UNIVERSIDADE FEDERAL DO PARANÁ
CURSO DE ESTATÍSTICA

Andressa Luiza Cordeiro GRR:20160218
Jayme Gomes dos Santos Junior GRR:20160210
Luciana Helena Kowalski GRR:20160231

MODELAGEM DE BANK MARKETING POR REGRESSÃO LOGÍSTICA

CURITIBA
2019

1 Resumo

O objetivo do trabalho foi realizar a modelagem de dados sobre marketing bancário vizando predizer a contratação do serviço pelo potencial cliente. Os dados foram extraídos do do site *mldata* (Bank Marketing). Para o ajuste do modelo, foi utilizado modelos lineares generalizados (GLM) da família binomial com resposta binária. Primeiramente foi ajustado o modelo com 80% das observações, após o mesmo foi validado com os 20% restantes. A seleção das covariáveis foi realizada através do *stepwise* com o critério de informação de Akaike (AIC). Devido oa elevado número de níveis em variáveis categóricas dos dados, foi utilizado o pacote *GAMLSS* visando agrupar níveis semelhantes. Para o cálculo da predição do modelo foi testado os pontos de corte de 0,3, 0,5 e 0,89. Posteriormente, foi calculada a sensibilidade e especificidade e a relação entre eles foi representada pela curva ROC. Foi realizada a regra de decisão para o modelo incorporando custos. O reagrupamente dos níveis de ambas as variáveis *month* e *job* gerou 5 níveis. O modelo gerado apresentou elevada especificidade e baixa sensibilidade. O ponto de corte da regra de decisão que gerou menor custo foi de 0,4.

2 Introdução

O marketing do setor bancário está em contante mudança, isso ocorre devido a facilidade de informação que acarreta em maior exigência por parte dos clientes. Com isso, os bancos constantemente estão alterando suas políticas de marketing e venda. No entanto, apesar da suma importância para conseguir novas fidelizações de clientes, é elevado o custo deste recurso. Além disso, nas últimas décadas aumentou significativamente a concorrência no setor bancário, o que maximiza a importância do marketing neste setor.

Neste contexto, o objetivo do trabalho foi realizar a modelagem de um programa de marketing telefônico bancário vizando predizer a adesão ou não de novos clientes a novos contratos.

3 Material e Métodos

A base de dados *Bank Marketing* utilizada no estudo foi extraída do site *"https://www.mldata.io/datasets/."*. Com as seguintes variáveis: **age** - idade do cliente(numérico inteiro); **job** - emprego do cliente = admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown(fator); **marital** - estado civil do cliente = divorced, married, single, unknown(fator); **education** - nível educacional do cliente = primary, secondary, tertiary, unknown(fator); **default** - cliente possui crédito = no, yes, unknown(fator); **balance**- balanço anual médio do cliente em euro(numérico); **housing** - cliente possui empréstimo habitacional = no, yes, unknown(fator); **loan** - cliente possui empréstimo pessoal = no, yes, unknown(fator); **contact** - forma de contato com o cliente = unknown, telephone, cellular(fator); **day** - dia do mês(inteiro entre 1 e 31); **month** - mês do ano = jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec(fator); **duration** - duração do último contato com o cliente(numérico em segundos); **campaing** - número de contatos feitos com o cliente(numérico inteiro); **pdays** - número de dias desde o último contato feito com o cliente em uma campanha passada(numérico inteiro onde -1 significa que não houve contato prévio); **y** - variável resposta = yes, no(fator).

As análises estatísticas foram realizadas através do software `R`.

Para realizar o ajuste do modelo foi utilizado GLM(*Generalized Linear Models*) da família binomial para resposta binária com as funções de ligação *logito*, *probito*, *complemento log-log* (*clog-log*) e *cauchit*.

A base de dados foi separada em duas, sendo uma para ajustar o modelo(com 80% das observações) e a outra para validação do ajuste(com os 20% restantes). Este procedimento foi realizado de forma aleatória.

Inicialmente, foram selecionada as covariáveis descritas anteriormente. Foram utilizados os método *forwar*, *backward* e *stepwise* para seleção de covariáveis, selecionando o modelo através do critério de informação de Akaike (AIC), utilizando o que produziu o menor valor (AIC).

Então foi utilizado um algorítmo de redução de níveis para ajustar as variáveis categóricas com muitos níveis. Assim ajustando novos modelos com as variáveis reajustadas para realizar uma nova seleção de modelo.

Posteriormente foi avaliada a predição do modelo através de pontos de corte, estabelidos como 0,5(de maneira intuitiva) e 0,89(baseado na proporção da variável resposta). Então, foram comparadas a sensibilidade e especificidade nos dois pontos de corte. Após, foi utilizada a curva *ROC* para representar a relação entre sensibilidade e especificidade para então usar regras de decisão baseadas em custo de classificação.

4 Resultado e Discussões

Após realizada a seleção das covariáveis, foi escolhido o modelo ajustado pelo algorítmo *stepwise* com o menor AIC.

Tabela 1: Modelos Binomiais, Suas Funções de Ligação e Valores AIC.

Função de Ligação	AIC	N° de Parâmetros
Probito	1821.418	34
Logito	1833.575	34
Cauchit	1941.460	23
Clog-log	1967.442	35

Muito embora o modelo **Binomial(probito)** tenha sido selecionado pelo menor AIC, todos os modelos candidatos se mostraram muito complicados(número grande de parãmetros) devido a covariáveis do tipo fator com muitos níveis(Tabela 1).

Para tentar lidar com este problema, será usado um método do pacote *GAMLSS* de combinar níveis parecidos e assim gerar novos níveis para variáveis do tipo fator baseado nas idéias de *Tutz(2013)* que, diferente de outros métodos que encolhem a diferença entre as estimativas dos níveis do fator em torno da média geral, ele classifica níveis em blocos de estimativas parecidas através de um valor $\boldsymbol{\lambda}$ reduzindo a quantidade de níveis.

4.1 Reagrupamento de Níveis

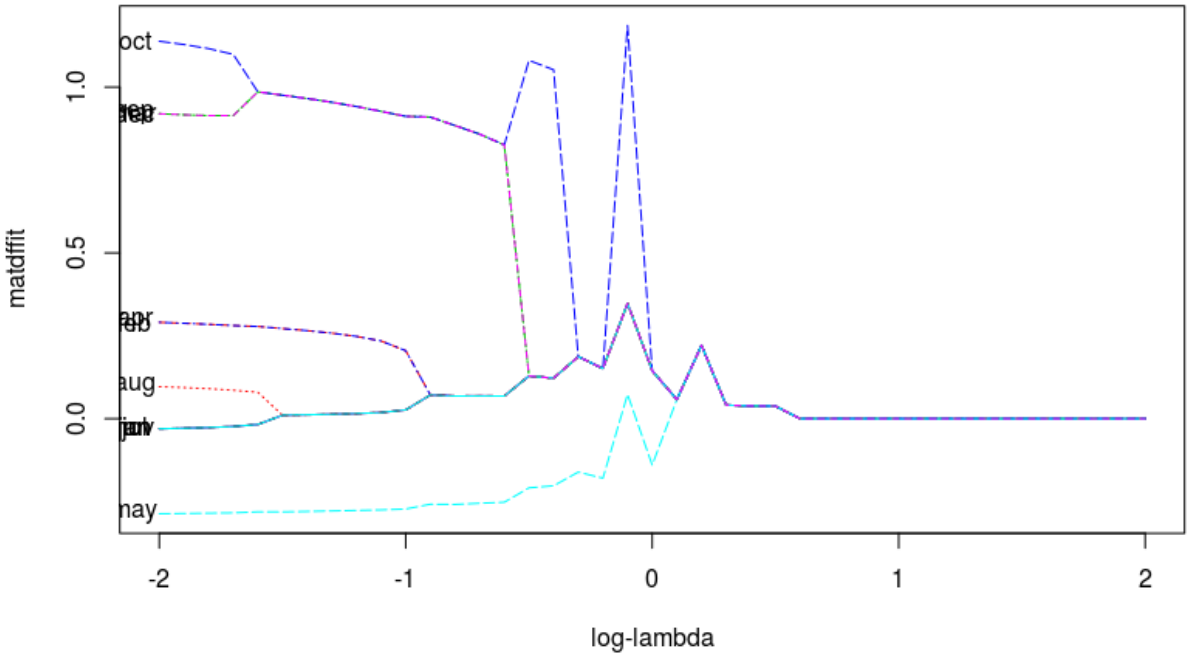

Figura 1: Gráfico de Reagrupamento de Níveis da variável *month*

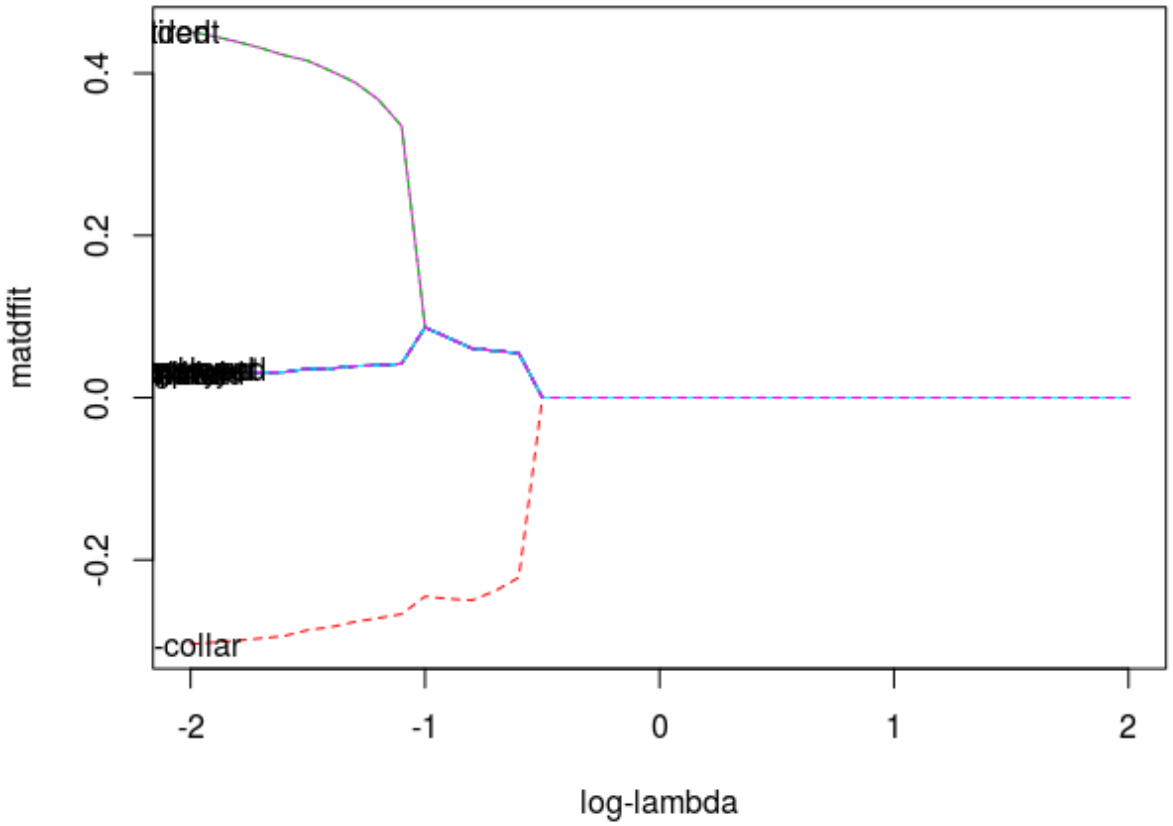

Figura 2: Gráfico de Reagrupamento de Níveis da variável *job*

É possível verificar que conforme o $\boldsymbol{\lambda}$ existe uma semelhança entre os níveis das variáveis *month*(Figura 1) e *job*(Figura 2).

O algorítmo de agrupamento retorna um coeficiente para cada nível da variável e coeficientes iguais indicam os niveis em cada novo agrupamento.

Tabela 2: Coeficientes Para Reagrupamentos dos Níveis da Variável *month*.

Niveis	Coeficientes
may	-0.28
jan	-0.03
jun	-0.03
jul	-0.03
nov	-0.03
aug	0.09
feb	0.29
apr	0.29
mar	0.91
sep	0.91
dec	0.91
oct	1.12

A variável *month* ficou agrupada em 5 novos níveis com $\boldsymbol{\lambda} = 0.165$(Tabela 2):

1. *may*;
2. *jan*, *jun*, *jul* e *nov*;
3. *aug*;
4. *feb* e *apr*;
5. *mar*, *sep* e *dec*.

Tabela 3: Coeficientes Para Reagrupamentos dos Níveis da Variável *job*.

Niveis	Coeficientes
blue-collar	-0.31
entrepreneur	-0.06
services	-0.06
technician	-0.06
admin.	0.10
housemaid	0.10
management	0.10
self-employed	0.10
unemployed	0.10
unknown	0.10
retired	0.44
student	0.56

E para a variável *job*, o novo agrupamento ficou dividido em 5 novos níveis com $\boldsymbol{\lambda} = 0.086$(Tabela 3).

1. *blue-collar*;
2. *entrepreneur*, *sevices* e *tehnician*;
3. *admin*, *housemaid*, *management*, *self-employed*, *unemployed* e *unknow*;
4. *retired*;
5. *student*.

4.2 Ajuste de Modelos com Níveis Reagrupados

Tabela 4: Comparação Entre o Modelo Selecionado(*) e os com Níveis Reagrupados.

Função de Ligação	AIC	N° de Parâmetros
Probito *	1821.418	34
Probito	1863.186	17
Logito	1872.035	21
Cauchit	1985.838	20
Complemento log-log	2012.095	18

Ao analisar os modelos com níveis reagrupados em relação ao selecionado anteriormente(Tabela 4), é notório que o menor (AIC) ainda é o do primeiro modelo escolhido, mas em virtude da redução em 17 parâmetros o modelo selecionado para continuar a análise foi o **Binomial** com funçãp de ligação **probito** e AIC = 1863.186, pois um acréscimo em 41.768 em AIC é razoável por uma simplicidade maior no modelo.

Tabela 5: Coeficientes do Modelo Ajustado e Seus Erros Padrões.

	Estimativa	Erro Padrão
(Intercept)	-1.895	0.200
duration	0.002	0.000
monthmonth2	0.156	0.101
monthmonth3	0.446	0.124
monthmonth4	0.402	0.119
monthmonth5	1.096	0.169
contacttelephone	0.054	0.131
contactunknown	-0.424	0.107
jobjob2	0.166	0.209
jobjob3	0.338	0.108
jobjob4	0.764	0.167
jobjob5	0.339	0.104
pdays	0.001	0.000
housingyes	-0.290	0.076
campaign	-0.051	0.016
loanyes	-0.339	0.107
age	-0.007	0.004

Modelo ajustado(Tabela 5) e sua equação.

$\mathbf{\Phi}^{-1}(\pi_i) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_8 + \hat{\beta}_9 x_9 + \hat{\beta}_{10} x_{10} + \hat{\beta}_{11} x_{11} + \hat{\beta}_{12} x_{12} + \hat{\beta}_{13} x_{13} + \hat{\beta}_{14} x_{14} + \hat{\beta}_{15} x_{15} + \hat{\beta}_{16} x_{16} + \hat{\beta}_{17} x_{17}$

4.3 Gráficos de Efeito do Preditor

Como o modelo selecionado temo função de ligação **probito**, uma forma de interpretar os parâmetros é através de gráficos de efeito(Figura 3 e Figura 4).

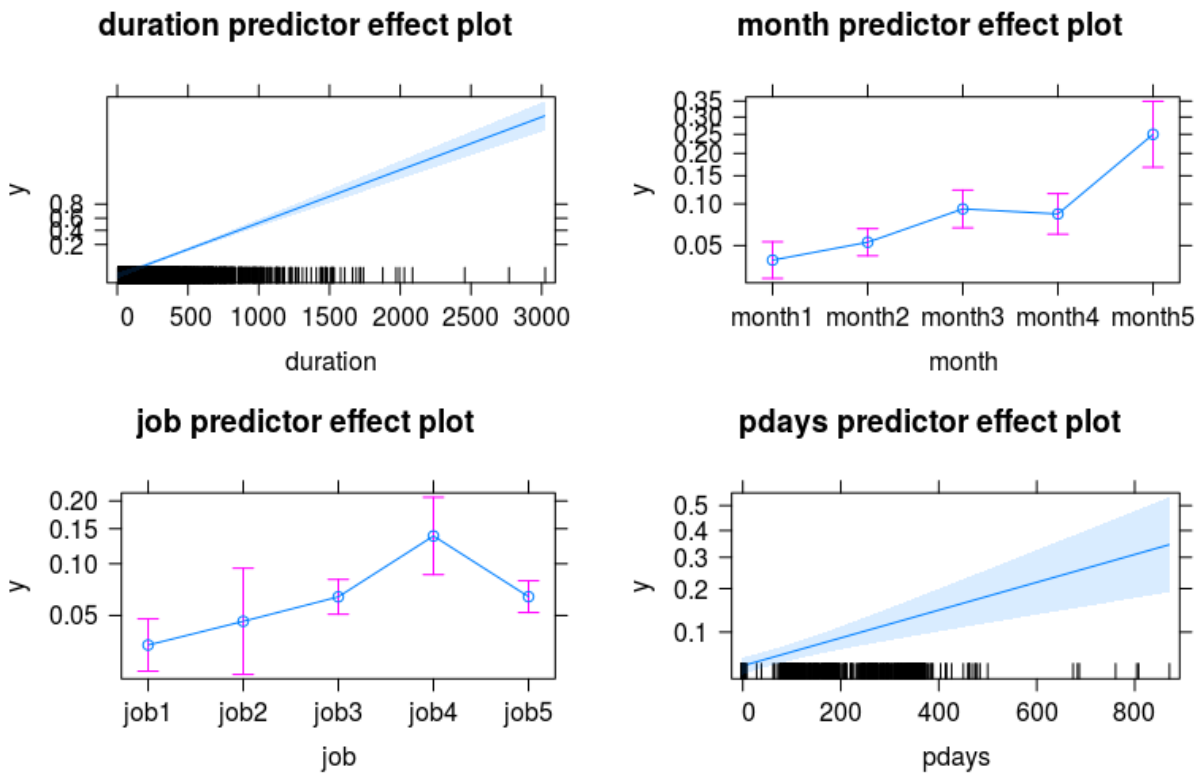

Figura 3: Gráficos de Efeito para *duration*, *month*, *job* e *pdays*

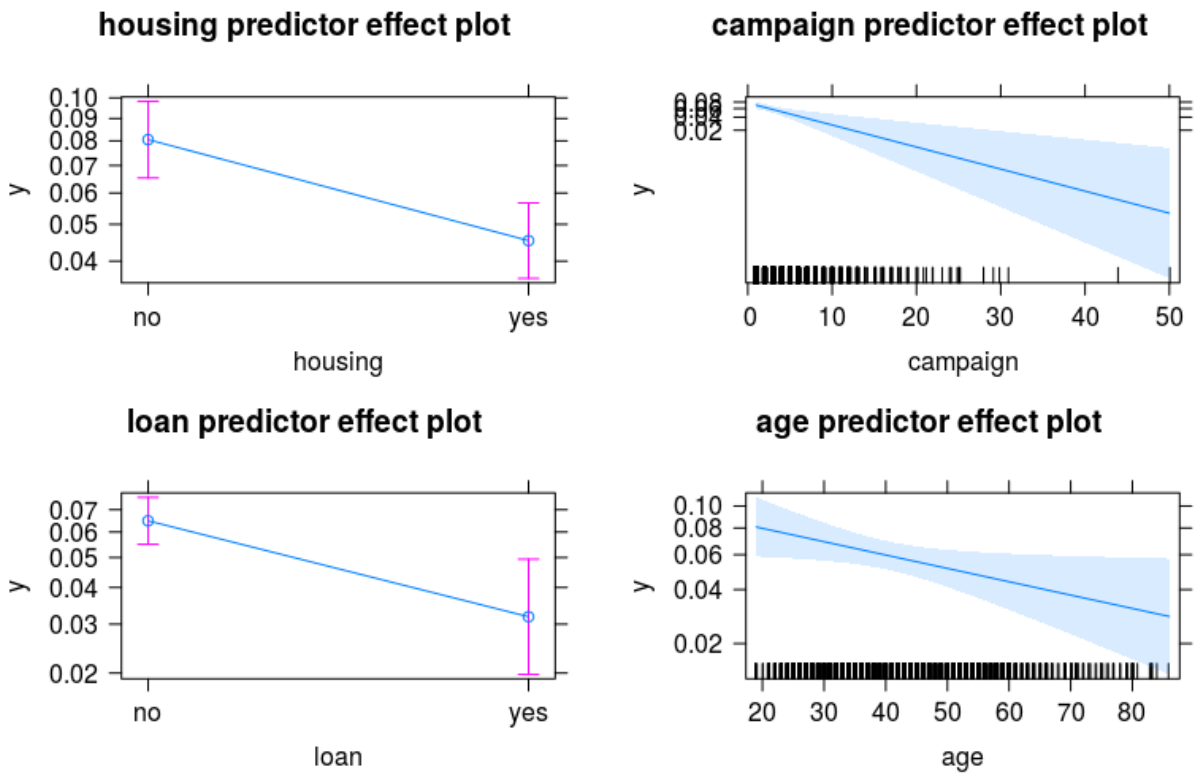

Figura 4: Gráficos de Efeito para *housing*, *campaign*, *loan* e *age*

4.4 Predição

A comparação entre as predições do modelo e os valores observados é feita através de tabelas de classificação, onde as prediçoes são baseadas em pontos de corte.

Os pontos de corte escolhidos para o estudo foram o tradicional **0.5**(Tabela 6), **0.89**(Tabela 7) baseado na proporção de sucessos observados e **0.3**(Tabela 8) como um valor mais baixo para comparação.

Tabela 6: Tabela de classificação com ponto de corte **0.5**.

	no	yes
no	17	18
yes	772	97

Tabela 7: Tabela de classificação com ponto de corte **0.89**.

	no	yes
no	4	2
yes	785	113

Tabela 8: Tabela de classificação com ponto de corte **0.3**.

	no	yes
no	52	46
yes	737	69

Pelas tabelas de classificação é possível perceber que, de modo geral, o modelo gera muitas prediçoes de *yes*, quando na realidade são *no*. Isso fica ainda mais claro na tabela comparando sensibilidade e especificidade para cada ponto de corte(Tabela 9). Evidenciando uma alta especificidade e uma baixíssima sensibilidade do modelo.

Tabela 9: Tabela de classificação com ponto de corte **0.3**.

	Sensibilidade	Especificidade
pc=0,89	0.0050697	0.9826087
pc=0,5	0.0215463	0.8434783
pc=0,3	0.0659062	0.6000000

4.5 Curva ROC

Usamos a curva ROC para comparar especificidade e sensibilidade ao longo dos pontos de corte(Figura 5).

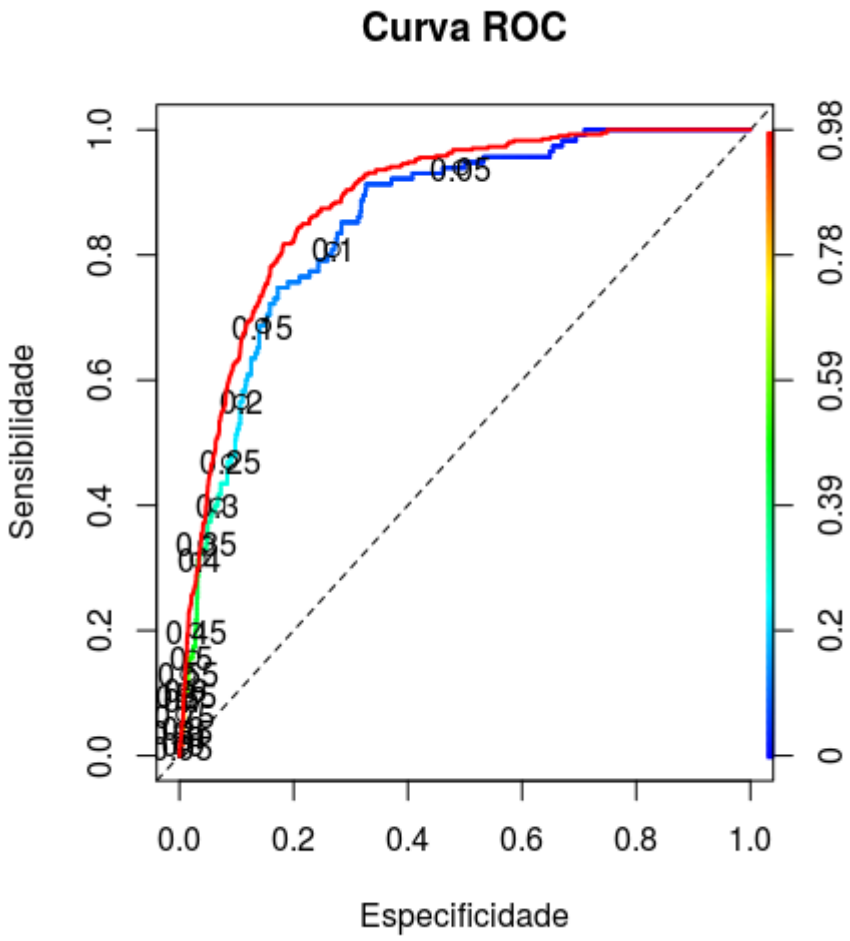

Figura 5:

4.6 Regra de Decisão Incorporando Custos

Vamos supor que o custo de classificar um falço não seja 1.5 vezes o de classificar um falso sim (C(S|N) = 1.5 * C(N|S)). Assim, para uma regra de classificação com probabilidades de classificação incorretas P(S|N) e P(N|S), o custo esperado de má classificação (ECMC) fica dado por:

ECMC = C(S|N) * P(S|N) * P(N) + C(N|S) * P(N|S) * P(S)

Como existe 11% de respostas sim na base, P(S) = 0,11; P(N) = 0,89.

Assim, ECMC = 0.89 * P(S|N) + 0,165 * P(N|S)

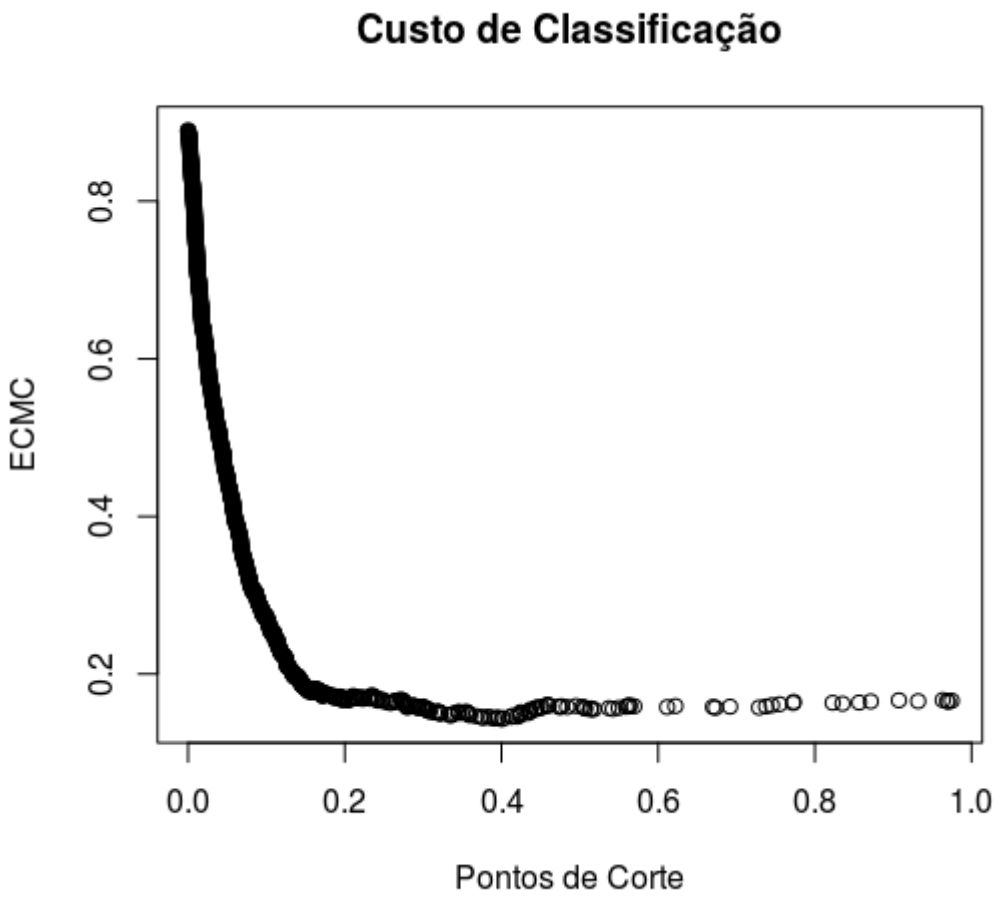

Figura 6: Gráfico de Custo de Classificação

O gráfico mostra que um ponto de corte em torno de 0.4 gera o menor custo de classificação(Figura 6).