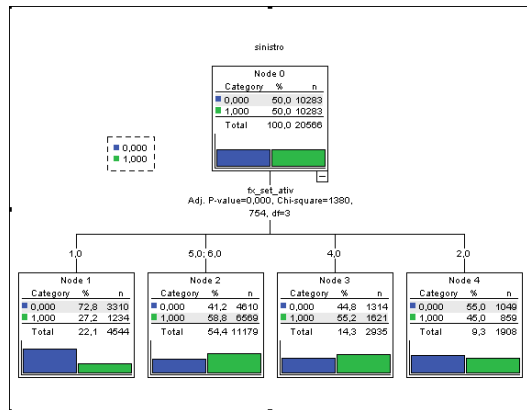


MÉTODOS DE ENSEMBLE LEARNING

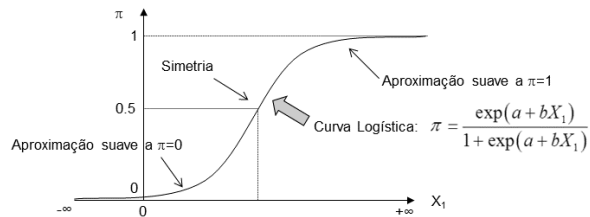
BAGGING

- Como funciona:
 - Treina **modelos individuais** usando uma amostra aleatória para cada;
 - **Agrega os modelos individuais depois de treinados** com suas respectivas amostras;
 - No caso de problemas de regressão usa a média e no caso de classificação a moda;
- Vantagem:
 - ajuda reduzir a variância (amostragem aleatória);
 - pode reduzir o viés (pois estamos usando média e moda para combinar os modelos);
 - fornece estabilidade e robustez (alto número de estimadores usados).
- Desvantagem:
 - tem um custo computacional alto (usa muito espaço e tempo - cada nova iteração é criada uma amostra diferente)
 - A técnica só funciona se o modelo base já tem uma boa performance. Usar o bagging em um modelo base ruim pode fazer com que o modelo final fique ainda pior. Como os modelos individuais usam o mesmo algoritmo, o bagging pode não reconhecer alguns padrões.

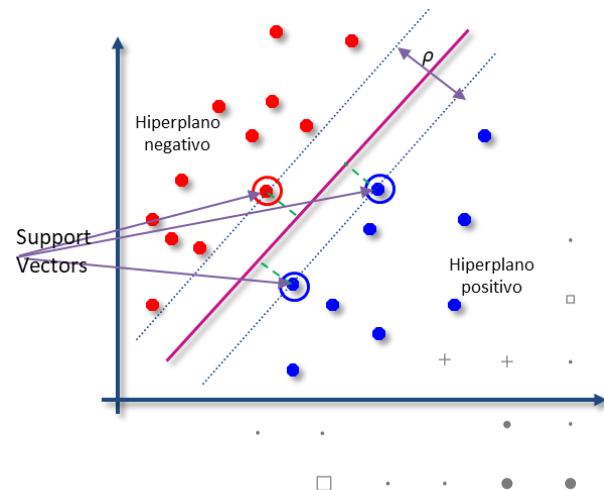
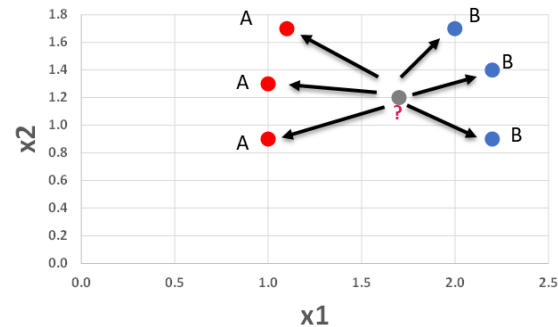
Técnicas de Classificação:



variável	categoria	Coefficientes
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



Qual a distância euclidiana entre os pontos?



TÉCNICAS DE CLASSIFICAÇÃO

REGRESSÃO
LOGÍSTICA

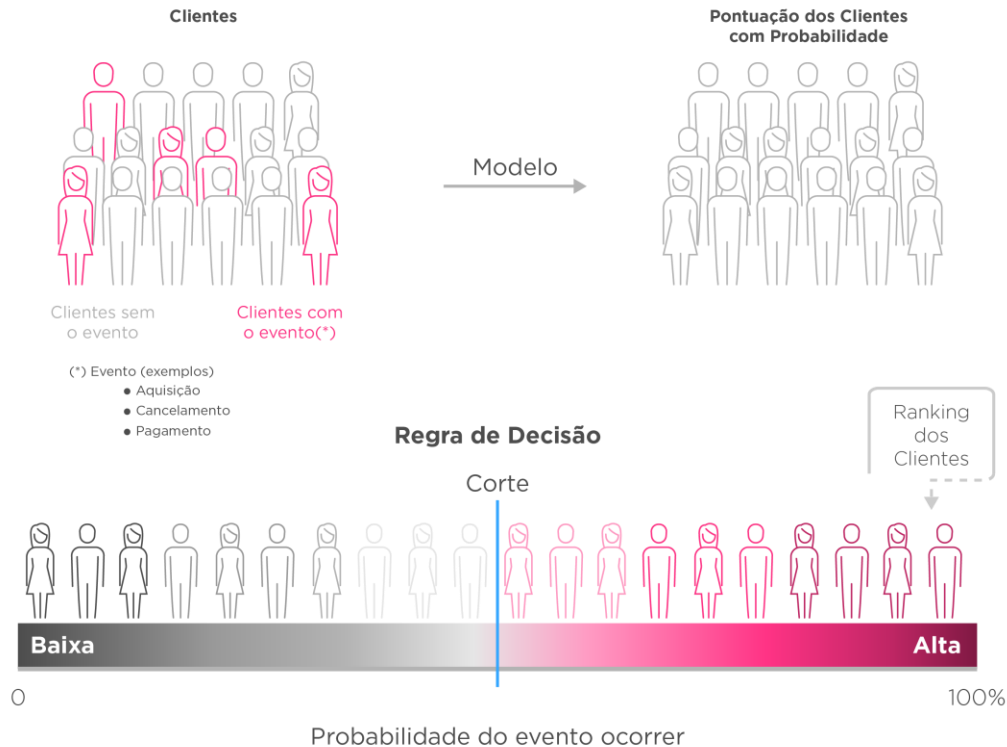
ANÁLISE DE DISCRIMINAÇÃO DE ESTRUTURA

REGRESSÃO LOGÍSTICA

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

ANÁLISE DE DISCRIMINAÇÃO DE ESTRUTURA

REGRESSÃO LOGÍSTICA



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
 - (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**:
 - $p1 / (1-p1) = (146/266) / (120/266) = 0,55 / 0,45 = 1,22$
- **Chance** de preferir o canal de futebol entre **mulheres**:
 - $p2 / (1-p2) = (110/234) / (124/234) = 0,47 / 0,53 = 0,89$
- **Razão de chances** de preferir canal de futebol **entre homens, em relação às mulheres**:
 - $[p1/(1-p1)] / [p2/(1-p2)] = 1,22 / 0,89 = 1,37$

ANÁLISE DE REGRESSÃO LOGÍSTICA

Modelo de Regressão Logística

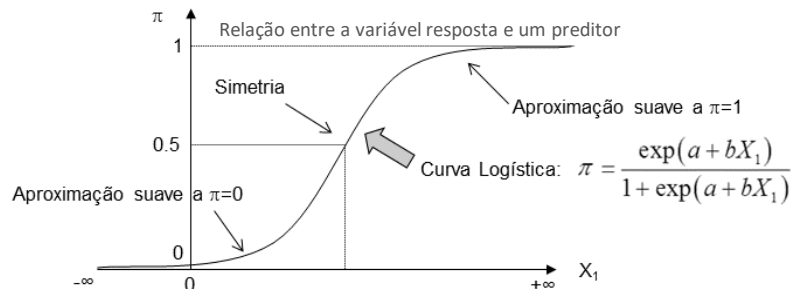
$$G = a + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$$

G: logit da resposta de preferência (sim) a :

Intersecção B_1, B_2, \dots, B_n : coeficientes logísticos

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

$$p = \frac{\exp(x)}{1 + \exp(x)}$$



ANÁLISE DE REGRESSÃO LOGÍSTICA

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 $[L(y, \vartheta)]$, que é a função de probabilidade de ocorrência de um específico conjunto de dados e estimar os parâmetros que a maximizam.

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 sequência 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)

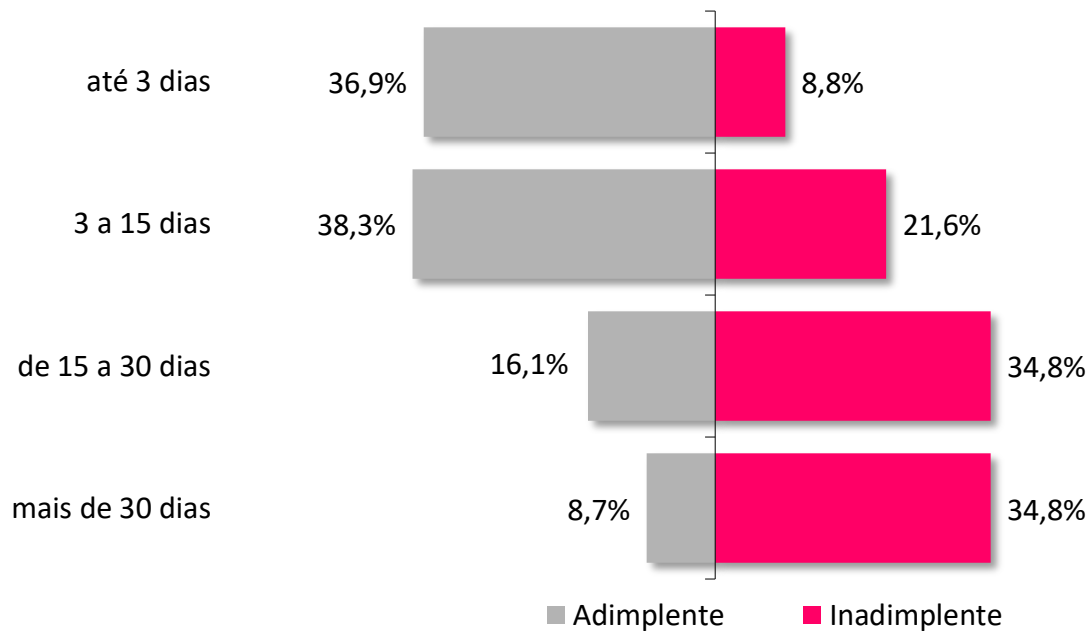
EXEMPLO – MODELO DE INADIMPLÊNCIA

Segmento: Cartões de Crédito

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

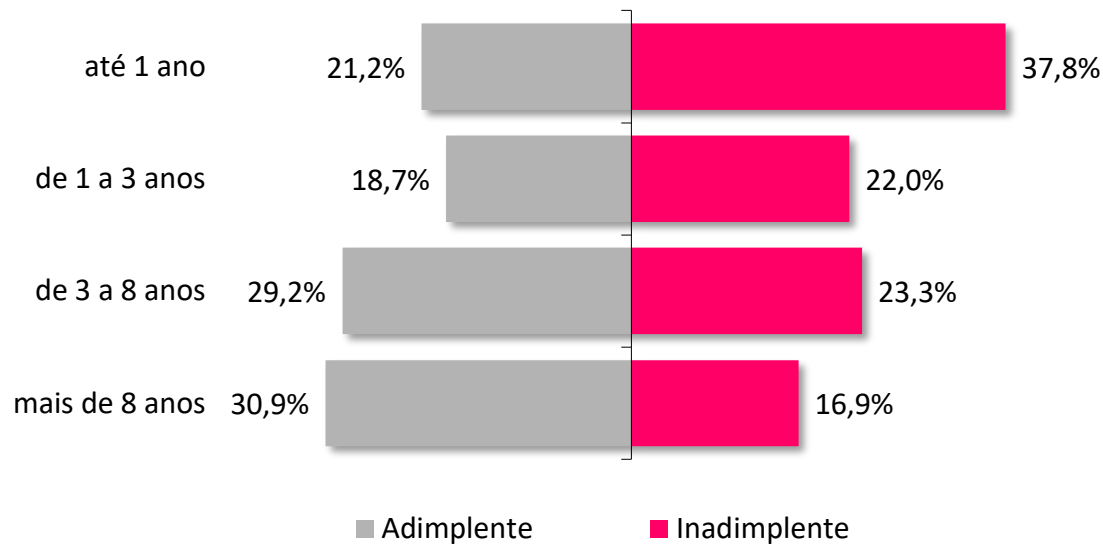
EXEMPLO – MODELO DE INADIMPLÊNCIA

Média de dias com pagamentos em atraso nos últimos 6 meses



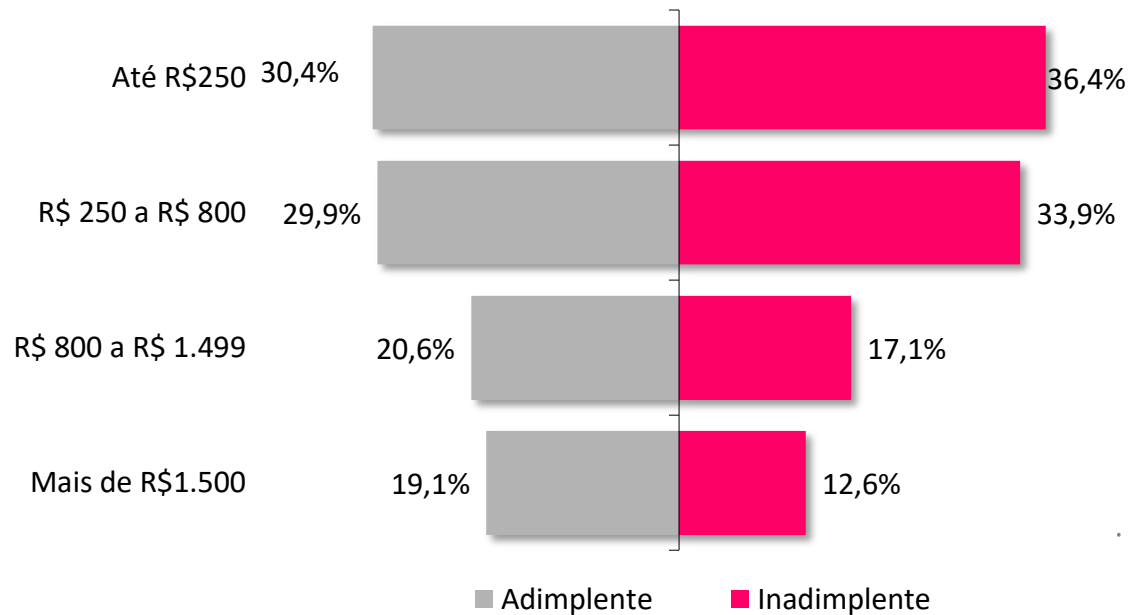
EXEMPLO – MODELO DE INADIMPLÊNCIA

Tempo de relacionamento em anos



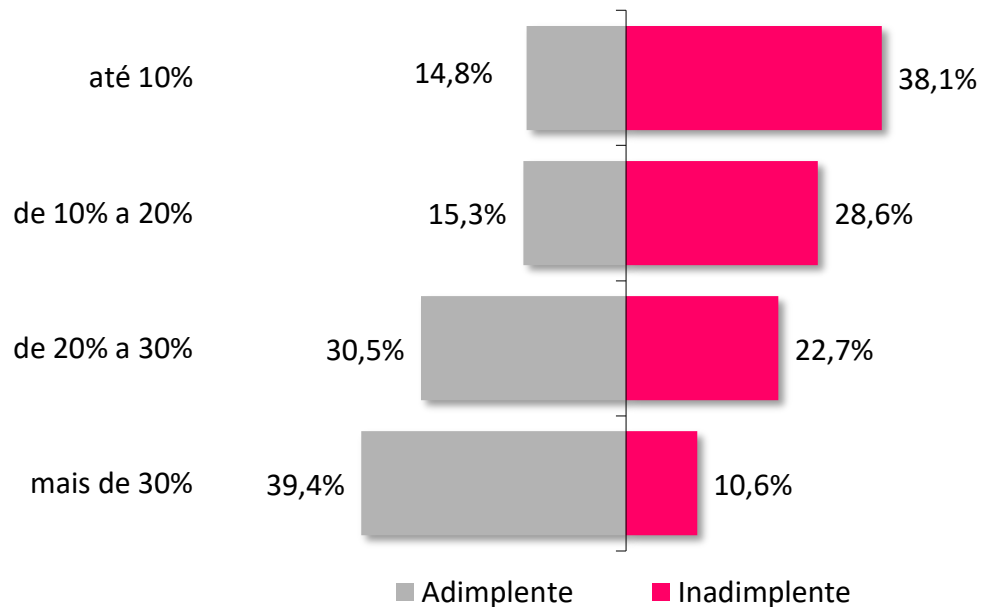
EXEMPLO – MODELO DE INADIMPLÊNCIA

Valor Médio da Fatura Mensal



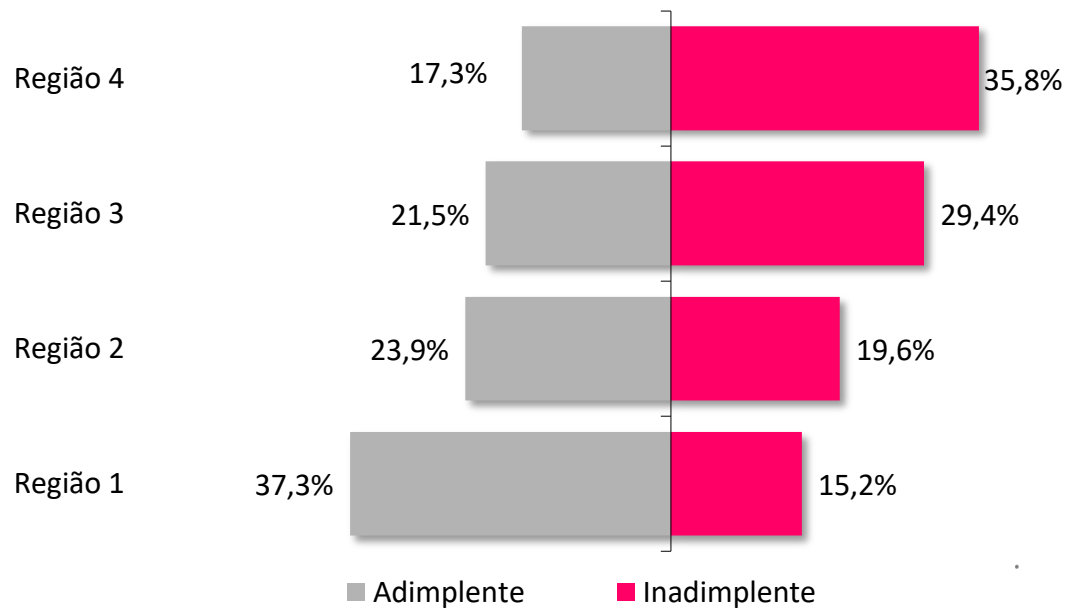
EXEMPLO – MODELO DE INADIMPLÊNCIA

Percentual dos gastos em alimentação



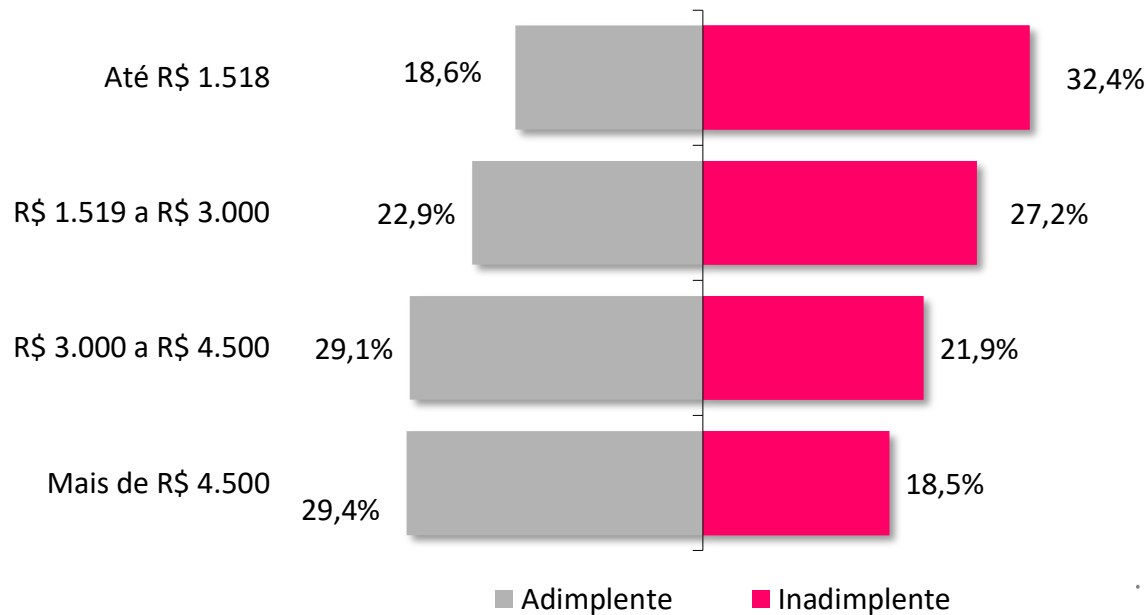
EXEMPLO – MODELO DE INADIMPLÊNCIA

Regiões de Risco



EXEMPLO – MODELO DE INADIMPLÊNCIA

Renda média mensal



EXEMPLO – MODELO DE INADIMPLÊNCIA

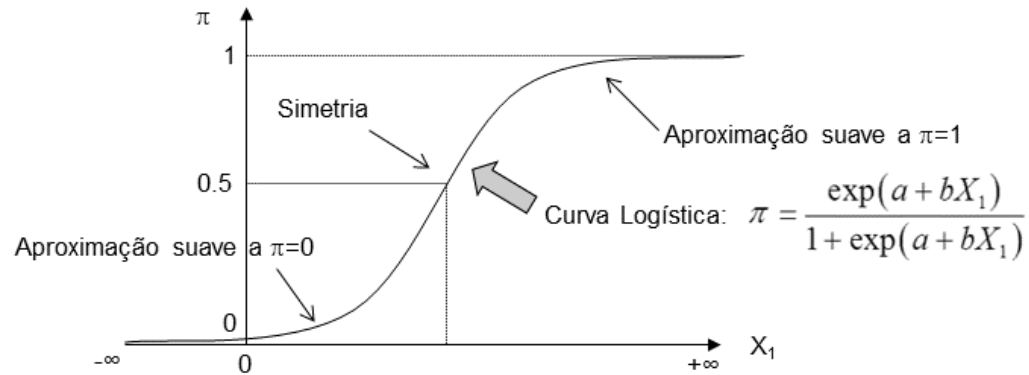
Tabela de Coeficientes do Modelo

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

Tabela de Coeficientes do Modelo

$$p = \frac{\exp(x)}{1 + \exp(x)}$$



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%

AVALIAÇÃO DO MODELO

- Exemplo

Classification			
Observed	Predicted		
	0	1	Percent Correct
0	5.137	999	83,7%
1	1.208	4.412	78,5%
Overall Percentage	81,0%	81,5%	81,2%
Growing Method: EXHAUSTIVE CHAID Dependent Variable: Resposta			

TÉCNICAS DE CLASSIFICAÇÃO

Qualificação do Ajuste do Modelo

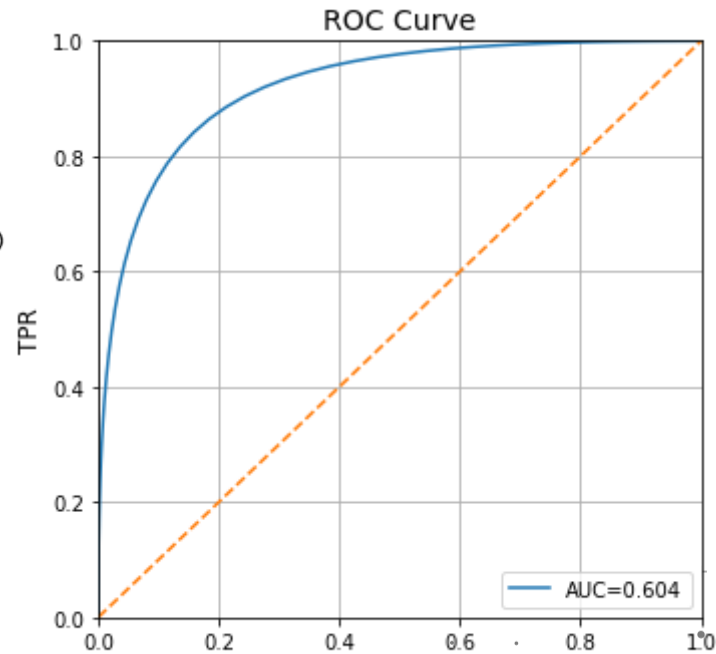
Qualificação do Ajuste do Modelo

		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 = $n4 / (n3+n4)$

- Acurácia: É a proporção de predições corretas: $(n1+n4) / (n1+n2+n3+n4)$
- A curva ROC plota (chamado de sensibilidade) versus (chamado de 1-especificidade) para todos os possíveis pontos de corte entre 0 e 1.
- Uma forma bastante utilizada para determinar o ponto de corte .



$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

TÉCNICAS DE CLASSIFICAÇÃO

Qualificação do Ajuste do Modelo

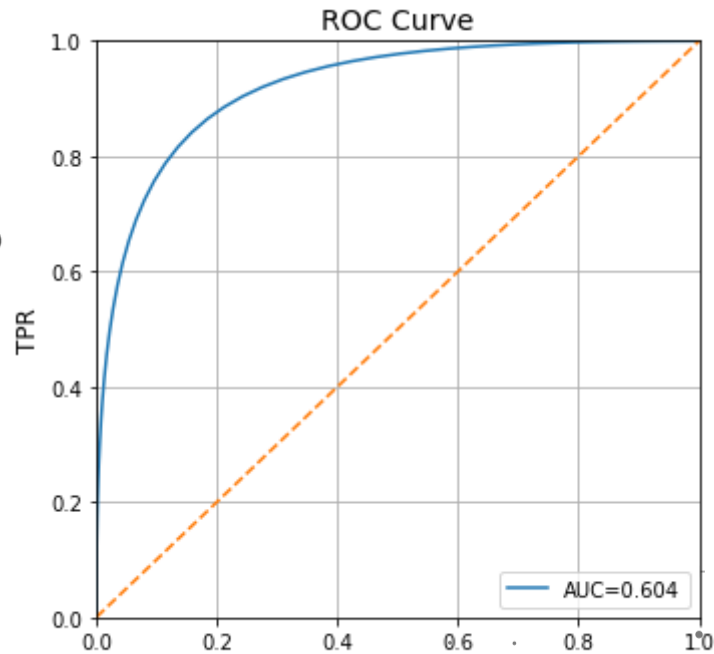
Qualificação do Ajuste do Modelo

		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 = $n4 / (n3+n4)$

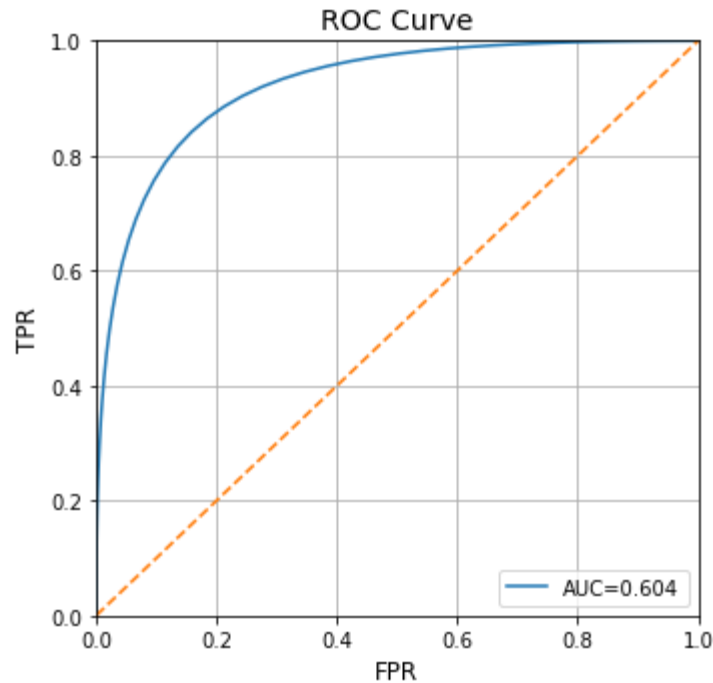
- Acurácia: É a proporção de predições corretas: $(n1+n4) / (n1+n2+n3+n4)$
- A curva ROC plota (chamado de sensibilidade) versus (chamado de 1-especificidade) para todos os possíveis pontos de corte entre 0 e 1.
- Uma forma bastante utilizada para determinar o ponto de corte.



$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

TÉCNICAS DE CLASSIFICAÇÃO

Qualificação do Ajuste do Modelo



- A curva ROC plota (chamado de sensibilidade) versus (chamado de 1-especificidade) para todos os possíveis pontos de corte entre 0 e 1.
- Uma forma bastante utilizada para determinar o ponto de corte .

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}.$$

TÉCNICAS DE CLASSIFICAÇÃO

Qualificação do Ajuste do Modelo

Matriz de
Confusão

		Classe Predita	
		positivo	negativo
Classe Esperada	positivo	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
	negativo	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

Medidas de Avaliação

- Sensibilidade ou taxa de verdadeiros positivos: $(VP / (VP + FN))$
- Especificidade ou taxa de verdadeiros negativos: $(VN / (FP + VN))$
- Taxa de falsos positivos: % de falsos positivos dentre todos que a classe esperada é a classe negativa: $(FP / (VN + FP))$
- Taxa de falsas descobertas: % de falsos positivos dentre a classe esperada é a classe positiva: $(FP / (VP + FP))$
- Preditividade positiva ou precisão: % de acertos ou verdadeiros positivos: $(VP / (VP + FP))$
- Preditividade negativa: % de verdadeiros negativos dentre todos classificados como negativos: $(VN / (VN + FN))$
- Acurácia: É a proporção de predições corretas, sem considerar o que é positivo e o que negativo e sim o acerto total. É dada por: $(VP+VN)/(VP+FN+FP+VN)$

ANÁLISE DE REGRESSÃO LOGÍSTICA

Exemplo: Modelo Cross-Selling

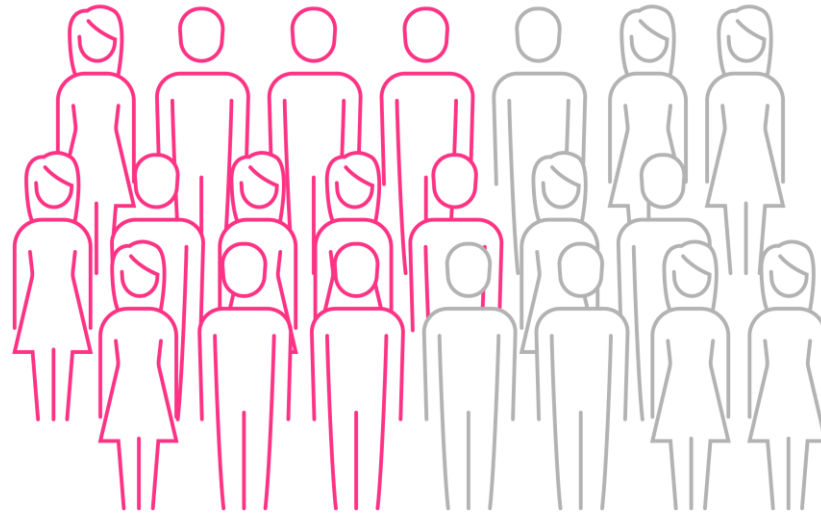
Propensão à Compra de um Produto

Objetivo:

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

MODELOS CROSS SELLING

- Propensão de compra do Produto X

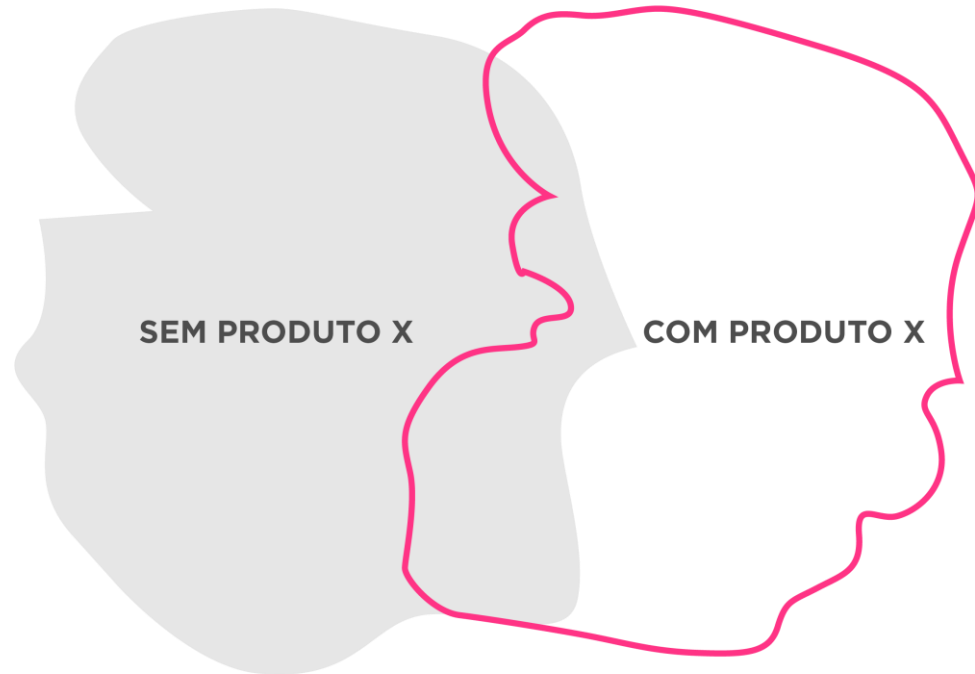


SEM PRODUTO X

COM PRODUTO X

MODELOS CROSS SELLING

- Propensão de compra do Produto X



MODELOS CROSS SELLING

Implementação

- Propensão de compra do Produto X

Algoritmo Matemático

Para associar uma probabilidade de compra de um produto X a 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.

$$\text{Probabilidade} = 100 \times e^{\frac{Y}{(1 + e)^Y}}$$

MODELOS CROSS SELLING

Implementação

- Propensão de compra do Produto X

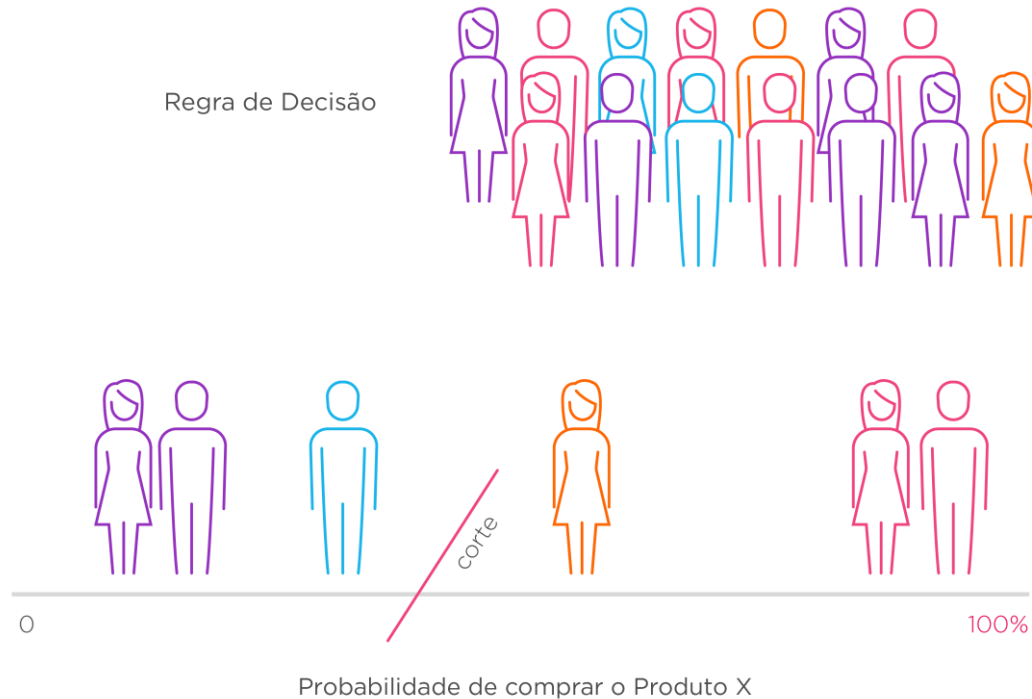
Regra de Decisão Estatística

Após associar a 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 essa 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.

MODELOS CROSS SELLING

- Propensão de compra do Produto X



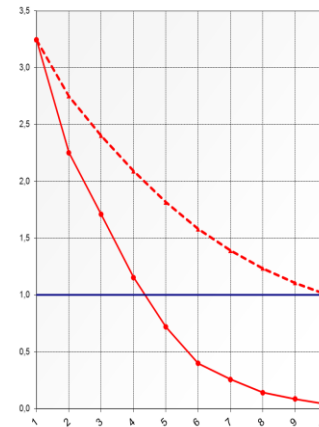
AVALIAÇÃO DO MODELO PREDITIVO

Decil	Clientes	Base sem utilização de modelos						
		Penetração			Lift		Capture	
	Qtde	Qtde	%	% Ac.	Lift	Lift Ac.	%	% Ac.
1	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	10,0%
2	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	20,0%
3	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	30,0%
4	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	40,0%
5	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	50,0%
6	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	60,0%
7	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	70,0%
8	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	80,0%
9	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	90,0%
10	5.300	612	11,5%	11,5%	1,00	1,00	10,0%	100,0%
Total	53.000	6.120	11,5%					

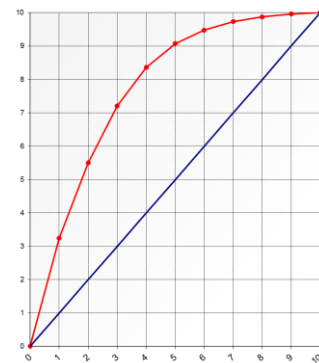
AValiação DO MODELO PREDITIVO

Decil	Clientes	Aplicando Modelo						
		Penetração			Lift		Capture	
	Qtde	Qtde	%	% Ac.	Lift	Lift Ac.	%	% Ac.
1	5.300	1.986	37,5%	37,5%	3,24	3,24	32,4%	32,4%
2	5.300	1.379	26,0%	31,7%	2,25	2,75	22,5%	55,0%
3	5.300	1.046	19,7%	27,7%	1,71	2,40	17,1%	72,1%
4	5.300	706	13,3%	24,1%	1,15	2,09	11,5%	83,6%
5	5.300	440	8,3%	21,0%	0,72	1,82	7,2%	90,8%
6	5.300	244	4,6%	18,2%	0,40	1,58	4,0%	94,8%
7	5.300	157	3,0%	16,1%	0,26	1,39	2,6%	97,3%
8	5.300	87	1,6%	14,3%	0,14	1,23	1,4%	98,8%
9	5.300	52	1,0%	12,8%	0,08	1,11	0,8%	99,6%
10	5.300	24	0,5%	11,5%	0,04	1,00	0,4%	100,0%
Total	53.000	6.120	11,5%					

LIFT CHART

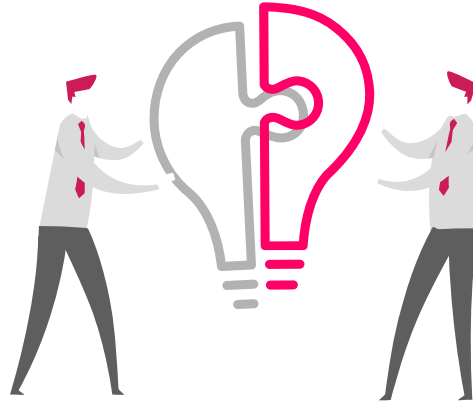


CAPTURE RESPONSE



EXERCITANDO

REGRESSÃO
LOGÍSTICA



Inadimplência

BIBLIOGRAFIA

- KUHN, M. / JOHNSON K. **Applied Predictive Modeling**, 1st ed. 2013, Corr. 2nd printing 2018 Edition
- LESKOVEC, RAJAMARAM, ULLMAN. **Mining of Massive Datasets**, 2014. <http://mmds.org>.
- HAIR, J.F. / ANDERSON, R.E. / TATHAN, R.L. / BLACK, W.C. **Análise multivariada de dados**, 2009
- TORGO, L. **Data Mining with R: Learning with Case Studies**, 2.a ed. Chapman and Hall/CRC , 2007
- MINGOTI, S.A.; **Análise de dados através de métodos de estatística multivariada**, UFMG, 2005
- CARVALHO, L.A.V., **Datamining – A mineração de dados no marketing, medicina, economia, engenharia e administração**. Rio de Janeiro: Editora Ciência Moderna, 2005.
- BERRY, M.J.A., LINOFF, G. **Data Mining Techniques For Marketing, Sales and Customer Support**. 3a. ed. New York: John Wiley & Sons, Inc., 2011.
- DUNHAM, M.H. **Data Mining - Introductory and Advanced Topics**. Prentice Hall, 2002.
- DINIZ, C.A.R. , NETO F.L. **Data Mining: Uma Introdução**. São Paulo: XIV Simpósio Nacional de Probabilidade e Estatística. IME-USP, 2000.

OBRIGADO



/AdelaideAlves



profadelaide.alves@fiap.com.br

FIAP

Copyright © 2022 | Professor (a) Adelaide Alves de Oliveira

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor.