Analytics & Inteligência Artificial

Tema da aula

Regressão Logística

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseada em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.


Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Conteúdo da Aula

1. *Information Value*
2. VIF
3. Ponto de Corte
4. KS e AUC
5. Exercícios

1. *Information Value*

Information Value

O **Valor da Informação,** ou ***Information Value (IV),*** é um indicador que mede a força da relação entre variáveis e é útil para realizar uma análise prévia das variáveis que têm maior potencial para o modelo.

O **IV** é indicador calculado por meio da análise bivariada entre as variáveis explicativas (categorizadas) versus a variável resposta (binária). Quanto maior seu valor, mais ele explica a resposta, variando na prática geralmente entre valores de 0 a 0,5.

Análise Bivariada

O propósito da **análise bivariada** é identificar a relação e o quão forte é a variável explicativa em relação à variável resposta. Esta análise investigativa é muito importante para identificar potenciais variáveis candidatas a compor o modelo de Regressão Logística.

- Por exemplo, um estudo pretende identificar as variáveis mais importantes para explicar se um cliente paga ou não paga um empréstimo.

- A covariável em análise nesse caso é a "quantidade de contatos da assessoria de cobrança com o cliente certo (chamado de CPC – contato com a pessoa certa)".

Quantidade de CPC	Cliente Pagou a Dívida?					
(contatos com a pessoa certa)	Não (0)		Sim (1)		Total	%
	#	%	#	%		
Zero	1.366.039	**66,30%**	23.128	**23,70%**	1.389.167	64,40%
Um	289.587	**14,10%**	20.356	**20,80%**	309.942	14,40%
Dois	139.243	**6,80%**	13.583	**13,90%**	152.826	7,10%
Três	82.422	**4,00%**	8.202	**8,40%**	90.624	4,20%
Quatro	103.220	**5,00%**	14.305	**14,60%**	117.525	5,40%
Cinco	79.775	**3,90%**	18.157	**18,60%**	97.932	4,50%
Total	2.060.286	**100%**	97.730	**100,00%**	2.158.015	100,00%

Analisando as distribuições de clientes que pagaram a dívida em comparação a aqueles não pagaram a dívida, verificamos que as distribuições são diferentes, concentrando um maior número de contatos para os clientes que pagaram a dívida. **Esta análise mostra que esta covariável parece ser relevante para explicar a resposta.**

Information Value e WOE

A análise dos percentuais revela que a incidência de pagamentos é mais baixa quando não se fala com o titular da dívida, enquanto que esse número se eleva consideravelmente à medida que a quantidade de CPC aumenta.

Para facilitar esse processo para um número maior de variáveis (dezenas, centenas de variáveis), a força da relação entre as variáveis pode ser traduzida por um indicador chamado **"valor da informação (VI ou IV)",** que contempla não apenas a diferença entre %bons e %maus para cada categoria, mas também considera, de algum modo, a penetração desse atributo na carteira.

$$VI = \sum_{i=0}^{n} \ln\left(\frac{\%Bons_i}{\%Maus_i}\right) * (\%Bons_i - \%Maus_i)$$

O indicador WOE (*weight of evidence*) ou "**peso da evidência**", é um indicador da força do atributo e é parte do cálculo do IV, como $WOE = \ln\left(\frac{\%Bons_i}{\%Maus_i}\right)$.

Information Value

$$IV = \sum_{i=0}^{n} \ln\left(\frac{\%Bons_i}{\%Maus_i}\right) * (\%Bons_i - \%Maus_i)$$

Quantidade de CPC (contatos com a pessoa certa)	Cliente Pagou a Dívida?								
	Não (0)		Sim (1)		Total	%	SIM/NÃO	WOE	INF. VALUE
	#	%	#	%					
Zero	1.366.039	66,3%	23.128	23,7%	1.389.167	64,4%	0,36	-103,0	0,4393
Um	289.587	14,1%	20.356	20,8%	309.942	14,4%	1,48	39,3	0,0266
Dois	139.243	6,8%	13.583	13,9%	152.826	7,1%	2,06	72,1	0,0515
Três	82.422	4,0%	8.202	8,4%	90.624	4,2%	2,10	74,1	0,0325
Quatro	103.220	5,0%	14.305	14,6%	117.525	5,4%	2,92	107,2	0,1032
Cinco	79.775	3,9%	18.157	18,6%	97.932	4,5%	4,80	156,8	0,2306
Total	2.060.286	100%	97.730	100,0%	2.158.015	100,0%	1,00	-	**0,8838**

O IV da variável quantidade de CPC é 0,8838.

Information Value

A partir do cálculo do VI gerado para todas as covariáveis do banco de dados, essa lista pode ser ordenada e uma maior importância inicial pode ser dada às variáveis mais fortes (com VI mais elevado), reduzindo sensivelmente o esforço de processamento na modelagem, dedicando atenção para um conjunto menor de variáveis.

Valor da informação (VI)	Classificação
<=0,02	Fraquíssima
Entre 0,02 e 0,10	Fraca
Entre 0,10 e 0,30	Média
Entre 0,30 e 0,50	Forte
> 0,50	Suspeita

Se o VI for superior a 0,50, a variável tem poder preditivo “suspeito” ou “alto demais” e é importante checar a consistência do dado, bem como a relevância prática dessa variável, assim como, eventualmente, o modo como foi gerada.

Information Value

Um valor de IV de 0,8838 é considerado com forte poder preditivo, sendo um excelente candidato a compor o modelo de Regressão Logística.

Vale a pena salientar, como esta é uma análise bidimensional, pode ocorrer de uma covariável apresentar alto valor de IV, dado a presença das demais covariáveis do modelo (visão multidimensional), uma determinada covariável não ser relevante .

Quantidade de CPC (contatos com a pessoa certa)	Cliente Pagou a Dívida?								
	Não (0)		Sim (1)		Total	%	SIM/NÃO	WOE	INF. VALUE
	#	%	#	%					
Zero	1.366.039	66,3%	23.128	23,7%	1.389.167	64,4%	0,36	-103,0	0,4393
Um	289.587	14,1%	20.356	20,8%	309.942	14,4%	1,48	39,3	0,0266
Dois	139.243	6,8%	13.583	13,9%	152.826	7,1%	2,06	72,1	0,0515
Três	82.422	4,0%	8.202	8,4%	90.624	4,2%	2,10	74,1	0,0325
Quatro	103.220	5,0%	14.305	14,6%	117.525	5,4%	2,92	107,2	0,1032
Cinco	79.775	3,9%	18.157	18,6%	97.932	4,5%	4,80	156,8	0,2306
Total	2.060.286	100%	97.730	100,0%	2.158.015	100,0%	1,00	-	**0,8838**
									FORTE

Case: Cancelamento

Um diretor de retenção ao cliente de uma Telecom deseja criar um modelo para calcular a probabilidade de cancelamento para decisão de retenção ativa.

a) Calcule o IV das variáveis. Quais variáveis possuem maior IV?

Arquivo Cancelamento.xlsx

2. Estatística VIF

VIF

Quando trabalhamos com dados multivariados podem existir variáveis que carreguem informações similares. Esse tipo de redundância, quando falamos de análise de regressão, chama-se **multicolinearidade**.

- ✓ Multicolinearidade é um efeito que ocorre quando variáveis fortemente correlacionadas são consideradas no mesmo modelo de regressão.

- ✓ Quando elas entram simultaneamente no modelo, a consequente redundância potencialmente eleva o nível de variabilidade dos pesos das variáveis correlacionadas, distorcendo sua interpretação e tornando o modelo instável e, muitas vezes, alterando a lógica esperada das variáveis envolvidas.

VIF

VIF (Variance Inflation Factor) ou Fator de Inflação da Variância: Mede o quanto variância dos coeficientes de regressão (pesos) são inflacionados por problemas de multicolinearidade.

Um valor de **VIF>5 já revela sinais de multicolinearidade**.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Case: Cancelamento

Um diretor de retenção ao cliente de uma Telecom deseja criar um modelo para calcular a probabilidade de cancelamento para decisão de retenção ativa.

a) Verifique se há multicolinearidade no modelo final por meio do VIF.

Arquivo Cancelamento.xlsx

3. Ponto de Corte

Pontos de Corte

A Tabela de Classificação apresenta o cruzamento da variável resposta observada em comparação com a variável resposta predita pelo modelo. Ela também é conhecida como Matriz de Confusão.

Um bom ajuste de modelo apresenta grande concentração de casos na diagonal principal.

Tabela de Classificação avaliada no ponto de corte:

		Variável Resposta Predita		Total
		0	1	
Variável Resposta Observada	0	VN	FP	VN+FP
	1	FN	VP	FN+VP
	Total	VN+FN	FP+VP	VN+FN+ FP+VP[1]

[1] *VP*: verdadeiro-positivo; *VN*: verdadeiro-negativo; *FP*: falso-positivo e *FN*: falso-negativo

Pontos de Corte

Tabela de Classificação avaliada no ponto de corte:

		Variável Resposta Predita		Total
		0	1	
Variável Resposta Observada	0	VN	FP	VN+FP
	1	FN	VP	FN+VP
	Total	VN+FN	FP+VP	VN+FN+ FP+VP[1]

[1] *VP*: verdadeiro-positivo; *VN*: verdadeiro-negativo; *FP*: falso-positivo e *FN*: falso-negativo

- Acurácia $$Acur = \frac{VP+VN}{VP+VN+FP+FN}$$
- Sensibilidade $$Sens = \frac{VP}{VP + FN}$$
- Especificidade $$Espec = \frac{VN}{FP + VN}$$

Os índices de **acurácia**, **sensibilidade** e **especificidade** variam de 0 a 1 (ou de 0% a 100%).

Esperamos que os índices sejam superiores a 50% (acima do acerto aleatório), sendo os valores mais próximos de 100% com maior poder preditivo.

Na prática, valores acima 70%-75% com ótimo desempenho.

Case: Cancelamento

Um diretor de retenção ao cliente de uma Telecom deseja criar um modelo para calcular a probabilidade de cancelamento para decisão de retenção ativa.

a) Obtenha diversos pontos de corte e calcule a acurácia, sensibilidade e especificidade para eles.

Arquivo Cancelamento.xlsx

Exemplo de Faixas de probabilidade

É possível realizar uma simulação para diversos pontos de corte e verificar as alterações nos valores de acurácia, sensibilidade e especificidade.

Ponto de Corte	Acurácia	Sensibilidade	Especificidade
0,1	0,3361	0,9543	0,1779
0,15	0,5565	0,8439	0,4830
0,2	0,6921	0,6706	0,6976
0,25	0,7593	0,5096	0,8232
0,3	0,7845	0,3775	0,8886

Pontos de Corte

O ponto de corte afeta os cálculos dos indicadores de qualidade de um modelo.

É possível calcular o ponto de corte que:

- Acurácia $$Acur = \frac{VP + VN}{VP + VN + FP + FN}$$ Maximiza a acurácia
- Sensibilidade $$Sens = \frac{VP}{VP + FN}$$
- Especificidade $$Espec = \frac{VN}{FP + VN}$$

Pontos de Corte

O ponto de corte afeta os cálculos dos indicadores de qualidade de um modelo.

É possível calcular o ponto de corte que:

- Acurácia $$Acur = \frac{VP + VN}{VP + VN + FP + FN}$$
- Sensibilidade $$Sens = \frac{VP}{VP + FN}$$
- Especificidade $$Espec = \frac{VN}{FP + VN}$$

Minimiza a diferença entre sensibilidade e especificidade, gerando maior equilíbrio entre os acertos dos 0 e 1.

Case: Cancelamento

Um diretor de retenção ao cliente de uma Telecom deseja criar um modelo para calcular a probabilidade de cancelamento para decisão de retenção ativa.

a) Obtenha o ponto de corte que maximiza a acurácia e que minimiza a diferença entre sensibilidade e especificidade.

Arquivo Cancelamento.xlsx

Exemplo

O ponto de corte que maximiza a acurácia é 0.8608, porém a diferença entre a sensibilidade (0,0005) e especificidade (1) é muito grande.

optimal_cutpoint	accuracy	acc	sensitivity	specificity	tp	fn	fp	tn
0.8608	0.7964	0.7964	0.0005	1	1	2036	0	7963

Exemplo

É possível buscar o maior equilíbrio, ou seja, a menor diferença entre sensibilidade e especificidade. No exemplo, o ponto de corte é 0,1967.

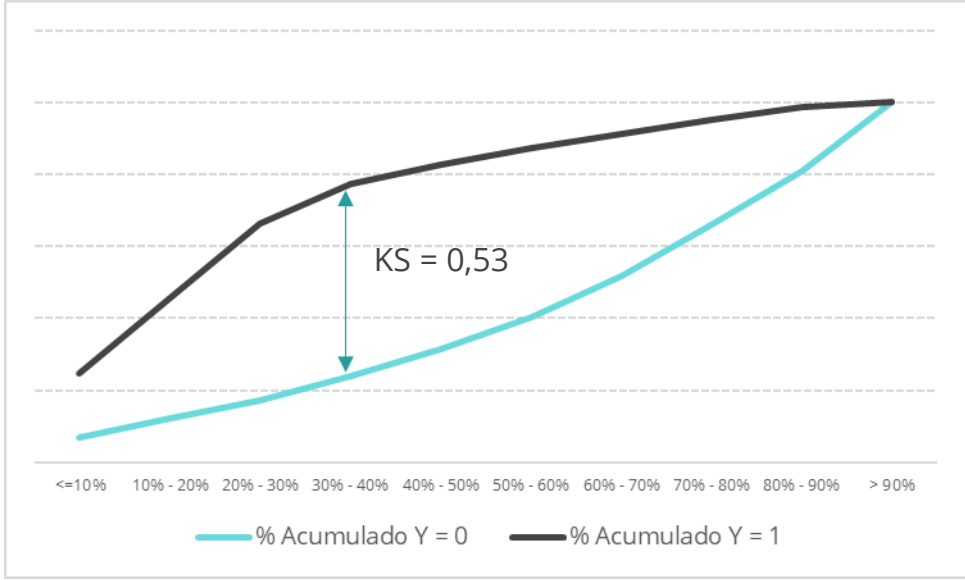
optimal_cutpoint	abs_d_sens_spec	acc	sensitivity	specificity	tp	fn	fp	tn
0.1967	0.0003	0.6846	0.6848	0.6845	1395	642	2512	5451

4. KS e AUC

KS

O indicador Kolmogorov-Smirnov (KS) reflete a máxima separação (em termos absolutos) entre as curvas acumuladas de dois grupos distintos quaisquer (virou consultor/não virou consultor ou inadimplente/adimplente). O indicador varia entre 0 (nenhuma separação) a 1 (separação completa).

Faixa de Probabilidade	% Y = 0	% Acumulado Y = 0	% Y = 1	% Acumulado Y = 1
<=10%	6,8%	6,8%	24,5%	24,5%
10% - 20%	5,4%	12,2%	21,1%	45,6%
20% - 30%	4,9%	17,1%	20,6%	66,3%
30% - 40%	6,9%	24,0%	10,9%	77,2%
40% - 50%	7,5%	31,5%	5,6%	82,8%
50% - 60%	8,9%	40,5%	4,6%	87,4%
60% - 70%	11,3%	51,8%	3,8%	91,2%
70% - 80%	14,2%	65,9%	3,8%	95,0%
80% - 90%	15,1%	81,0%	3,5%	98,5%
> 90%	19,0%	100,0%	1,5%	100,0%



Estatística KS

Kolmogorov-Smirnov propuseram uma metodologia para comparação das distribuições de uma variável aleatória em duas populações [Conover, 1999].

Sejam as populações de distribuição:

- S1(*c*): distribuição acumulada do evento 1 nas *c* faixas de probabilidades.
- S0(*c*): distribuição acumulada do evento 0 nas *c* faixas de probabilidades.

✓**Objetivo:** comparar os valores da distribuição S1(*c*) com os valores de S0(*c*); ou seja, nesse caso, a estatística KS será:

$$KS = \max_{1 \le c \le q} \left| S1(c) - S0(c) \right|,$$

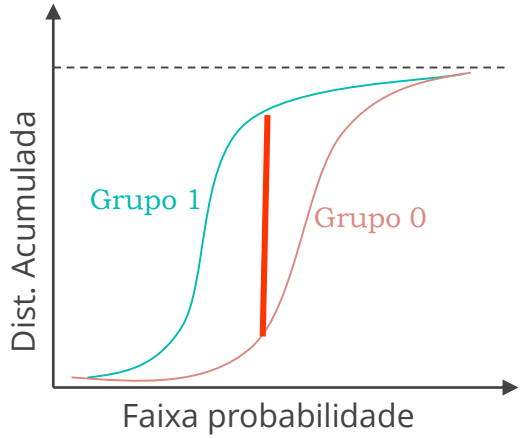
em que *c* é o índice que percorre as faixas de probabilidades de 1 a q.

Estatística KS

Faixa	0	1	KS
1	5,3%	0,4%	4,9%
2	10,4%	0,8%	9,6%
3	15,7%	1,1%	14,6%
4	21,0%	1,5%	19,5%
5	26,1%	1,8%	24,3%
6	31,4%	2,4%	29,0%
7	36,7%	3,1%	33,6%
8	41,9%	3,9%	38,0%
9	47,1%	5,0%	42,1%
10	52,3%	6,1%	46,2%
11	57,5%	7,3%	50,2%
12	62,6%	9,0%	53,6%
13	67,8%	11,1%	56,6%
14	72,9%	13,6%	59,3%
15	78,0%	16,6%	61,4%
16	83,1%	20,6%	62,5%
17	88,1%	25,3%	62,8%
18	93,0%	32,2%	60,8%
19	97,7%	43,6%	54,0%
20	100,0%	100,0%	0,0%

O máximo da diferença será o KS do modelo.

Para cada faixa de score, é calculada a diferença absoluta da distribuição acumulada do grupo 0 e do grupo 1.

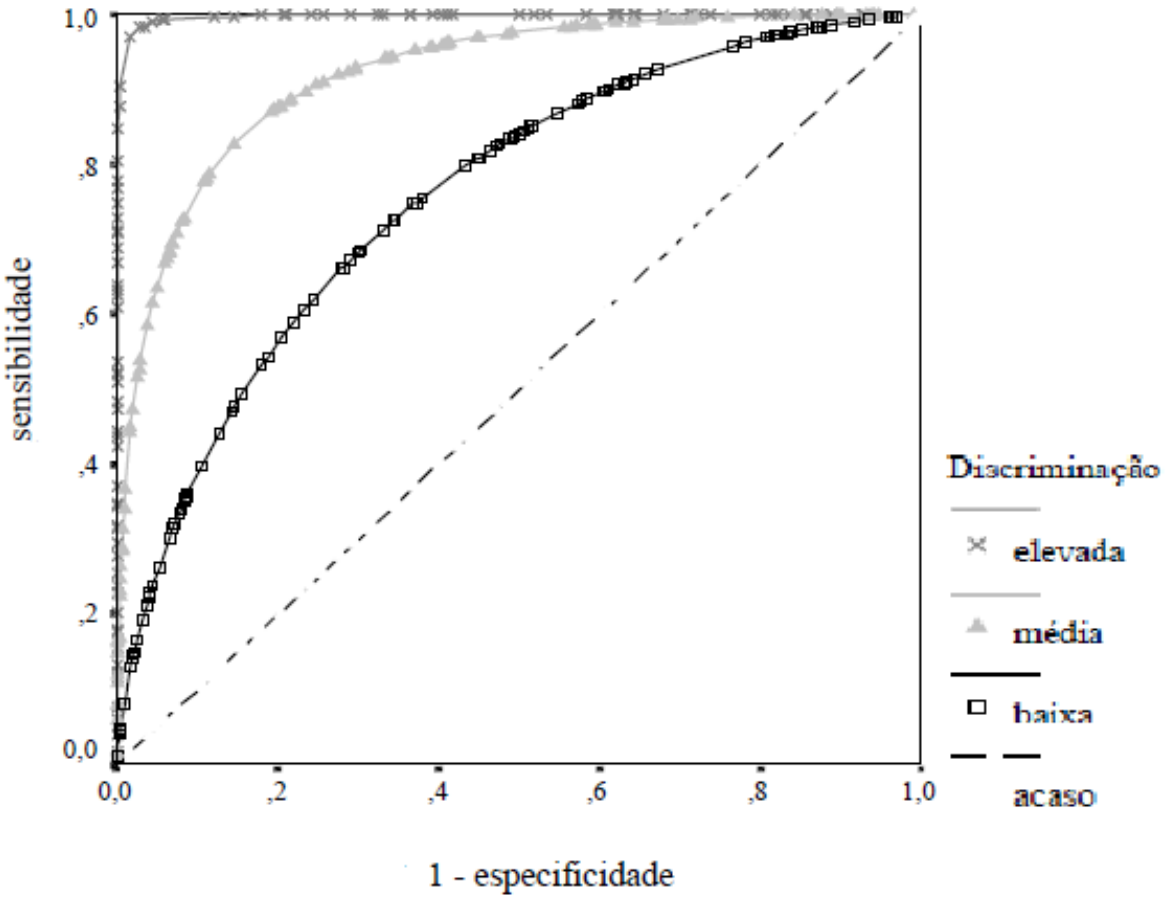


Curva ROC

- A análise ROC (*Reciever Operating Curve*) foi desenvolvida entre 1950 e 1960 para avaliar a detecção de sinais em radar e na psicologia sensorial.
- O maior problema da **Sensibilidade** e da **Especificidade** é que estas medidas dependem de um valor de corte, o qual é, às vezes, definido arbitrariamente.
- Assim, mudando o critério, pode-se aumentar a **Sensibilidade** com o consequente detrimento da **Especificidade**, ou vice-versa.
- A curva ROC permite estudar a probabilidade do verdadeiro-positivo (Sensibilidade) em função do falso-positivo (1 - Especificidade) para diferentes valores de corte.

Curva ROC

Gráfico bidimensional: **Sensibilidade** x **1 – Especificidade**.



Case: Cancelamento

Um diretor de retenção ao cliente de uma Telecom deseja criar um modelo para calcular a probabilidade de cancelamento para decisão de retenção ativa.

a) Obtenha o KS e a curva ROC do modelo final.

Arquivo Cancelamento.xlsx

5. Exercícios

Case: Cancelamento

Um diretor de retenção ao cliente de uma Telecom deseja criar um modelo para calcular a probabilidade de cancelamento para decisão de retenção ativa.

a) Calcule o IV das variáveis. Quais variáveis possuem maior IV?

a) Verifique se há multicolinearidade no modelo final por meio do VIF.

b) Obtenha o ponto de corte que maximiza a acurácia e que minimiza a diferença entre sensibilidade e especificidade.

c) Obtenha o KS e a curva ROC do modelo final.

Arquivo Cancelamento.xlsx

Case: Avaliação de risco de um empréstimo bancário

Considere os dados provenientes de um banco de varejo, referente a 5.000 propostas passadas de crédito, geradas para solicitação de um empréstimo. A base traz dados como idade, nível de instrução, tempo de experiência, tempo no endereço e renda, além da variável “classif” (0=bom, 1=mau). Vamos avaliar o potencial preditivo inicial dessas variáveis para predizer se um cliente será um mau pagador.

a) Calcule o IV das variáveis. Quais variáveis possuem maior IV?

a) Verifique se há multicolinearidade no modelo final por meio do VIF.

b) Obtenha o ponto de corte que maximiza a acurácia e que minimiza a diferença entre sensibilidade e especificidade.

c) Obtenha o KS e a curva ROC do modelo final.

Arquivo CasoUso_02_Emprestimo_Bancario.xlsx

Case: Fatores de influência no valor da remuneração mensal

Considere os dados de 534 profissionais, seus salários e algumas informações sociodemográficas. A ideia é tentar entender os fatores que mais influenciam na possibilidade do profissional ganhar um salário superior a 10 s.m. (salários mínimos). A partir da análise dessa base de dados, vamos tentar entender alguns elementos em torno desse tema, com a seguinte base de dados:

a) Calcule o IV das variáveis. Quais variáveis possuem maior IV?

a) Verifique se há multicolinearidade no modelo final por meio do VIF.

b) Obtenha o ponto de corte que maximiza a acurácia e que minimiza a diferença entre sensibilidade e especificidade.

c) Obtenha o KS e a curva ROC do modelo final.

Arquivo CasoUso_03_Fatores_Impacto_Salario.xlsx

Case: Cancelamento de seguro auto

Considere os dados de 2.143 segurados do produto seguro automóvel, de uma certa seguradora. A ideia nesse estudo é entender o que influencia o cancelamento do seguro auto e construir um modelo que explique a probabilidade de cancelamento, para estabelecer estratégias de retenção mais efetivas.

a) Calcule o IV das variáveis. Quais variáveis possuem maior IV?

a) Verifique se há multicolinearidade no modelo final por meio do VIF.

b) Obtenha o ponto de corte que maximiza a acurácia e que minimiza a diferença entre sensibilidade e especificidade.

c) Obtenha o KS e a curva ROC do modelo final.

Arquivo
CasoUso_04_Cancelamento_Seguro_Auto.xlsx

Referências

LIVROS-TEXTO & ARTIGOS

- Agresti, A. (2002). *Categorical data analysis* (Vol. 359). Wiley-Interscience.
- Conover, W. J. (1999). *Practical nonparametric statistics*. New York: Wiley.
- Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861-874.
- Hosmer, D. W. e Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd ed. New York: Wiley.