Análise estatística de dados com objetivo de identificar clientes propensos a cancelar produtos de um banco

Daniel V. F. Falbel

16 de Novembro de 2014

Resumo

Para auxiliar um banco na criação de ações publicitárias para retenção de clientes que possuem um certo cartão de crédito, este trabalho apresenta uma análise estatística que permite a identificação dos clientes mais propensos a cancelar o produto. Através de uma análise exploratória dos dados, foi identificado que a evolução dos saldos do cliente (credor, devedor ou da poupança) além do seu salário eram variáveis importantes para a explicação do tempo até o cancelamento do cartão de crédito.

1 Descrição do problema

Um banco deseja fazer ações de marketing para reter os seus clientes, evitar que eles cancelem seus produtos. Para poder fazer ações mais acertivas, o banco deseja saber qual é o perfil dos clientes com maior propensão a cancelar um certo cartão de crédito.

As variáveis que podem ajudar na identificação dos perfis estão listadas abaixo:

- sexo: M-masculino; F-feminino
- modulo: Segmentação de clientes; valores mais baixos representam clientes com menor renda ou investimento, valores mais altos representam clientes mais interessantes para a instituição
- cheque: Classificação da conta corrente; quanto maior o valor, mais "especial" o cliente
- evolcredor: Evolução do saldo credor médio trimestral (A: aumentou, D: diminuiu, M: manteve)
- evoldevedor: Evolução do saldo devedor médio trimestral (A: aumentou, D: diminuiu, M: manteve)
- evolpoup: do saldo da poupança trimestral (A: aumentou, D: diminuiu, M: manteve)
- idade: idade em anos
- salario: está categorizado em 10 categorias (quanto maior, maior o salário)
- cartaocancel: Cancelamento do cartão de credito pelo banco 0 - não cancelou; 1 - cancelou
- bancsal: 0: não recebe salário pelo banco; 1: recebe salário pelo banco
- tempo: Tempo de permanência com um determinado produto, em meses
- status: 1: tempo refere-se ao tempo da contratação até o cancelamento do produto; 0: tempo da contratação até término de acompanhamento sem cancelamento (censura)

2 Análise Descritiva

Para verificar se existem possíveis inconsistências nos dados, foi feita uma análise exploratória. Além disso, podemos já ter uma ideia do que pode influencia no tempo até o cancelamento do cartão de crédito.

Pelos gráficos da figura 1, podemos observar que 60% dos clientes são do sexo masculino. A maior parte dos clientes teve um aumento do saldo credor e do saldo devedor. Também podemos analisar que a maior parte dos clientes manteve estável o saldo da poupança. Observamos que a maior parte dos clienes estão no segmento 110 do

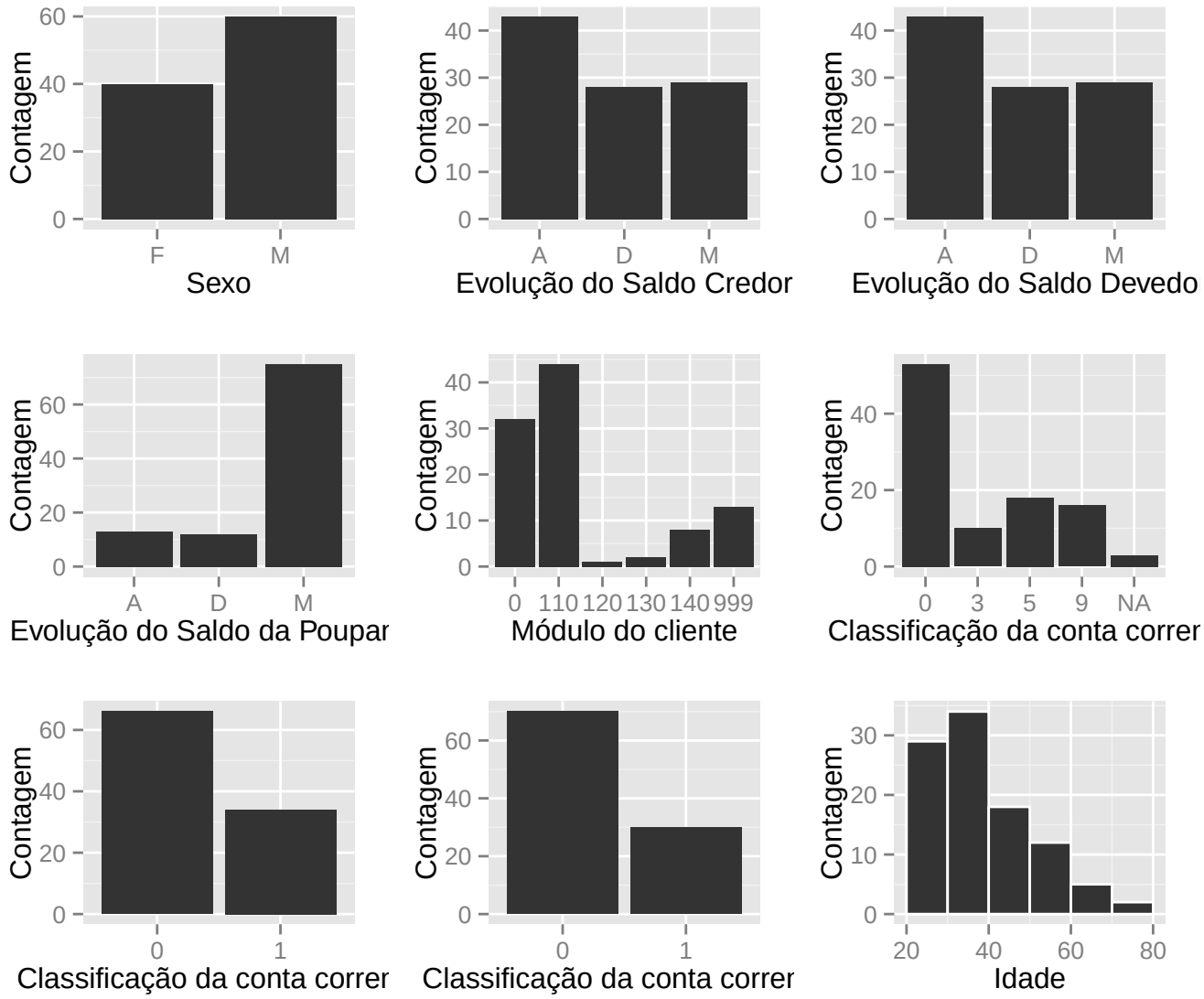

Figure 1: (i) quantidade de clientes em cada categoria da variável sexo, (ii) quantidade de clientes em cada categoria de evolução do saldo credor, (iii) quantidade de clientes em cada categoria de evolução do saldo devedor, (iv) quantidade de clientes em cada categoria de evolução do saldo da poupança, (v) quantidade de clientes em cada módulo (segmento) criado pelo banco, (vi) quantidade de clientes em cada classificação da conta corrente, (vii) quantidade de clientes que cancelaram o cartão (1), (viii) quantidade de clientes que recebem o salário por meio do banco, (ix) histograma da idade dos clientes

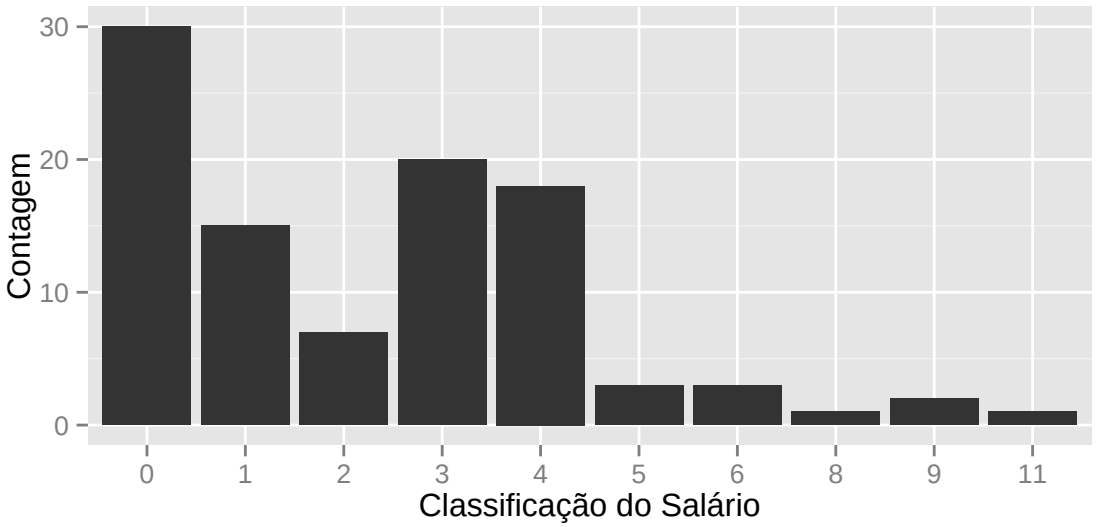

Figure 2: Quantidade de clientes em cada classificação do salário

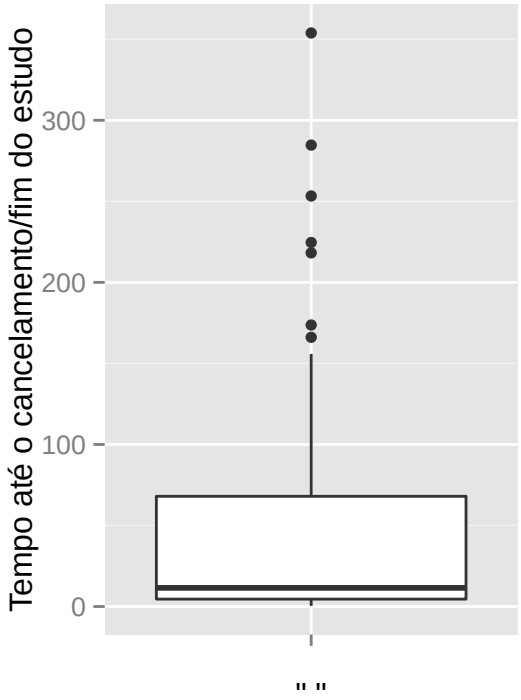

Figure 3: Boxplot do tempo até o cancelamento do cartão de crédito ou fim do acompanhamento

módulo, e que cerca de 50% dos cliente estão na classificação 0 da conta corrente. Note que no gráfico (vi) existe a ocorrência de uma observação ausente, isto é, não temos a informação de que tipo de conta o cliente ppossui.

Pela figura 2 observamos que cerca de 30% dos clientes ganham até R$300,00, em seguida as categorias com mais indivíduos são a 3 e a 4 que contém pessoas que ganham de R$501,00 a R$1500,00.

Na figura 3, vemos que os clientes em mediana têm o produto por 11 meses. Aproximadamente 75% dos clientes tem tempo até 68 meses. O indivíduo que possui o produto a mais tempo, o tem a 353 meses. No estudo observamos que 18 clientes foram censurados, ou seja, ainda não tinham cancelado o cartão de crédito até a data de fim do acompanhamento.

Como não encontramos nenhuma irregularidade nas variáveis, passamos para uma análise descritiva bivariada. Aqui cruzamos as variáveis do banco de dados com o tempo até o cancelamento do produto, para assim termos

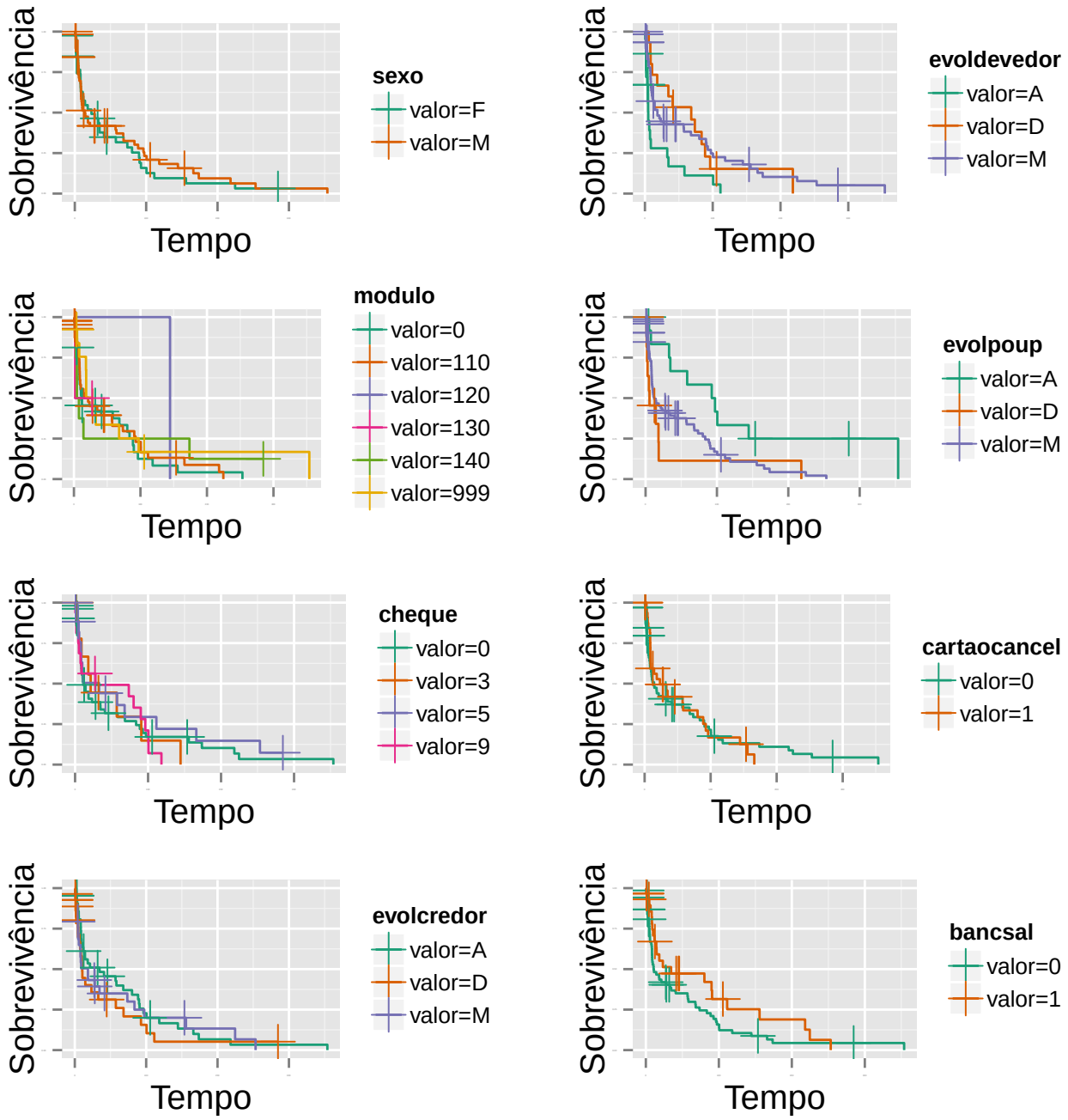

Figure 4: Curvas de sobrevivência estimadas por Kaplan-Meier para cada variável do banco de dados

dicas de como as variáveis estão relacionadas ao tempo que o cliente permanece com o cartão.

Pelos gráficos da figura 5 podemos ver que não parece existir associação entre o sexo e o tempo, já que para as duas categorias as linhas se sobrepoem. O módulo também não parece influenciar, mas o tipo de conta e as evoluções de saldo devedor, creditor ou da poupança parecem apresentar diferenças no tempo até o cancelamento dependendo das categorias. Os clientes que recebem o salário pelo banco tendem, aparentemente a ter tempos maiores.

Nos gráficos da figura 6, é possível observar que o salário do cliente parece ser correlacionado positivamente com o tempo atée o cancelamento do cartão e que a idade não apresenta esse comportamento.

Após a análise descritiva, acreditamos que as principais variáveis para prever o tempo até o cancelamento do cartão são os saldos (credor, devedor e de poupança) e a categoria de salário do cliente.

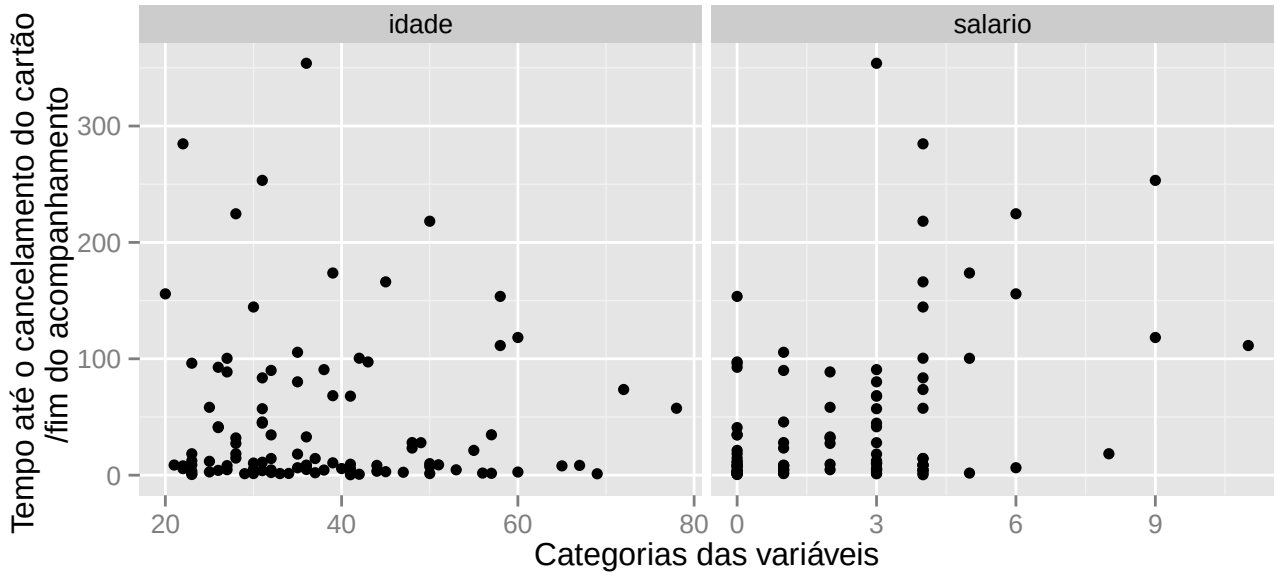

Figure 5: Gráficos de dispersão do tempo até o cancelamento do cartão/fim do acompanhamento pela idade e pela categoria de salário do cliente.

3 Análise Inferencial

Inicialmente vamos propor um modelo de Cox da forma:

$$\alpha(t|X) = \alpha_0(T)exp(t\beta)$$

Com X um vetor com as covariáveis apresentadas na análise descritiva e β os parâmetros associados a essas variáveis. O ajuste do modelo foi feito no `R` usando o comando `coxph` do pacote `survival`. Já a seleção das variáveis explicativas foi feita usando o método 'AIC', isto é, selecionamos o modelo com o menor AIC. Como apenas três observações possuiam valores omisso em algumas das variáves, optamos por apenas excluí-las da análise. Optamos por não utilizar devido as duas variáveis "saldo credor" e "saldo devedor" juntas no modelo, pois por terem forte associação, a estimação dos parâmetros pode ser prejudicada. Por isso utilizamos apenas a variável evolcredor.

```
library(survival)
modelo <- coxph(Surv(tempo,status) ~ sexo + modulo + as.factor(cheque) + evolcredor +
                  evolpoup + idade + salario + cartaocancel + bancsal,
                data = na.omit(dados))


modelo.a <- step(modelo)
```

As variáveis explicativas selecionadas foram a evolução do saldo credor e da poupança além do salário e do indicador se o cliente recebe o salário pelo banco. Todas as variáveis explicativas selecionadas pelo método AIC são significativas a nível de confiança 5% pelo teste de Wald marginal.

Com o modelo escolhido, fizemos uma análise de diagnóstico para verificar a qualidade do ajuste. Analisamos o resíduo de Cox-Snell, que da uma ideia da qualidade geral do ajuste do modelo.

O gráfico dos resíduos de Cox-Snell indicou que o ajuste geral do modelo parece ser adequado, já que os pontos estao próximos a reta referência. Não tendo problema com o ajuste geral do modelo, fizemos o gráfico do preditor linear pelo resíduo deviance, para assim poder identificar observações aberrantes que poderiam interferir nas estimativas do parâmetro.

A partir da observação do gráfico podemos concluir que não existem pontos aberrantes. Além disso temos mais um indicativo de que o modelo de Cox está bem ajustado já que os valores dos resíduos estão aleatoriamente distribuidos de acordo com o preditor linear. Não observamos nenhum padrão, o que indicaria que o modelo não está bem ajustado.

A seguir apresentamos as estimativas dos parâmetros do modelo e discutiremos a interpretação dos resultados.

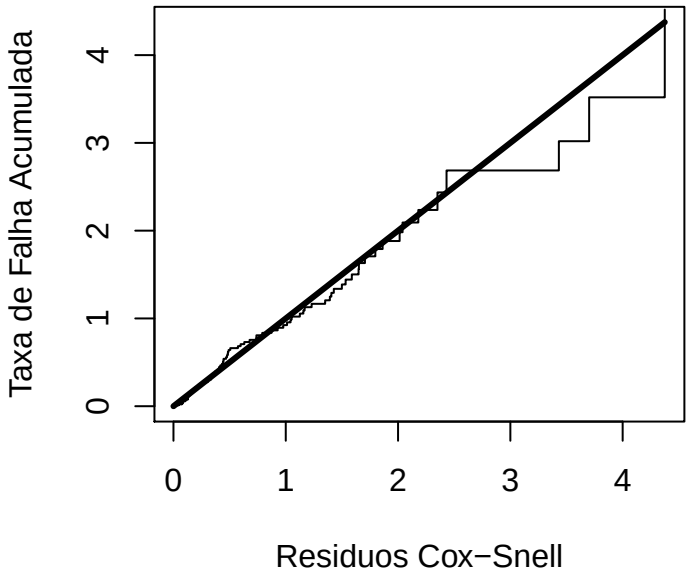

Figure 6: Gráfico da taxa de falha acumulada pelos resíduos de Cox-Snell, era esperado os pontos estivessem sobre/próximos a reta referência

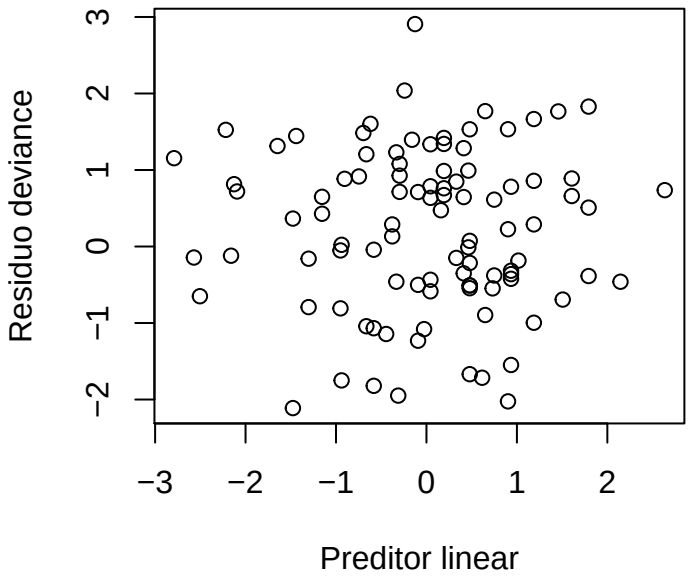

Figure 7: Gráfico do preditor linear pelo resíduo deviance. Esperamos que eles estejam distribuídos aleatoriamente e que não haja resíduos com valores em módulo maiores do que 2.

Para facilitar a interpretação utilizaremos a tabela abaixo que apresenta as taxas de falha relativa estimadas bem como seus intervalos de confiança.

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
evolcredorD	1.31	3.72	0.33	4.04	0.00
evolcredorM	0.71	2.03	0.29	2.43	0.02
evolpoupD	3.25	25.88	0.60	5.38	0.00
evolpoupM	2.13	8.38	0.44	4.87	0.00
salario	-0.29	0.75	0.06	-4.48	0.00
bancsal	-0.78	0.46	0.28	-2.73	0.01

Table 1: Estimativas dos parâmetros do modelo ajustado

	exp(coef)	exp(-coef)	lower .95	upper .95
evolcredorD	3.72	0.27	1.97	7.04
evolcredorM	2.03	0.49	1.15	3.60
evolpoupD	25.88	0.04	7.92	84.63
evolpoupM	8.38	0.12	3.56	19.69
salario	0.75	1.33	0.66	0.85
bancsal	0.46	2.17	0.26	0.80

Table 2: Estimativas das razões de chance com base nos parâmetros do modelo ajustado

Utilizando as estimativas do modelo, podemos concluir que o fator mais importante para o cancelamento do cartão é a evolução do saldo da poupança. O cliente que diminui o saldo da poupança tem 24 vezes mais risco de cancelar do que um cliente como na referência (que aumentou o saldo credor, o saldo de poupança, não recebe o salário pelo banco e tem um salário baixo). O risco de cancelar quando o cliente manteve o saldo da poupança é 8 vezes o da referência. O cliente que teve aumento no saldo devedor tem 3 vezes o risco de cancelar o cartão se comparado com a referência, já o que manteve o saldo tem 2 vezes esse risco.

Um fato interessante é o que o cliente que recebe o salário pelo banco tem 2 vezes menos risco de cancelar o cartão quando comparado a referência. Quanto maior o salário, menor o risco de cancelamento também, a cada aproximadamente 500 reais o risco diminui cerca de 30%.

References

[1] Enrico Antônio Colosimo e Suely Ruiz Giolo. *Análise de Sobrevivência Aplicada*. 2006.