Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Classificação de Bons e Baus pagadores RegreAnálise de Risco

Douglas de Paula Nestlehner

Setembro, 2022

Capítulo 1

Problema Apresentado

Em corporações financeiras, é de extremo interesse identificar o perfil de seus clientes, ou possíveis clientes, no intuito de predizer bons ou maus pagadores. Permitindo assim, traçar estratégias e tratamentos específicos para cada grupo.

Uma das técnicas mais utilizadas em situações de classificações de clientes (bons ou maus pagadores por exemplo), é a regressão logística, que consiste em descrever os dados e explicar a relação entre uma variável binária dependente e uma ou mais variáveis independentes.

Teremos como objetivo nesse trabalho, ajustar um modelo de regressão logística no intuito de classificar maus pagadores, para a base de dados disponibilizada, a qual contem informações de cliente de um determinado banco.

1.1 Dados

A base disponibilizada contem informações de 1000 clientes de um determinado banco, sendo essas informações apresentadas por meio de 20 covariáveis, e a variável resposta.

- **V1**: Variável indicativa de "Status de existência de conta corrente" ($A11 : ... < 0$ unidades monetárias; $A12 : 0 \leq ... < 200$ unidades monetárias; $A13 : ... \geq 200$ unidades monetárias/atribuições salariais por pelo menos 1 ano; $A14$: sem conta corrente);

- **V2**: Variável numérica da duração em meses da conta corrente;

- **V3**: Variável indicativa do "Histórico de crédito" (A30: nenhum crédito recebido/todos os créditos pagos devidamente; A31: todos os créditos neste banco foram devidamente pagos; A32: créditos existentes pagos devidamente até agora; A33: atraso no pagamento no passado; A34: conta crítica/outros créditos existentes;

- **V4**: Variável indicativa do "Propósito" (A40: carro (novo); A41 : carro (usado); A42: móveis/equipamentos; A43: rádio/televisão; A44: eletrodomésticos; A45 : reparos; A46: educação; A47: férias; A48: reciclagem; A49: negócios; A410: outros);

- **V5**: Variável numérica da quantidade de crédito da conta corrente;
- **V6**: Variável indicativa da "Conta poupança/títulos" (A61: ... ¡ 100 unidades monetárias; $A62 : 100 \leq ... < 500$ unidades monetárias; $A63 : 500 \leq ... < 1000$ unidades monetárias; $A64 : .. \geq 1000$ unidades monetárias; A65 : desconhecido;
- **V7**: Variável indicativa do "Emprego atual desde" (A71: desempregado; $A72 : ... <$ 1 ano; $A73 : 1 \leq ... < 4$ anos; $A74 : 4 \leq ... < 7$ anos; $A75 : .. \geq 7$ anos);
- **V8**: Variável continua "Taxa de prestação em percentagem do rendimento disponível";
- **V9**: Variável indicativa do "Status pessoal e sexo" (A91: masculino, divorciado/separado; A92: mulher, divorciada/separada/casada; A93: masculino, solteiro; A94: masculino, casado/viúvo; A95: feminino, solteiro);
- **V10**: Variável indicativa de "Outros devedores/fiadores" (A101: nenhum; A102: co-requerente; A103: fiador);
- **V11**: Variável numérica indicativa tempo em meses na residencia atual;
- **V12**: Variável indicativa de "Propriedade" (A121: imóveis; A122: se não A121: acordo de poupança da sociedade de construção/seguro de vida; A123: se não A121/A122 : carro ou outro, não na Variável 6; A124: desconhecido/sem propriedade);
- **V13**: Variável numérica indicativa da idade em anos;
- **V14**: Variável indicativa de "Outros planos de parcelamento" (A141: banco; A142: lojas; A143: nenhum);
- **V15**: Variável indicativa de "Residência" (A151: aluguel; A152: própria; A153: de graça);
- **V16**: Variável numérica indicativa do numero de crédito disponivel no banco;
- **V17**: Variável indicativa de "Trabalho" (A171: desempregado/não qualificado - não residente; A172: não qualificado – residente; A173: funcionário/funcionário qualificado; A174 : gestão/autônomo/funcionário/diretor altamente qualificado);
- **V18**: Variável numérica indicativa do numero de pessoas responsáveis pela manutenção;
- **V19**: Variável indicativa de "Telefone" (A191: nenhum; A192: sim, registrado em nome do cliente);
- **V20**: Variável indicativa de "Trabalhador estrangeiro" A201: sim; A202: não).
- **Y**: Variável resposta, indicativa se o cliente é 0: Bom pagador ou 1: Mau pagador.

Capítulo 2

Resultados

Nessa seção iremos apresentar todos os procedimentos realizados no intuito de se obter um modelo de classificação, para os dados apresentados.

2.1 Análise exploratória

Para se ter mais conhecimento das variáveis presentes na base, calculamos algumas medidas descritivas (min., max., média, mediana e quartis, para as variáveis continuas, e os fatores e suas respectivas frequências, para as variáveis categóricas). Na figura 2.1 temos representadas essas medidas.

```
  V1          V2          V3          V4          V5          V6          V7
A11:274   Min.   : 4.0   A30: 40   A43   :280   Min.   :  250   A61:603   A71: 62
A12:269   1st Qu.:12.0   A31: 49   A40   :234   1st Qu.: 1366   A62:103   A72:172
A13: 63   Median :18.0   A32:530   A42   :181   Median : 2320   A63: 63   A73:339
A14:394   Mean   :20.9   A33: 88   A41   :103   Mean   : 3271   A64: 48   A74:174
          3rd Qu.:24.0   A34:293   A49   : 97   3rd Qu.: 3972   A65:183   A75:253
          Max.   :72.0             A46   : 50   Max.   :18424
                                   (Other): 55
      V8          V9          V10          V11          V12          V13          V14
Min.   :1.000   A91: 50   A101:907   Min.   :1.000   A121:282   Min.   :19.00   A141:139
1st Qu.:2.000   A92:310   A102: 41   1st Qu.:2.000   A122:232   1st Qu.:27.00   A142: 47
Median :3.000   A93:548   A103: 52   Median :3.000   A123:332   Median :33.00   A143:814
Mean   :2.973   A94: 92              Mean   :2.845   A124:154   Mean   :35.55
3rd Qu.:4.000                        3rd Qu.:4.000              3rd Qu.:42.00
Max.   :4.000                        Max.   :4.000              Max.   :75.00

  V15          V16          V17          V18          V19          V20                Y
A151:179   Min.   :1.000   A171: 22   Min.   :1.000   A191:596   A201:963   Bom pagador:700
A152:713   1st Qu.:1.000   A172:200   1st Qu.:1.000   A192:404   A202: 37   Mau pagador:300
A153:108   Median :1.000   A173:630   Median :1.000
           Mean   :1.407   A174:148   Mean   :1.155
           3rd Qu.:2.000              3rd Qu.:1.000
           Max.   :4.000              Max.   :2.000
```

Figura 2.1: Medidas descritivas.

Para tornar a visualização dessas informações mais fáceis, construímos gráficos de proporções para as variáveis categóricas, e histogramas para as variáveis continuas, permitindo ter uma visão mais clara sobre cada variável da base. Temos representado na figura 2.2 os gráficos.

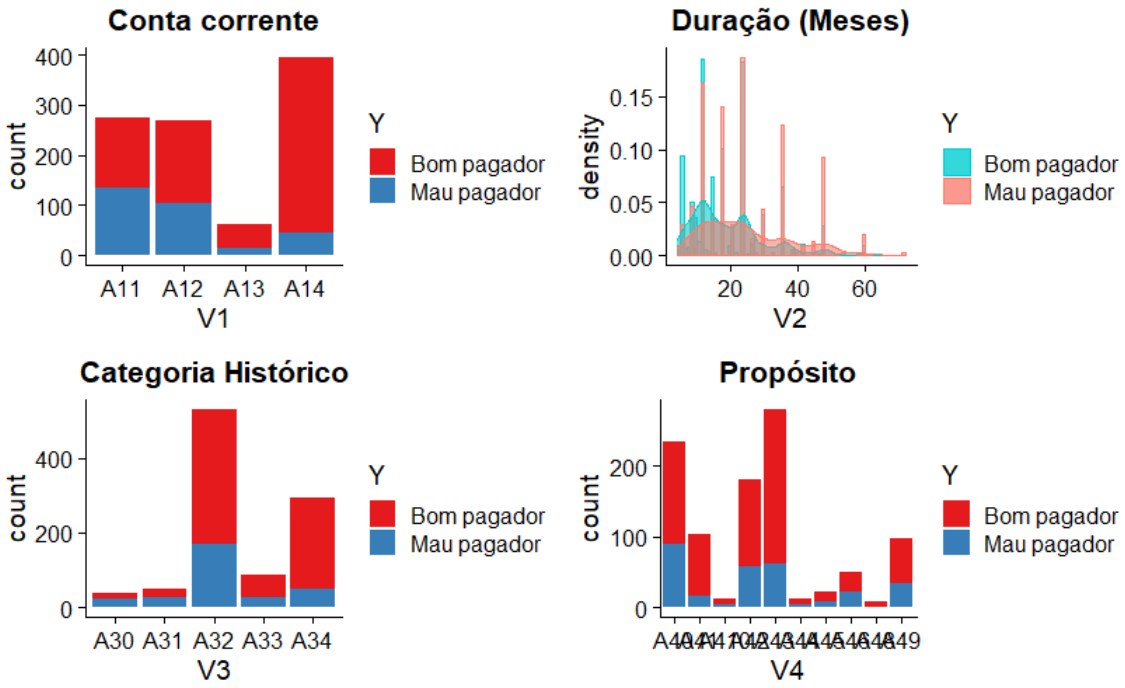

Figura 2.2: Descritiva quatro primeiras variáveis.

Observando as 4 primeiras variáveis, podemos notar características que apresentam maior frequência em bons e maus pagadores. Exemplo: na variável V1, apesar de maior frequência de conta correte "A14" existe uma maior frequência de maus pagadores quando a conta é "A11" e "A12". Na variável continua V2, notamos que a maioria de bons pagadores estão relacionados a um tempo de duração pequeno, enquanto os maus pagadores tendem a ter uma duração maior. A seguir temos os gráficos para as demais variáveis.

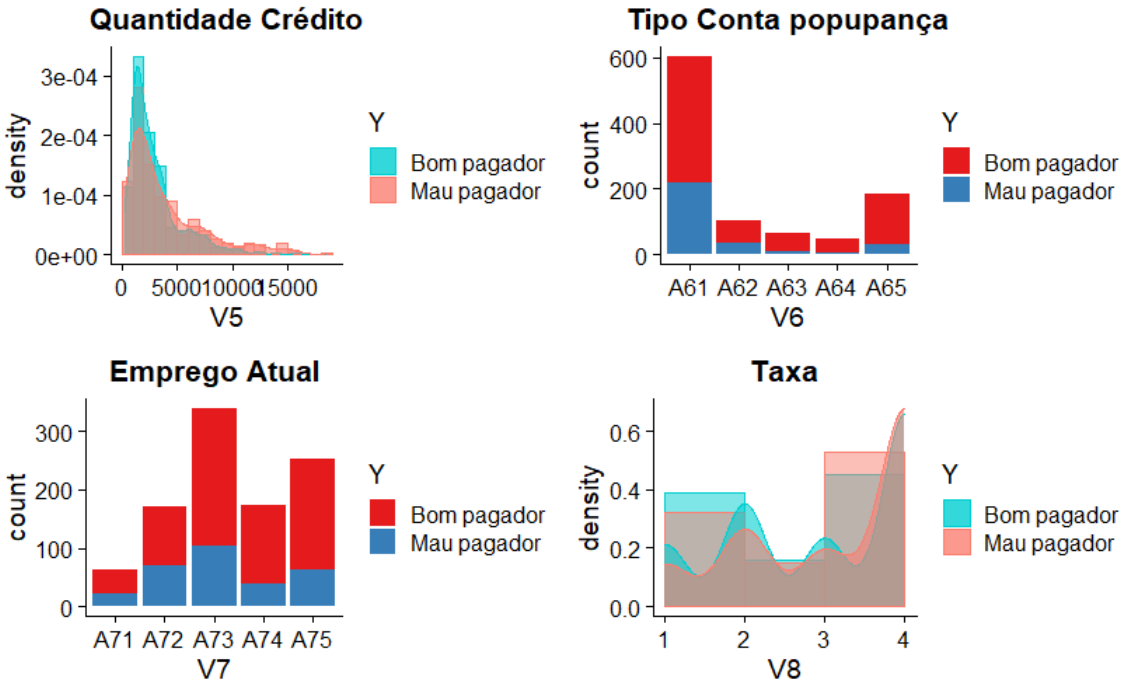

Figura 2.3: Descritiva V5 - V8

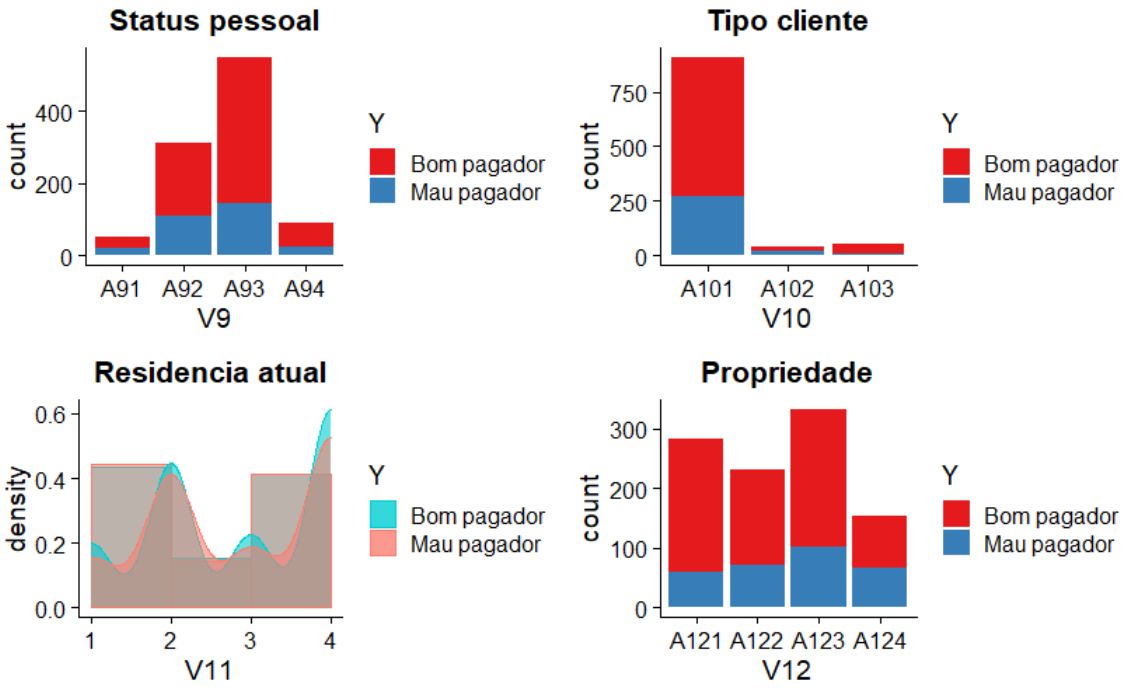

Figura 2.4: Descritiva V9 - V12

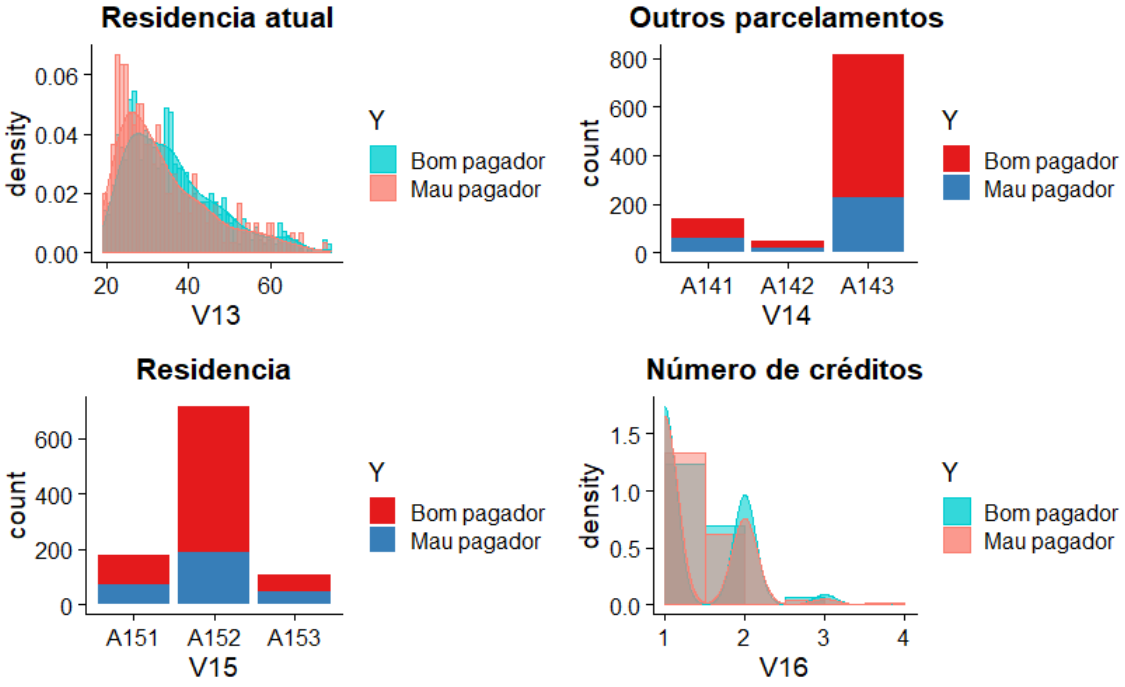

Figura 2.5: Descritiva V13 - V16

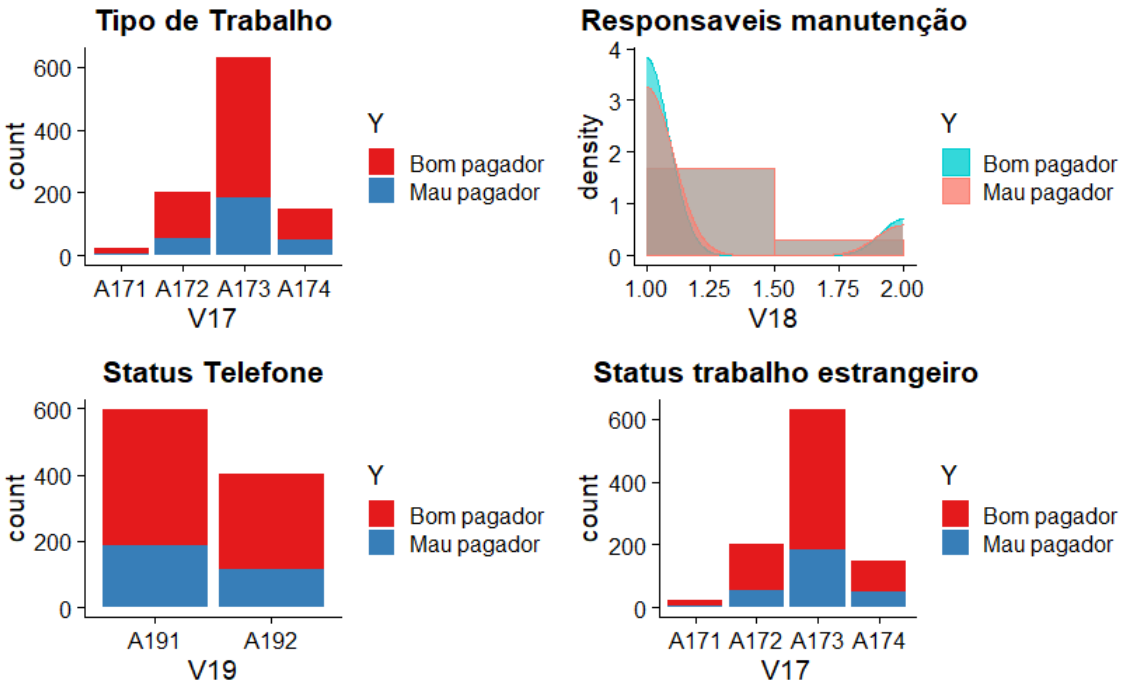

Figura 2.6: Descritiva V17 - V20

De modo geral, observamos que todas as variáveis possuem um certo comportamento (por menor que seja) que difere bons e maus pagadores. Assim sendo, o uso das mesmas, para a modelagem de classificação (iremos usar regressão logística) permitira trazer bons resultados.

Apenas no intuito de observarmos a frequência da variável resposta, construímos a figura **??**.

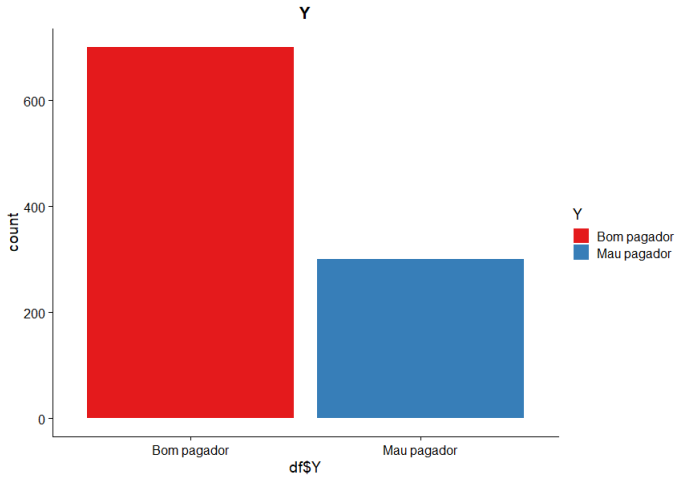

Figura 2.7: Descritiva Y.

2.2 Ajuste do Modelo

Em modelos de regressão logística, podemos nos deparar com situações em que a variável resposta é binaria, ou categórica ordenada, ou categórica desordenada (não existe uma hierarquia). Para o caso em estudo, temos o caso de regressão logística binaria, em que a variável resposta binaria (1: bom pagador, 2:mau pagador).

Para realizar a previsão da variável resposta (no caso binario), teremos que valores com probabilidade acima de 0.50 sejam classificados ao grupo de interesse, e caso contrario no outro grupo. Para isso, realizamos a estimação dos coeficientes das varaiveis independentes, pelo logit ou razão de desiguladades, dados por:

$$Logit_i = log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * V1 + \beta_{2V}2 + ... + \beta_{20} * V20$$

Em que p é a probabilidade do evento. Lembrando que a descrição do modelo acima é apenas uma representação, como temos variáveis categóricas devemos representa-las com dummys (tendo muito mais β's)

Para poder treinar o modelo e verificar se ele realmente é eficiente, iremos dividir (de forma aleatória) os dados em 70% para treinamento, e 30% para validação.

Desse modo, utilizando a função "glm()" do R, declarando distribuição da variável resposta como "binomial", realizamos o primeiro ajuste considerando todas as variáveis presentes na base. Na figura 2.8 temos representado as estimativas obtidas no modelo (Modelo 1).

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.582e+00  1.442e+00   1.097 0.272698
V1A12       -4.211e-01  2.997e-01  -1.405 0.159949
V1A13       -1.107e+00  5.242e-01  -2.112 0.034700 *
V1A14       -1.937e+00  3.090e-01  -6.269 3.63e-10 ***
V2           2.494e-02  1.282e-02   1.946 0.051678 .
V3A31        3.072e-01  7.754e-01   0.396 0.691936
V3A32       -7.057e-01  6.084e-01  -1.160 0.246098
V3A33       -6.810e-01  6.661e-01  -1.022 0.306593
V3A34       -1.385e+00  6.100e-01  -2.271 0.023133 *
V4A41       -2.310e+00  5.296e-01  -4.362 1.29e-05 ***
V4A410      -1.447e+00  9.824e-01  -1.473 0.140768
V4A42       -8.799e-01  3.383e-01  -2.601 0.009294 **
V4A43       -1.068e+00  3.289e-01  -3.247 0.001167 **
V4A44       -1.773e+00  1.755e+00  -1.011 0.312185
V4A45        8.442e-01  7.123e-01   1.185 0.235962
V4A46        4.300e-03  5.047e-01   0.009 0.993203
V4A48       -1.678e+01  7.580e+02  -0.022 0.982337
V4A49       -1.012e+00  4.645e-01  -2.178 0.029425 *
V5           1.451e-04  5.843e-05   2.484 0.012994 *
V6A62       -3.811e-01  3.954e-01  -0.964 0.335061
V6A63       -1.325e-01  4.980e-01  -0.266 0.790277
V6A64       -1.563e+00  6.899e-01  -2.265 0.023507 *
V6A65       -8.788e-01  3.475e-01  -2.529 0.011437 *
V7A72        4.386e-01  5.897e-01   0.744 0.457015
V7A73        5.236e-02  5.561e-01   0.094 0.924992
V7A74       -7.431e-01  6.210e-01  -1.197 0.231476
V7A75       -1.154e-02  5.519e-01  -0.021 0.983320
V8           4.230e-01  1.187e-01   3.564 0.000366 ***
V9A92       -4.961e-01  5.235e-01  -0.948 0.343309
V9A93       -9.808e-01  5.221e-01  -1.878 0.060316 .
V9A94       -4.637e-01  6.173e-01  -0.751 0.452563
V10A102      9.655e-01  5.340e-01   1.808 0.070600 .
V10A103     -1.184e+00  5.589e-01  -2.119 0.034109 *
V11          3.506e-02  1.144e-01   0.307 0.759138
V12A122      4.003e-01  3.338e-01   1.199 0.230483
V12A123      3.838e-01  3.166e-01   1.212 0.225493
V12A124      8.531e-01  6.144e-01   1.389 0.164979
V13         -3.294e-02  1.244e-02  -2.648 0.008090 **
V14A142     -3.805e-01  5.776e-01  -0.659 0.509985
V14A143     -6.516e-01  3.369e-01  -1.934 0.053097 .
V15A152     -7.301e-01  3.156e-01  -2.314 0.020693 *
V15A153     -3.242e-01  6.941e-01  -0.467 0.640481
V16          3.339e-01  2.593e-01   1.288 0.197916
V17A172      4.613e-02  1.052e+00   0.044 0.965009
V17A173     -1.695e-01  1.021e+00  -0.166 0.868106
V17A174     -1.488e-01  1.017e+00  -0.146 0.883640
V18          1.230e-01  3.696e-01   0.333 0.739203
```

Figura 2.8: Estimativas modelo 1.

Afim de verificar quais parâmetros (efeitos fixos: variáveis) do modelo estão sendo significativos, realizamos o teste anova do tipo II, representado na Figura 2.9.

```
> anova(M1, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   652     791.07
V1    3   94.665       649     696.40 < 2.2e-16 ***
V2    1   19.330       648     677.07 1.100e-05 ***
V3    4   16.304       644     660.77  0.002637 **
V4    9   41.257       635     619.51 4.492e-06 ***
V5    1    1.403       634     618.11  0.236175
V6    4   12.479       630     605.63  0.014121 *
V7    4   15.074       626     590.56  0.004550 **
V8    1   10.504       625     580.05  0.001191 **
V9    3    6.875       622     573.18  0.075977 .
V10   2    9.214       620     563.96  0.009981 **
V11   1    0.729       619     563.23  0.393368
V12   3    4.996       616     558.24  0.172063
V13   1    7.801       615     550.44  0.005221 **
V14   2    3.119       613     547.32  0.210236
V15   2    6.169       611     541.15  0.045748 *
V16   1    1.923       610     539.22  0.165489
V17   3    0.600       607     538.62  0.896458
V18   1    0.153       606     538.47  0.695702
V19   1    0.659       605     537.81  0.416867
V20   1    7.272       604     530.54  0.007004 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Figura 2.9: Estimativas Anova modelo 1.

Nota-se que com a inclusão da variável V5 o não foi significativa, poderíamos obtar por retira-la e observar novamente o comportamento do ajuste, entretanto observamos na descaritativa que a variável V5 é uma informação importante sobre os clientes, desse modo, seguiremos com o modelo completo.

Para poder verificar se o ajuste obtido apresenta bons resultados, realizamos a predição considerando os dados de testes, e comparamos os resultados obtidos. Na figura 2.11 temos representado uma tabela de contingencia, que relaciona a classificação de bons e maus pagadores reais, e as classificações obtidas pelo modelo (para os dados de teste).

```
classe        Bom pagador Mau pagador
  Bom pagador         175          36
  Mau pagador          37          46
>
```

Figura 2.10: Classificações reais x preditas

Notamos que, dentro das 212 classificações de bons pagadores o modelo conseguiu acertar 175, e dentro das 82 classificações de maus pagadores o modelo conseguiu acertar 46. Para poder quantificar o desempenho do modelo, calculamos o risco, que totalizou 0.2482, ou seja, o modelo vai estar errando aproximadamente 24,82% das classificações.

Outro método que permite verificar o desempenho do modelo, é a contrução da curva ROC, que tem como intuito mensurar a capacidade de predição do modelo proposto, através das predições da sensibilidade e da especificidade. Na figura **??**, temos representado a curva ROC estimada para o modelo.

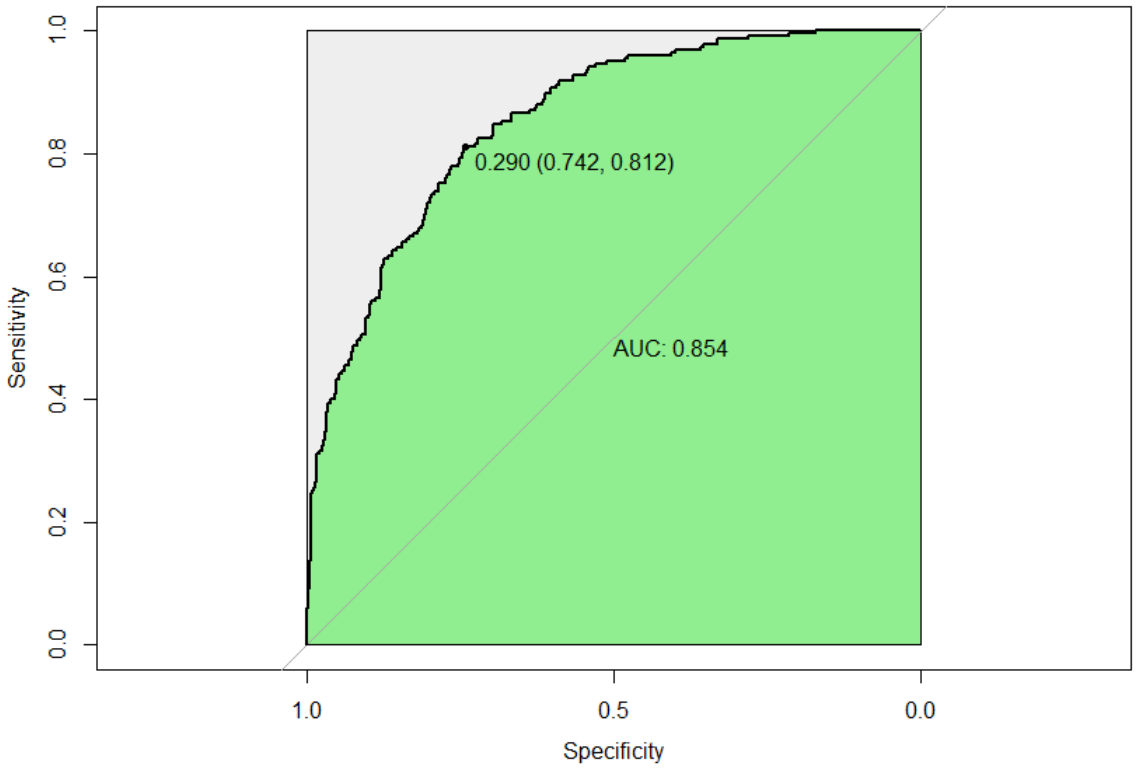

Figura 2.11: Curva ROC

De modo geral, assim como foi observado na tabela de contingencia e pelo risco estimado, a curva ROC mostra que o modelo teve um desempenho relativamente bom.

Para facilitar na interpretação dos resultados, podemos calcular a razão de chances (ODDS ratio), a qual permite trazer inferências sobre a influencia de cada parâmetro do modelo. Na figura 2.12 temos representado as estimativas obtidas:

```
> exp(M1$coefficients)
(Intercept)        V1A12        V1A13        V1A14           V2        V3A31        V3A32        V3A33
  0.4421476    0.8409828    0.4142938    0.1631715    1.0401148    1.3918845    0.6177583    0.3196918
       V3A34        V4A41       V4A410        V4A42        V4A43        V4A44        V4A45        V4A46
  0.1297148    0.2151018    0.2795028    0.3487143    0.3396475    0.7319968    0.8443534    1.6933023
       V4A48        V4A49           V5        V6A62        V6A63        V6A64        V6A65        V7A72
  0.1087163    0.3595853    1.0001357    0.8965488    1.1677598    0.3295416    0.2507228    0.7688844
       V7A73        V7A74        V7A75           V8        V9A92        V9A93        V9A94       V10A102
  0.6739711    0.3749039    0.7862752    1.3258111    1.2355235    0.5739223    0.9286889    1.4293774
      V10A103          V11      V12A122      V12A123      V12A124          V13      V14A142      V14A143
  0.3155736    0.8385358    1.2757487    1.0767897    2.6838985    0.9946509    0.7029613    0.5416552
      V15A152      V15A153          V16      V17A172      V17A173      V17A174          V18      V19A192
  0.5558433    0.2533975    1.6244741    6.2069642    6.0764623    4.7885552    1.3454656    0.6197308
      V20A202
  0.3459776
> |
```

Figura 2.12: ODDS

Como exemplo de interpretação, podemos observar a variável V2 (Duração em Meses) com *odds* = 1.04, isso indica que para uma alteração em uma unidade (1 mes) em V2, a chance de que a observação seja classificada como sendo igual a 1 (mau pagador) aumenta em 4% ((1,04-1)*100).

Como exemplo para a variáveis categóricas, podemos observar variável V17A172 (variável 17 fator "A172") com *odds* = 6.0764, isso indica que a chance da resposta ser igual a 1 (mau pagador) é 6.0764 vezes maior quando a observação tem na variável V17 o fator "A172".