Análise sobre renda pessoa global em 1994

Paulo Henrique Brasil Ribeiro
UFSCar, ICMC

Milene Regina dos Santos
UFSCar, ICMC

Resumo Na atualidade, o volume de dados existentes sobre as pessoas como um todo tem sido cada vez maior. Essas informações podem ser utilizadas entre outras coisas para construir informações e posteriormente análises e ilações sobre os indivíduos estudados. Dados mais "completos" podem ser utilizados para captação de padrões, possibilitam deduzir sobre os demais. No presente trabalho, utilizando técnicas de Regressão Logística num banco de treinamento no intuito de inferir se a renda das pessoas no banco de validação excede 50 mil doláres. Também faremos a apuração de precisão destas predições e demais métricas referentes ao modelo ajustado.

Keywords: stylesheet, glossa, article

1 Introdução

Seção 1

As presentes informações foram extraídas por Barry Becker no banco de dados Censo de 1994 e um dos locais onde está disponível disponibilizada em University of California, School of Information and Computer Science - Machine Learning Repository [1].

Os dados de treinamento são compostos inicialmente de 32561 observações e 13 variáveis:

- Idade
- Trabalho
- Escolaridade
- Anos de estudo
- Estado civil
- Profissição
- Raça
- Sexo
- Ganho de capital
- Perda de capital
- Horas trabalhadas (por semana)
- Nacionalidade
- Renda anual
- Grupo (<=50k, >50k) - y

Enquanto os dados de teste ou validação compõe as mesma variáveis porém com 16282 observações. Considerado uma quantidade relativamente grande de observações em ambos as partições, decidiu-se manter essa proporção de 67% e 33%

Sendo variável resposta (y) uma variável binária categória (relacionada ao evento de interesse a Renda anual ser maior do que \$50.000), decidiu-se efetuar a modelagem com o através de regressão logística. Para os aspectos inferenciais a seguir, admitiu o nível de significância ($\alpha = 0,01$) visto a quantidade de expressiva observações.

[1] Acessado em 09/jul/2022 - 11h40m

1.1 Regressão Logística

Assim como as demais distribuições abrangidas pelas técnicas de MLG, a distribuição binomial pertence a família exponencial conforme as equações (1),

$$f(y;\pi) = \binom{m}{n}\pi^y(1-\pi)^{m-y}, \quad \pi \in [0;1], \quad y = 0,1,\ldots,m \tag{1}$$

$$\phi = 1, \;\; \theta = log\left(\frac{\pi}{1-\pi}\right) = \left(\frac{\mu}{m-\mu}\right) \implies \mu = \frac{me^{\theta}}{1+e^{\theta}}$$
$$b(\theta) = -m\, log(1-\pi) = m\, log(1+e^{\theta}), \;\; c(y,\phi) = log\binom{m}{y}$$

Utilizar a função de ligação canônica "logit" possue como principais vantagens ser vastamente difundida nas mais diversas áreas devido à fácil interpretação de seus coeficientes. O comportamento dessa função, vide Figura 1.

Aplicada a função logit ao modelo tem-se,

$$\text{logit}(Y) = log\left(\frac{p}{1-p}\right) = X\beta \tag{2}$$

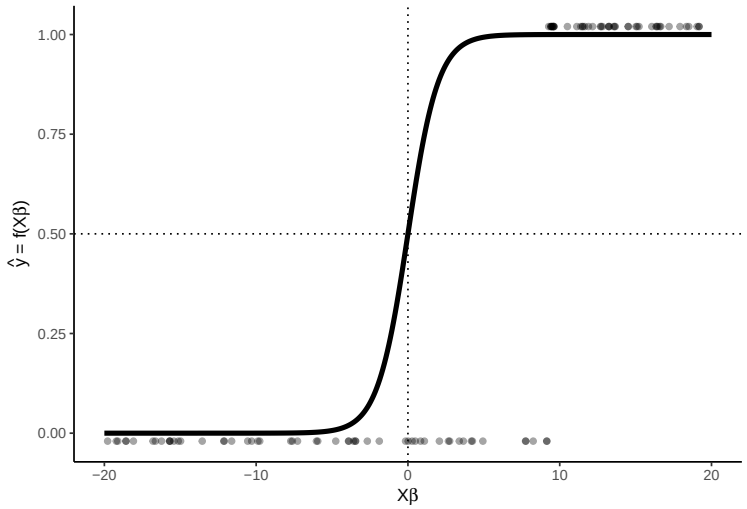

Figura 1: Função Logística (sigmoide)

A interpretabilidade facilitada por este tipo de modelo vem da maneira em que podem ler lidos os coeficientes do ajuste, razão de chances $(e^{\beta_i}, i = 1,\ldots,p)$, . A razão de chances quando vista numa variável regressora categórica é uma medida de associação que mensura a proporção em que o indivíduo detentor de uma certa característica possue em estar no grupo do evento de interesse, i.e. y=1, quando comparada com característica de referência desta mesma variável categórica.

De forma similar, para uma variável explicativa contínua a interpretação se dá como para cada acréscimo na variável (ε)o indíviduo possuem um acréscimo de $\dfrac{e^{\beta_i(x_i+\varepsilon)}}{e^{\beta_i x_i}} = e^{\beta_i \cdot \varepsilon}$ em estar no grupo do evento de interesse. Logo tem-se as seguintes alternativas de interpretação para cada β_i:

- Para $\beta_i > 0 \implies e^{\beta_i} > 1$, conforme a variável x_i aumentar o probabilidade de evento tende a aumentar,
- Para $\beta_i = 0 \implies e^{\beta_i} = 1$, conforme a variável x_i aumentar o probabilidade de evento tende a permanecer a mesma,
- Para $\beta_i < 0 \implies e^{\beta_i} < 1$, conforme a variável x_i aumentar o probabilidade de evento tende a diminuir.

Essas interpretações tem ampla exploração na Seção 4.

1.2 Motivação

Na atualidade, o volume de dados existentes sobre as pessoas como um todo tem sido cada vez maior. Essas informações podem ser utilizadas entre outras coisas para construir informações e posteriormente análises e ilações sobre os indivíduos estudados. Dados mais "completos" podem ser utilizados para captação de padrões, possibilitam deduzir sobre os demais.

No presente trabalho, utilizando técnicas de Regressão Logística num banco de treinamento no intuito de inferir se a renda das pessoas no banco de validação excede 50 mil doláres. Também faremos a apuração de precisão destas predições e demais métricas referentes ao modelo ajustado.

2 Análise inicial e Tratamento dos dados

2.1 Transformação e Seleção de variáveis

Reduziu-se a capilaridade de muitas das variáveis categóricas para que a estimação fosse possível e além do que diminui uma quantidade de parâmetro que poderia impossibilitar o bom entendimento do modelo, além do que algumas das categóricas pouco ou nada acrescentavam na elucidação das relações entre as covariáveis entre si e delas com a variável resposta.

Essa transformação ocorreu com a ajuda da técnica CHAID (Chi-Squared Automatic Interaction Detector) já bem estabelecida na literatura, porém o maior peso para as decisões foi proveniente do bom senso na aglomeração dos atributos, por exemplo a variáveis "países" foi sumarizada em "continentes" (cntl), a variável profissão resumiu-se na área de atuação do profissional (administrativo, serviço, etc) e outros agrupamentos similares. Um exemplo da técnica CHAID em anexo

2.2 Associação

Os coeficientes do modelo também podem ser calculados de forma simples através de uma tabela de contingência dos dados onde $\frac{P(Y=1|X=x_i)}{1-P(Y=1|X=x_i)} = \frac{p}{1-p}$, consequência direta da equação (2). Portanto avaliar a forma como a variável se distribui em suas pares não deixar de ser uma opção de ter-se uma vaga idéia de como cada variável iria se comportar isolada numa regressão.

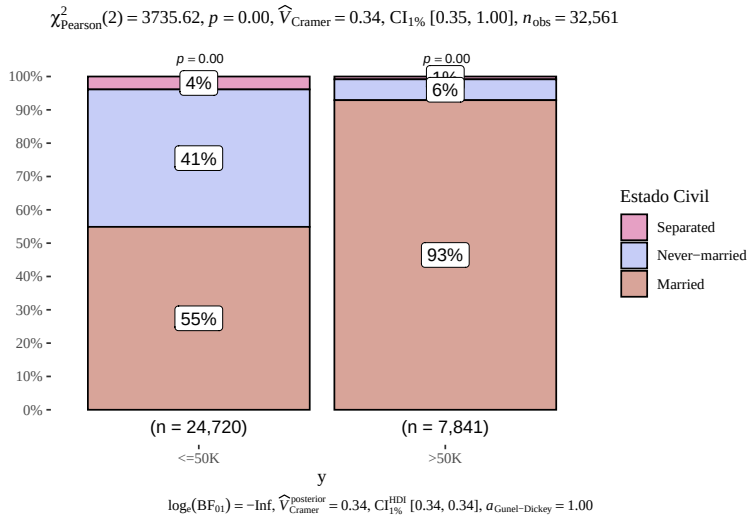

Figura 2: est

Por meio disso temos uma idéia de qual variável tem chance de ser uma ótima preditora (vide Figura 2), devido à disparidade no proporção de eventos de interesse nas suas categórias. E quais influenciaram entre pouco a nada na resposta (vide Figura (3)), lembrando sempre de que essa forma de análise deve ser usada sempre com extrema cautela, sendo somente um leve guia durante a construção do ajuste e que ao adicionar outras preditoras na regressão todo esse pode mudar por completo devido a interações e etc.

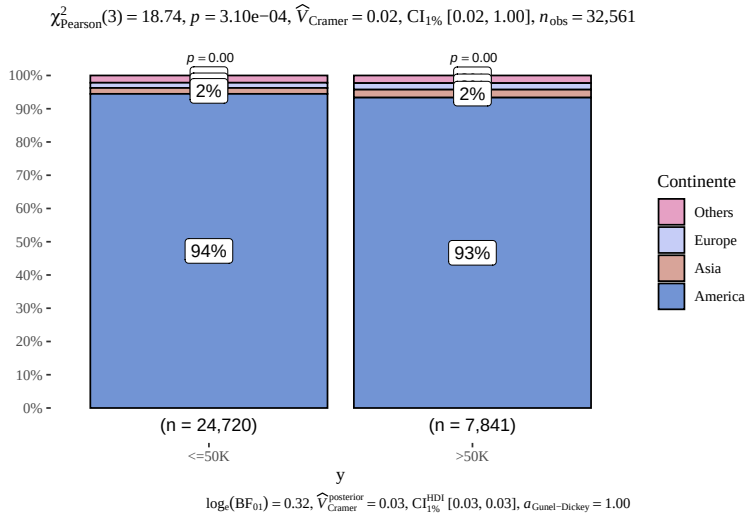

Figura 3: cntn

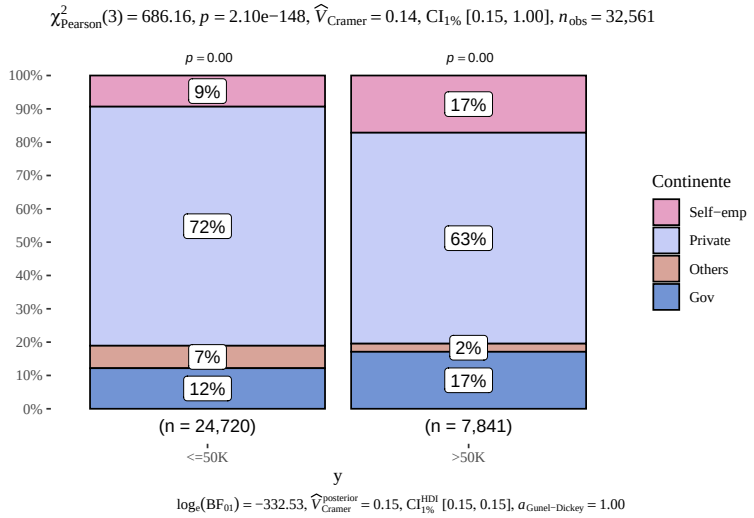

Figura 4: trbl

2.3 Modelo encaixados - Aula 21 - 45

Para seleção do modelo utilizou-se a estratégia com uma sequência de modelos encaixados nas duas "direções":

- Começar do modelo nulo e adicionar uma nova variável categórica, que é escolhida por ser aquela que mais aumenta a 'Deviance' [vide equação (2.3)]entre os os modelos, efetuado isso repetidas vezes até que a adição de nenhuma variável seja significativa através do teste de F para um dado nivel de significância e após isso avalia-se as variáveis quantitavias,
- A partir do modelo saturado e remover uma nova variável categórica, que é escolhida por ser aquela que menos aumenta 'Deviance' entre os modelos, efetuado isso repetidas vezes até que a remoção nenhuma variável seja significativa através do teste de F para um dado nivel de significância, após isso avalia-se as variáveis quantitavias. Conforme exemplificado nas Tabelas **??** e **??** nos Anexos.

Onde a estatística de teste se dá por:

$$D^* = -2\sum_{i=1}^{n}\left[y_i\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_1)\log\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right]$$

Outro método similar de seleção método de variáveis regressoras muito utilizado que também aplicou-se nesse estudo foi o stepwise AIC. Este método possibilita a determinação de um conjunto de variáveis estatisticamente significantes ao utilizar os critérios de AIC num conjunto de ajustes avaliados.

Estes três procedimentos (ANODEV forward, ANODEV backwar e stepwise AIC) foram aplicados aos modelos de primeira e segunda ordem (com termos quadráticas e interações das covariáveis 2 a 2).

2.4 Ajuste Final

Ao final das análises houve quatro ajustes indicados, após às 5 comparações em pares, segue na Tabela **??** última comparação entre os dois ajuste mais plaúsiveis.

Tabela 1: Comparação entre os dois modelos últimos indicados pelas técnicas usadas

Ajuste	Resid Df	Resid. Dev	Grau de lib.	Deviance	Estatística F	Valor de P
Modelo primeira ordem	32538	23935,22	NA	NA	NA	NA
Modelo segunda ordem	32517	23801,39	21	133,832	6,373	0

Este modelo de segunda ordem segue com preditor linear composto pelas covariáveis e seus respectivos coeficientes a seguir,

$$
\begin{aligned}
\hat{\mu} =& X\beta \\
=& \beta_0 + \beta_{1a} \cdot \text{estNever-married} + \beta_{1b} \cdot \text{estSeparated} + \\
& \beta_2 \cdot \text{ans} + \beta_3 \cdot \text{cptl} + \beta_4 \cdot \text{sexMale} + \beta_{8c} \cdot \text{trblSelf-emp} + \\
& \beta_{9a} \cdot \text{raceAsian-Pac-Islander} + \beta_{9b} \cdot \text{raceBlack} + \beta_{9b} \cdot \text{raceOther} + \\
& \beta_{9c} \cdot \text{raceWhite} + \beta_{10a} \cdot \text{edccLiberal} + \beta_{10b} \cdot \text{edccNon-Grad} + \\
& \beta_{11a} \cdot \text{estNever-married:trblOthers} + \beta_{11b} \cdot \text{estNever-married:trblPrivate} + \\
& \beta_{11c} \cdot \text{estNever-married:trblSelf-emp} + \beta_{12a} \cdot \text{estSeparated:trblOthers} + \\
& \beta_{12b} \cdot \text{estSeparated:trblPrivate} + \beta_{12c} \cdot \text{estSeparated:trblSelf-emp} + \\
& \beta_{13} \cdot \text{estNever-married:sexMale} + \beta_{14} \cdot \text{estSeparated:sexMale} + \\
& \beta_{15a} \cdot \text{sexMale:trblOthers} + \beta_5 \cdot \text{hr} + \beta_{6a} \cdot \text{classeOthers} + \\
& \beta_{6b} \cdot \text{classeProduction} + \beta_{6c} \cdot \text{classeService} + \beta_7 \cdot \text{idad} + \\
& \beta_{8a} \cdot \text{trblOthers} + \beta_{8b} \cdot \text{trblPrivate} + \beta_{8c} \cdot \text{trblSelf-emp} + \\
& \beta_{15a} \cdot \text{sexMale:trblOthers} + \beta_{15b} \cdot \text{sexMale:trblPrivate} + \\
& \beta_{15c} \cdot \text{sexMale:trblSelf-emp} + \beta_{16a} \cdot \text{sexMale:classeOthers} + \\
& \beta_{16b} \cdot \text{sexMale:classeProduction} + \beta_{16c} \cdot \text{sexMale:classeService} + \\
& \beta_{17a} \cdot \text{trblOthers:edccLiberal} + \beta_{17b} \cdot \text{trblOthers:edccNon-Grad} + \\
& \beta_{18a} \cdot \text{trblPrivate:edccLiberal} + \beta_{18b} \cdot \text{trblPrivate:edccNon-Grad} + \\
& \beta_{19a} \cdot \text{trblSelf-emp:edccLiberal} + \beta_{19b} \cdot \text{trblSelf-emp:edccNon-Grad} + \\
& \beta_{20a} \cdot \text{estNever-married:edccLiberal} + \beta_{20b} \cdot \text{estNever-married:edccNon-Grad} + \\
& \beta_{21a} \cdot \text{estSeparated:edccLiberal} + \beta_{21b} \cdot \text{estSeparated:edccNon-Grad}
\end{aligned}
\tag{3}
$$

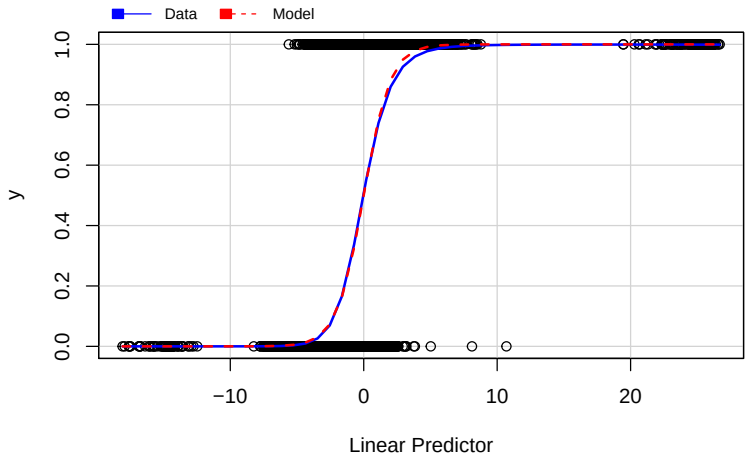

Figura 5: Sigmoide de probabilidade ajustado contra valores reais de y

3 Análise e diagnóstico

3.1 Multicolineariedade

Ao utilizar se da constância limite $k = 5$, pode-se concluir que não há evidência estatísticas suficientes afirmar que existe multicolineariedade no ajuste. Conforme Tabela **??**.

3.2 Resíduo - Aula 22 50min

Tabela 2: Estatística VIF para os parâmetros do modelo

Parâmetros	GV1IF	Grau de liberdade	GVIF^1/(2*Df)
sex:classe	1022.763578	3	3.174163
edcc	97.451588	2	3.141935
classe	664.251119	3	2.953852
est_	53.957506	2	2.710273
sex	7.098680	1	2.664335
sex:trbl	241.397108	3	2.495296
trbl	134.718473	3	2.264146
est_:sex	10.682382	2	1.807869
trbl:edcc	123.318485	6	1.493662
est_:trbl	37.193358	6	1.351674
ans_	1.749465	1	1.322673
idad	1.196625	1	1.093904
est_:edcc	1.819401	4	1.077683
hr___	1.106560	1	1.051931
cptl	1.017424	1	1.008675
race	1.039242	4	1.004823

3.3 Ponto de corte

Um ponto muito importante na regressão logística é a escolha do ponto de corte (cut-off) que será utilizado para definir apartir de qual probabilidade estimada será considerado

como evento de interesse ou não. Essa medida impacta de forma decisiva as medidas de análise dos modelos que serão analisadas na Seção 3.5, entre elas: prevalência, acurácia, sensibilidade, poder preditivo positivo, KS, correlação de Matheus, AUC.

Devido sua importância dentre os métodos aplicados, para maior brevidade comentar-se-á aqui sobre somente três destes métodos.

3.4 Curva Receiver Operating Characteristic (ROC) e outros

O primeiro método é através da visualização e análise da curva ROC. A curva ROC (Zweig & Campbell, 1993) tabula os pontos de corte, ao longo da amplitude dos escores fornecidos pelos modelos, e uma curva é contruída no plano cartesiano 'Sensibilidade vs (1-Especificidade)', eixo horizontal e vertical respectivamente. A fim de referenciar a qualidade das medidas, em geral, coloca-se uma reta onde simula a predição de y ao acaso, ou seja sem a utilização nenhuma ferramenta de predição. Com isso a área entre a curva e esta reta diagonal (Area Under the Curve - AUC) é usada como uma métrica que pode auxilia a avaliar a qualidade do modelo.

Segue na Figura 6 a curva ROC do ajuste aos dados presentes, Com AUC de 0,497 há indício de que o modelo tenha um bom poder preditivo.

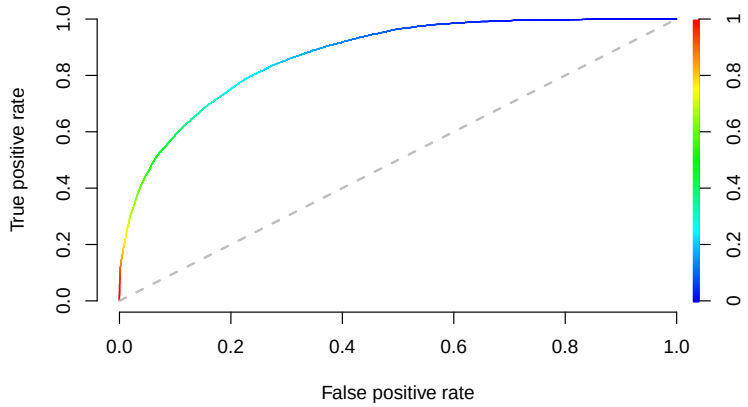

Figura 6: Curva ROC, com AUC = 0,8701 e ponto ótimo estipulado pela função custo de 0,497

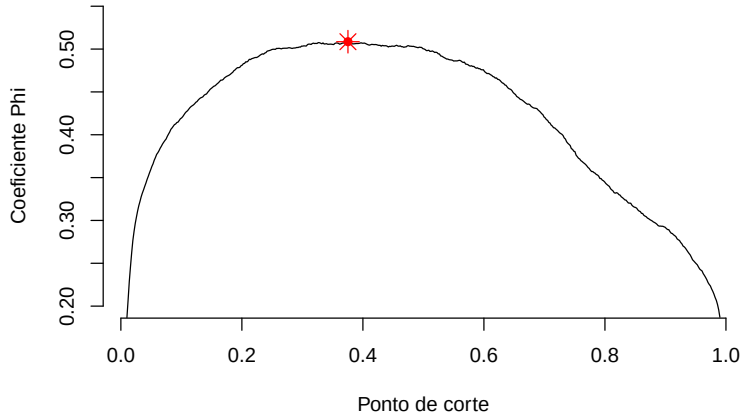

Figura 7: Busca de máximo no Coeficiente de Phi ao longo dos pontos de corte

Procurou-se também pontos de cortes que maximizassem (Figura **??**) a Correlação de Mathews (Coeficiente de Phi) obtendo um valor de 0,308, algo relativamente próximo do resultado anterior.

3.5 Teste KS e teste de Hosmer Lemeshow

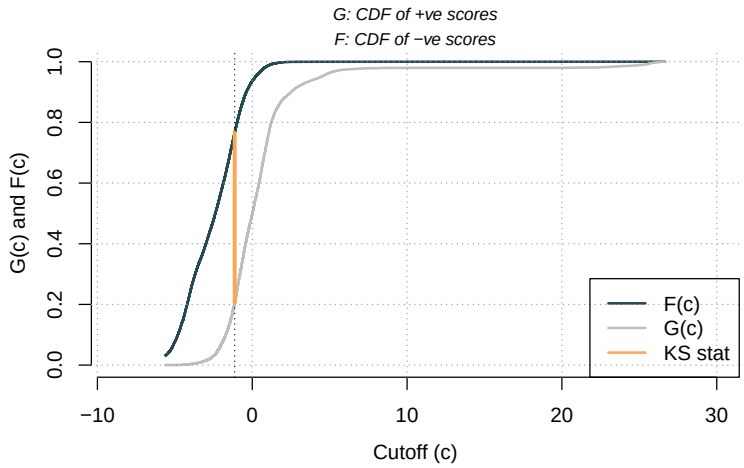

Figura 8: caption

```
## KS: 0.5624863
```

KS: 0.5624863
Segundo Louzada, Diniz et al:

- KS $<$ 10%: indica que n˜ao h´a discrimina¸c˜ao entre os perfis de bons e maus clientes;
- 10% $<$ KS $<$ 20%: indica que a discrimina¸c˜ao ´e baixa;
- KS $>$ 20%: indica que o modelo discrimina o perfil de bons e maus.

3.6 Outliers e Pontos de influência

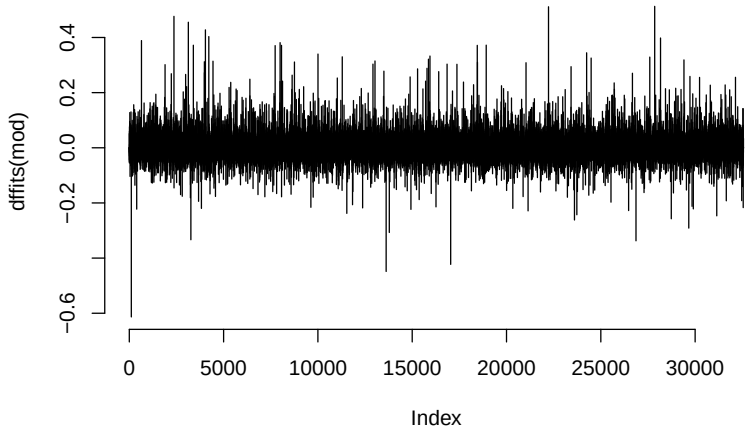

Figura 9: caption

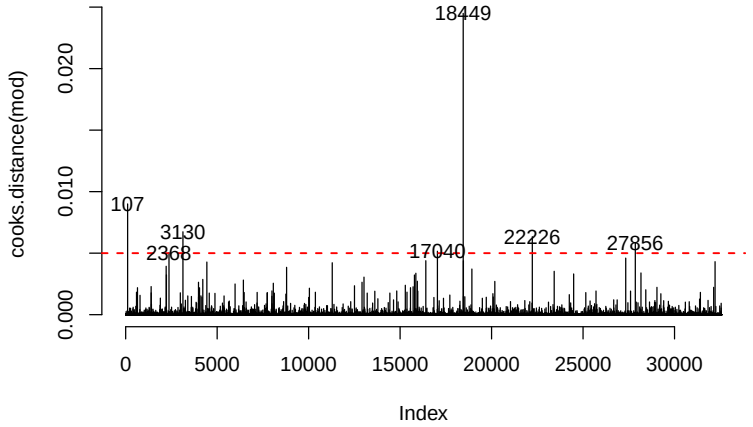

Figura 10: caption

3.7 Comparação nas estimativas com a retirada dos pontos influentes

Tabela 3: eopvno

	Modelo Final	Modelo sem pontos influentes	Diferença (%)
(Intercept)	-7,937	-7,936	0
est_Never-married	-1,795	-1,789	0,003
est_Separated	-1,94	-2,25	-0,16
ans_	0,351	0,351	0
cptl	0	0	-0,015
sexMale	1,369	1,375	-0,005
hr___	0,029	0,029	0,002
classeOthers	-0,227	-0,232	-0,022
classeProduction	-0,141	-0,143	-0,01
classeService	-0,562	-0,564	-0,002
idad	0,021	0,021	0,005
trblOthers	-0,112	-0,093	0,169
trblPrivate	0,037	0,042	-0,125
trblSelf-emp	0,143	0,135	0,055
raceAsian-Pac-Islander	0,556	0,554	0,004
raceBlack	0,411	0,418	-0,016
raceOther	-0,028	-0,025	0,116
raceWhite	0,713	0,717	-0,006
edccLiberal	-0,254	-0,28	-0,104
edccNon-Grad	-0,019	-0,021	-0,113
est_Never-married:trblOthers	-0,834	-0,679	0,186
est_Separated:trblOthers	-9,764	-10,332	-0,058
est_Never-married:trblPrivate	-0,172	-0,177	-0,032
est_Separated:trblPrivate	0,074	0,269	-2,645
est_Never-married:trblSelf-emp	0,634	0,632	0,004

	Modelo Final	Modelo sem pontos influentes	Diferença (%)
est_Separated:trblSelf-emp	-0,233	-0,177	0,243
est_Never-married:sexMale	-0,534	-0,541	-0,013
est_Separated:sexMale	0,407	0,612	-0,502
sexMale:trblOthers	-0,708	-0,735	-0,039
sexMale:trblPrivate	-0,036	-0,045	-0,268
sexMale:trblSelf-emp	-0,54	-0,535	0,009
sexMale:classeOthers	-0,162	-0,158	0,027
sexMale:classeProduction	-0,667	-0,668	-0,003
sexMale:classeService	0,027	0,027	0
trblOthers:edccLiberal	-0,208	-0,312	-0,501
trblPrivate:edccLiberal	-0,188	-0,151	0,198
trblSelf-emp:edccLiberal	0,562	0,604	-0,076
trblOthers:edccNon-Grad	-0,546	-0,417	0,236
trblPrivate:edccNon-Grad	0,023	0,02	0,16
trblSelf-emp:edccNon-Grad	0,904	0,905	0
est_Never-married:edccLiberal	0,684	0,656	0,041
est_Separated:edccLiberal	0,735	0,374	0,491
est_Never-married:edccNon-Grad	0,721	0,782	-0,084
est_Separated:edccNon-Grad	-0,713	-10,933	-14,34

3.8 Matriz de confusão e predições

Com o aux´ılio da curva ROC podemos escolher um ponto de corte igual a 0,29. Assim, as medidas relacionadas ‘a capacidade preditiva do modelo s˜ao: SENS = 0, 75, SP EC = 0, 76, V P P = 0, 58, V P N = 0, 87, CAT = 0, 76 e MCC = 0, 48, o que ´e indicativo de uma boa capacidade preditiva. Esta conclus˜ao ´e corroborada pela curva ROC apresentada na Figura 2.1.

4 Conclusões e interpretações dos parâmetros (razão de chances)

[2]

[2] Examples in footnotes are numbered with lower case Roman numerals enclosed between brackets:

(i) a. Colorless green ideas sleep furiously.
b. *The child seems sleeping.

More text can follow the example.

Anexo

Tabela 4: Exemplo de ANODEV, modelo nulo contra ajuste com estado civil e ajuste com estado civil e sexo

Resid Df	Resid. Dev	Grau de lib.	Deviance	Estatística F	Valor de P
32560	35948.08	NA	NA	NA	NA
32558	31457.66	2	4490.4191	2245.2096	0
32557	30513.21	1	944.4547	944.4547	0

Tabela 5: caption

	Parâmetros	Modelo nulo	Modelo com Estado civil	Modelo com Estado civil e sexo
1	(Intercept)	-1.14824625533759 ***	-0.622107617298178 ***	-1.42581989862586 ***
2		(0.0129610010863722)	(0.0145253090608884)	(0.0326961537290307)
3	est_Never-married		-2.41080661085251 ***	-2.29262322014075 ***
4			(0.0484286615530055)	(0.0487489387923965)
5	est_Separated		-2.05412871550575 ***	-1.7615489070328 ***
6			(0.128082388235369)	(0.129375612117635)
7	sexMale			1.03405734944672 ***
8				(0.0357227570989128)
1.1	N	32561	32561	32561
2.1	logLik	-17974.0397176114	-15728.8301530312	-15256.6028255339
3.1	AIC	35950.0794352229	31463.6603060625	30521.2056510678