Big Data na Prática 4 - Customer Churn Analytics

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 655
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
##  $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
##  $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...
##  $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...
##  $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...
##  $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...
##  $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...
##  $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
##  $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
##  $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

Os dados brutos contém 7043 linhas (clientes) e 21 colunas (recursos). A coluna “Churn” é o nosso alvo.

```
##       customerID           gender    SeniorCitizen          Partner
##                0                0                0                0
##       Dependents           tenure     PhoneService    MultipleLines
##                0                0                0                0
##  InternetService   OnlineSecurity     OnlineBackup DeviceProtection
##                0                0                0                0
##      TechSupport      StreamingTV  StreamingMovies         Contract
##                0                0                0                0
## PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
##                0                0                0               11
##            Churn
##                0
```

1. Vamos mudar “No internet service” para “No” por seis colunas, que são: “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “streamingTV”, “streamingMovies”.
2. Vamos mudar “No phone service” para “No” para a coluna “MultipleLines”

Como a permanência mínima é de 1 mês e a permanência máxima é de 72 meses, podemos agrupá-los em cinco grupos de posse (tenure): “0-12 Mês”, “12–24 Mês”, “24–48 Meses”, “48–60 Mês” Mês ”,“> 60 Mês”

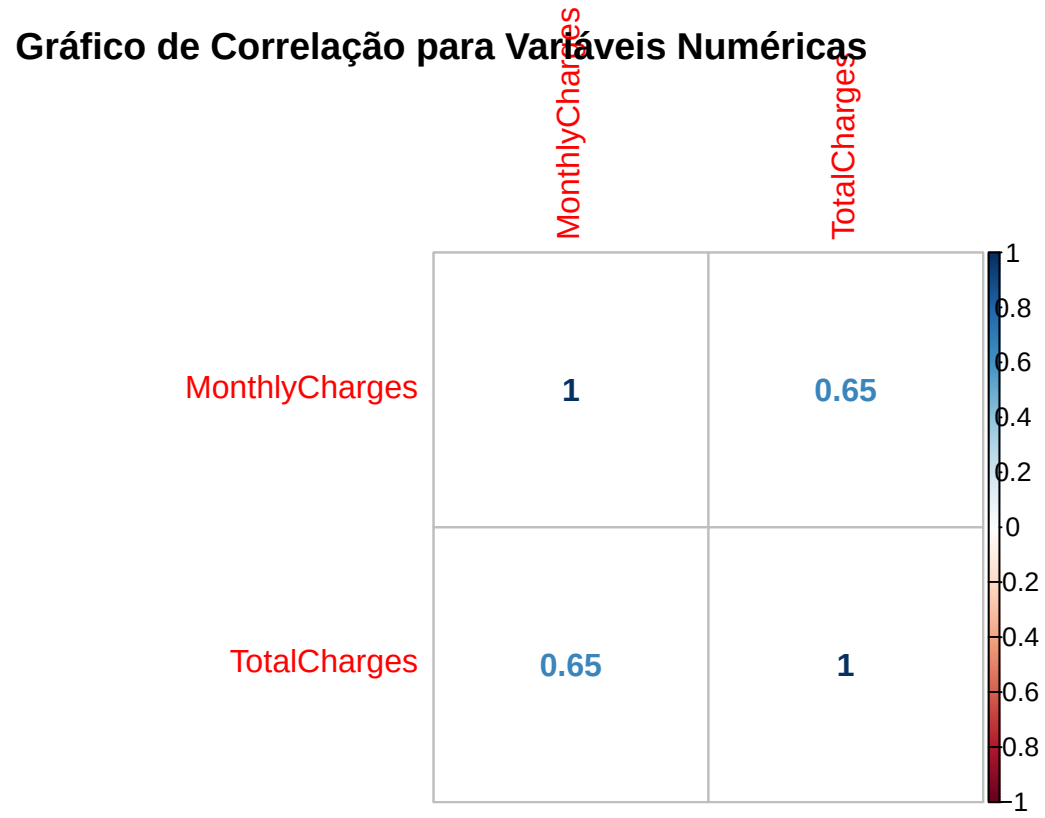
```
## [1] 1
```

```
## [1] 72
```

Alteramos os valores na coluna “SeniorCitizen” de 0 ou 1 para “No” ou “Yes”.

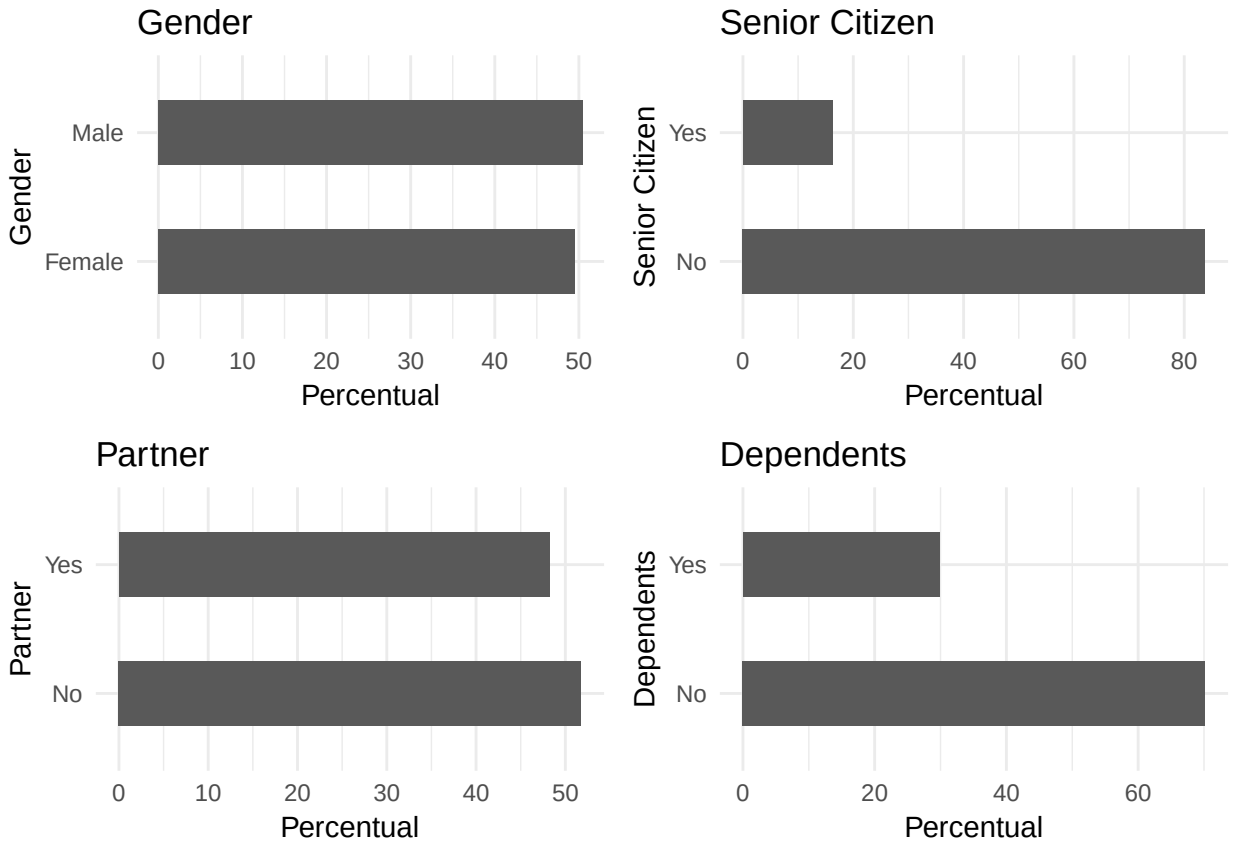
Removemos as colunas que não precisamos para a análise.

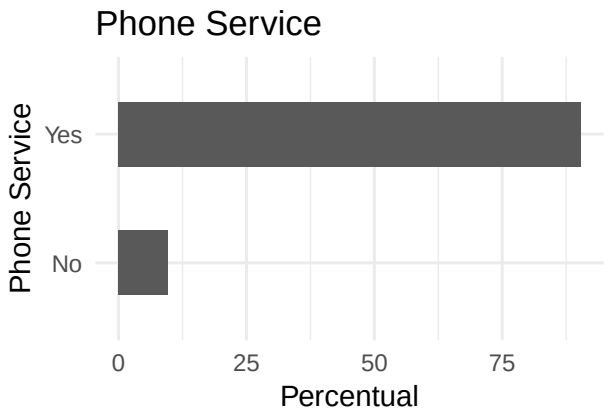
Análise exploratória de dados e seleção de recursos

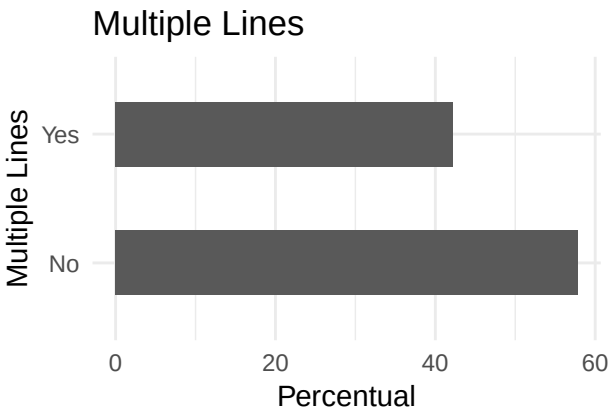


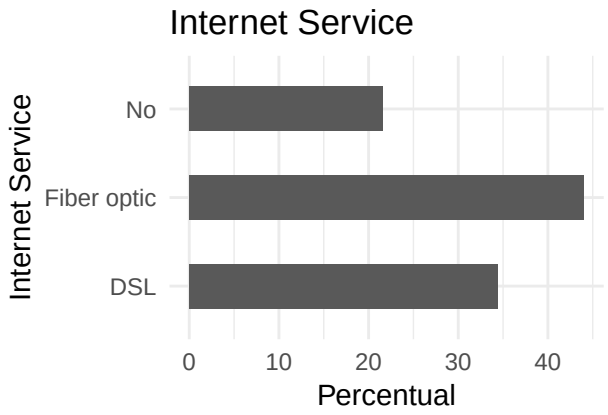
Os encargos mensais e os encargos totais estão correlacionados.

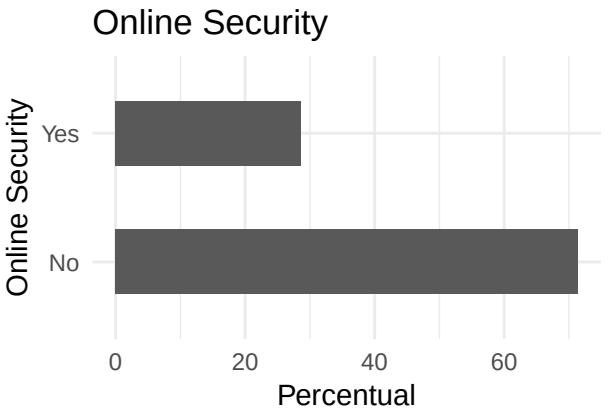
Gráficos de barra de variáveis categóricas

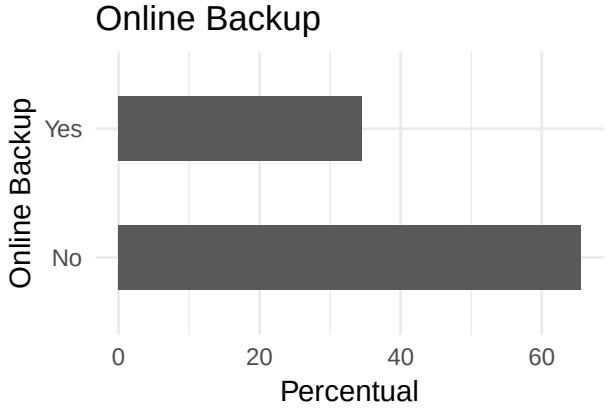


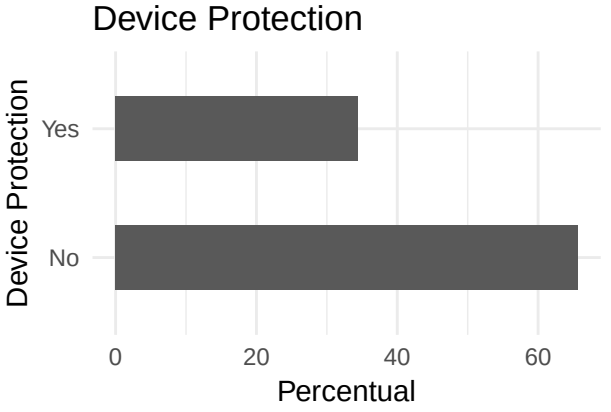

Phone Service
Phone Service
Yes
No
0
25
50
75
Percentual

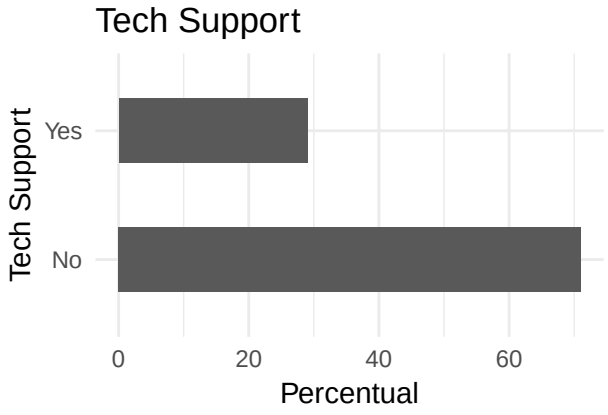

Multiple Lines
Multiple Lines
Yes
No
0
20
40
60
Percentual

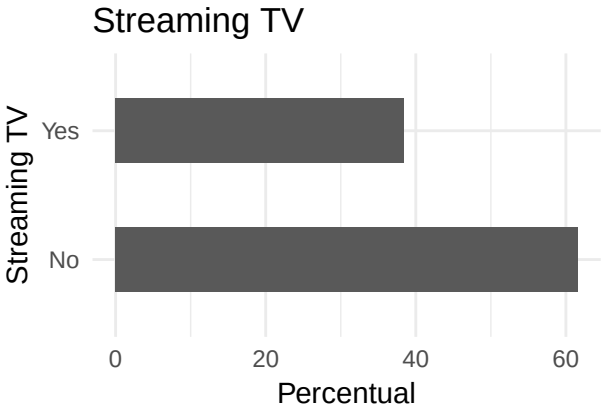

Internet Service
Internet Service
No
Fiber optic
DSL
0
10
20
30
40
Percentual

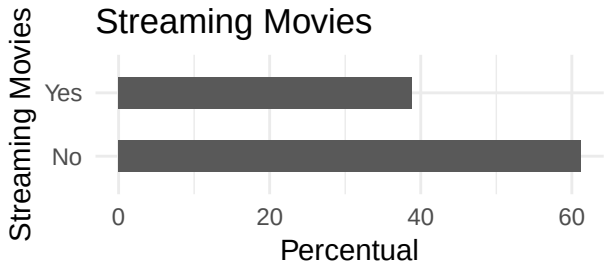

Online Security
Online Security
Yes
No
0
20
40
60
Percentual

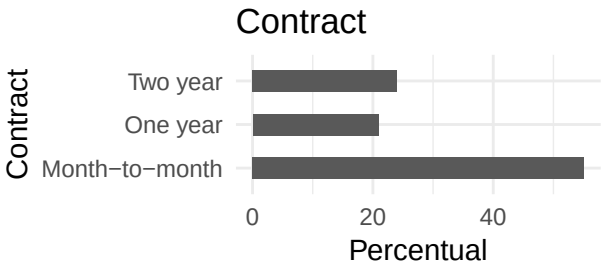

Online Backup
Online Backup
Yes
No
0
20
40
60
Percentual

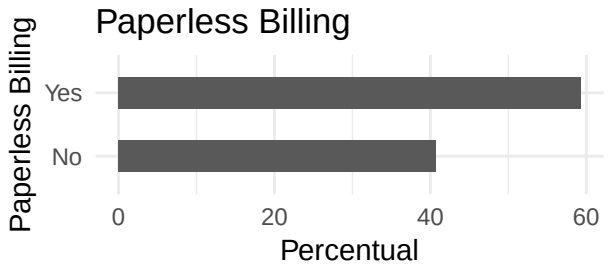

Device Protection
Device Protection
Yes
No
0
20
40
60
Percentual

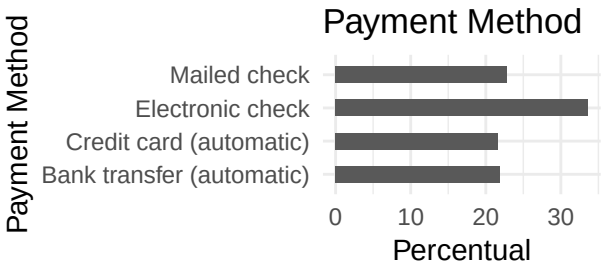

Tech Support
Tech Support
Yes
No
0
20
40
60
Percentual

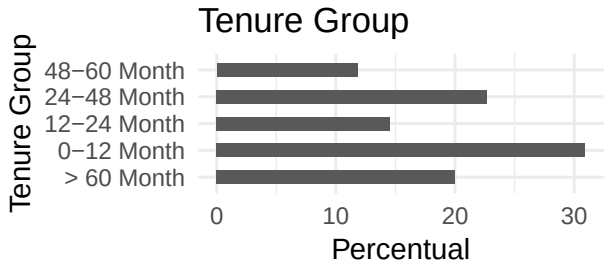

Streaming TV
Streaming TV
Yes
No
0
20
40
60
Percentual











Todas as variáveis categóricas parecem ter uma distribuição razoavelmente ampla, portanto, todas elas serão mantidas para análise posterior.

Regressão Logística

Primeiro, dividimos os dados em conjuntos de treinamento e testes

Confirme se a divisão está correta

```
## [1] 4924   19
```

```
## [1] 2108   19
```

Treinando o modelo de regressão logística

```
##
## Call:
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0474  -0.6755  -0.3052   0.6456   3.0392
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.48923    0.98512  -0.497 0.619457
## genderMale                       -0.01980    0.07748  -0.256 0.798313
## SeniorCitizenYes                  0.39039    0.10115   3.859 0.000114
## PartnerYes                       -0.06753    0.09239  -0.731 0.464772
```

```
## DependentsYes                          -0.04785    0.10702  -0.447 0.654790
## PhoneServiceYes                         0.72975    0.77648   0.940 0.347315
## MultipleLinesYes                        0.59532    0.21093   2.822 0.004767
## InternetServiceFiber optic              2.54438    0.95525   2.664 0.007732
## InternetServiceNo                      -2.38472    0.96467  -2.472 0.013434
## OnlineSecurityYes                      -0.15912    0.21454  -0.742 0.458276
## OnlineBackupYes                         0.18331    0.20900   0.877 0.380456
## DeviceProtectionYes                     0.26773    0.21132   1.267 0.205179
## TechSupportYes                          0.04331    0.21602   0.200 0.841102
## StreamingTVYes                          0.79399    0.39000   2.036 0.041765
## StreamingMoviesYes                      0.87442    0.39169   2.232 0.025585
## ContractOne year                       -0.63504    0.12912  -4.918 8.73e-07
## ContractTwo year                       -1.40751    0.20485  -6.871 6.37e-12
## PaperlessBillingYes                     0.33081    0.08914   3.711 0.000206
## PaymentMethodCredit card (automatic) -0.07053    0.13548  -0.521 0.602627
## PaymentMethodElectronic check           0.28852    0.11219   2.572 0.010116
## PaymentMethodMailed check               0.06800    0.13637   0.499 0.618032
## MonthlyCharges                         -0.06203    0.03794  -1.635 0.102091
## tenure_group0-12 Month                  1.80966    0.20067   9.018  < 2e-16
## tenure_group12-24 Month                 0.76770    0.19556   3.926 8.65e-05
## tenure_group24-48 Month                 0.45292    0.17858   2.536 0.011207
## tenure_group48-60 Month                 0.29783    0.19309   1.542 0.122955
##
## (Intercept)
## genderMale
## SeniorCitizenYes                       ***
## PartnerYes
## DependentsYes
## PhoneServiceYes
## MultipleLinesYes                       **
## InternetServiceFiber optic             **
## InternetServiceNo                      *
## OnlineSecurityYes
## OnlineBackupYes
## DeviceProtectionYes
## TechSupportYes
## StreamingTVYes                         *
## StreamingMoviesYes                     *
## ContractOne year                       ***
## ContractTwo year                       ***
## PaperlessBillingYes                    ***
## PaymentMethodCredit card (automatic)
## PaymentMethodElectronic check          *
## PaymentMethodMailed check
## MonthlyCharges
## tenure_group0-12 Month                 ***
## tenure_group12-24 Month                ***
## tenure_group24-48 Month                *
## tenure_group48-60 Month
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 5702.8  on 4923  degrees of freedom
## Residual deviance: 4124.0  on 4898  degrees of freedom
## AIC: 4176
##
## Number of Fisher Scoring iterations: 6
```

Análise de recursos:

1. Os três principais recursos mais relevantes incluem Contrato, Faturamento sem papel e grupo de posse, todos variáveis categóricas.

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Churn
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                4923     5702.8
## gender            1     0.27      4922     5702.5 0.6054342
## SeniorCitizen     1   127.97      4921     5574.5 < 2.2e-16 ***
## Partner           1   112.16      4920     5462.4 < 2.2e-16 ***
## Dependents        1    24.46      4919     5437.9 7.581e-07 ***
## PhoneService      1     0.45      4918     5437.4 0.5000830
## MultipleLines     1     5.86      4917     5431.6 0.0154598 *
## InternetService   2   477.08      4915     4954.5 < 2.2e-16 ***
## OnlineSecurity    1   183.48      4914     4771.0 < 2.2e-16 ***
## OnlineBackup      1    63.28      4913     4707.7 1.792e-15 ***
## DeviceProtection  1    46.77      4912     4661.0 7.971e-12 ***
## TechSupport       1    60.74      4911     4600.2 6.526e-15 ***
## StreamingTV       1     0.04      4910     4600.2 0.8363094
## StreamingMovies   1     0.51      4909     4599.7 0.4753631
## Contract          2   248.13      4907     4351.6 < 2.2e-16 ***
## PaperlessBilling  1    13.73      4906     4337.8 0.0002115 ***
## PaymentMethod     3    31.19      4903     4306.6 7.741e-07 ***
## MonthlyCharges    1     3.17      4902     4303.5 0.0751157 .
## tenure_group      4   179.45      4898     4124.0 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analisando a tabela de variância, podemos ver a queda no desvio ao adicionar cada variável uma de cada vez. Adicionar InternetService, Contract e tenure_group reduz significativamente o desvio residual. As outras variáveis, como PaymentMethod e Dependents, parecem melhorar menos o modelo, embora todos tenham valores p baixos.

Avaliando a capacidade preditiva do modelo

```
## [1] "Logistic Regression Accuracy 0.806925996204934"
```

Confusion Matrix

```
## [1] "Confusion Matrix Para Logistic Regression"
```

```
##
##     FALSE TRUE
##   0  1417  131
##   1   276  284
```

Odds Ratio

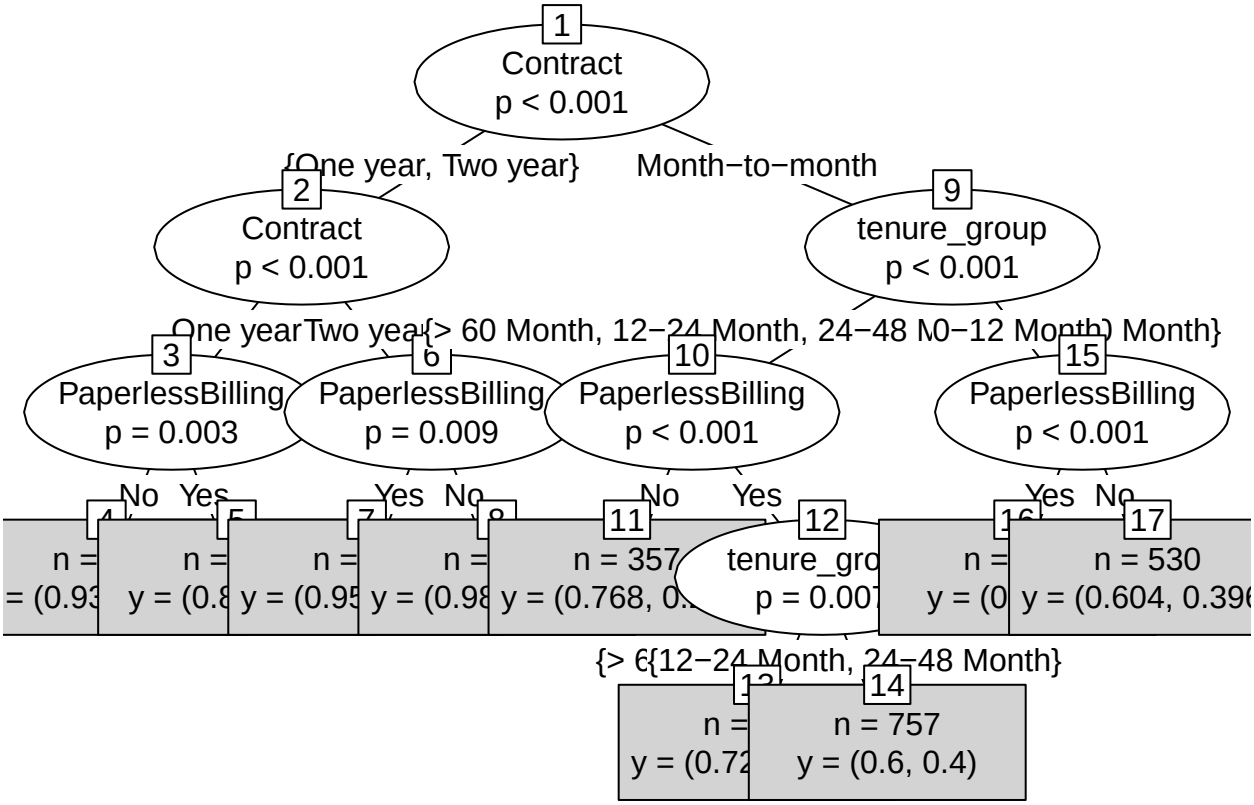
Uma das medidas de desempenho interessantes na regressão logística é Odds Ratio. Basicamente, odds ratio é a chance de um evento acontecer.

```
##                                           OR      2.5 %     97.5 %
## (Intercept)                          0.61309810 0.08883668  4.2285052
## genderMale                           0.98039732 0.84225707  1.1412103
## SeniorCitizenYes                     1.47755866 1.21175034  1.8016182
## PartnerYes                           0.93469503 0.77986897  1.1203098
## DependentsYes                        0.95327795 0.77241514  1.1751709
## PhoneServiceYes                      2.07455696 0.45315412  9.5177295
## MultipleLinesYes                     1.81361997 1.20005734  2.7441337
## InternetServiceFiber optic          12.73532097 1.96318731 83.1158200
## InternetServiceNo                    0.09211477 0.01387209  0.6093966
## OnlineSecurityYes                    0.85289122 0.55979261  1.2982994
## OnlineBackupYes                      1.20118113 0.79751875  1.8098847
## DeviceProtectionYes                  1.30699262 0.86383154  1.9783034
## TechSupportYes                       1.04426022 0.68361014  1.5946876
## StreamingTVYes                       2.21219539 1.03078653  4.7567536
## StreamingMoviesYes                   2.39749268 1.11358124  5.1728766
## ContractOne year                     0.52991651 0.41030059  0.6808574
## ContractTwo year                     0.24475084 0.16200486  0.3621221
## PaperlessBillingYes                  1.39210012 1.16931664  1.6585194
## PaymentMethodCredit card (automatic) 0.93189737 0.71426170  1.2151030
## PaymentMethodElectronic check        1.33445610 1.07183342  1.6641509
## PaymentMethodMailed check            1.07036736 0.81964198  1.3991870
## MonthlyCharges                       0.93985566 0.87240199  1.0123499
## tenure_group0-12 Month               6.10834018 4.13518592  9.0849939
## tenure_group12-24 Month              2.15481370 1.47206884  3.1702768
## tenure_group24-48 Month              1.57290007 1.11114750  2.2391577
## tenure_group48-60 Month              1.34693699 0.92252051  1.9681023
```

Para cada aumento de unidade no encargo mensal (Monthly Charge), há uma redução de 2,5% na probabilidade do cliente cancelar a assinatura.

Decision Tree

Para fins de ilustração, vamos usar apenas três variáveis para plotar árvores de decisão, elas são “Contrato”, “tenure_group” e “PaperlessBilling”.



1. Das três variáveis que usamos, o Contrato é a variável mais importante para prever a rotatividade de clientes ou não.
2. Se um cliente em um contrato de um ano ou de dois anos, não importa se ele (ela) tem ou não a PapelessBilling, ele (ela) é menos propenso a se cancelar a assinatura.
3. Por outro lado, se um cliente estiver em um contrato mensal, e no grupo de posse de 0 a 12 meses, e usando o PaperlessBilling, esse cliente terá mais chances de cancelar a assinatura.

```
## [1] "Confusion Matrix for Decision Tree"
```

```
##           Actual
## Predicted   No  Yes
##       No  1395  346
##       Yes  153  214
```

```
## [1] "Decision Tree Accuracy 0.763282732447818"
```

Random Forest

```
##
## Call:
##  randomForest(formula = Churn ~ ., data = training)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 20.92%
## Confusion matrix:
##       No Yes class.error
## No  3247 368   0.1017981
```

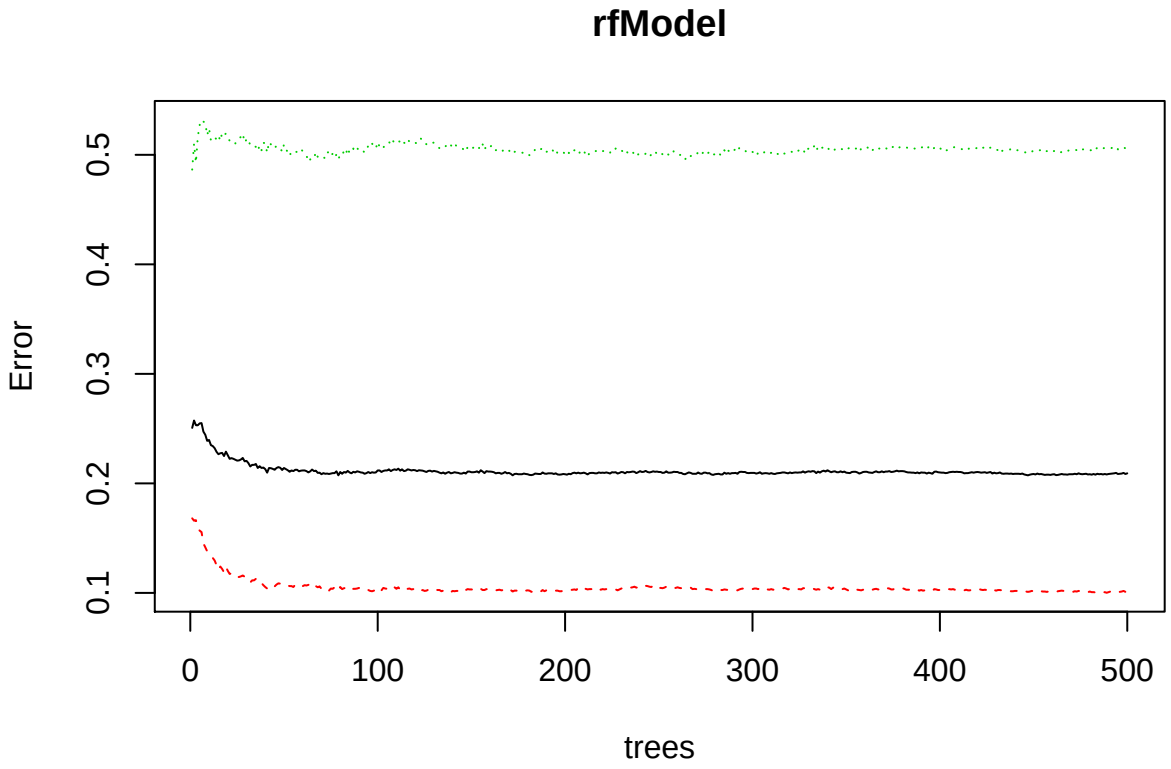
```
## Yes  662 647   0.5057296
```

A previsão é muito boa ao prever "Não". A taxa de erros é muito maior quando se prevê "sim".

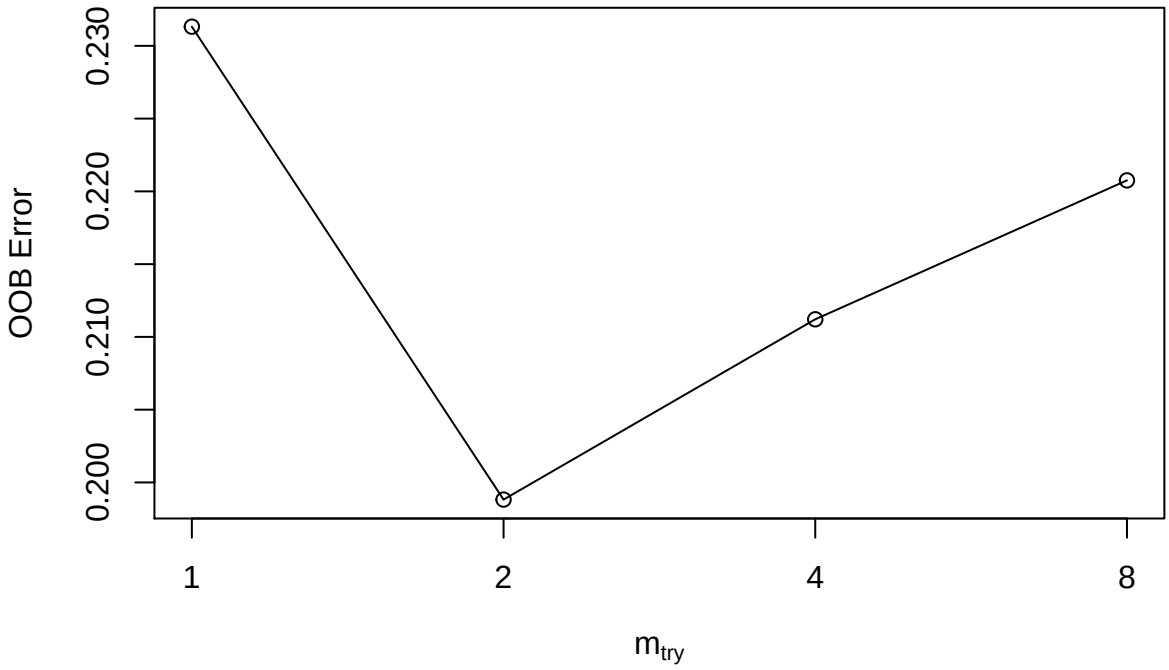
Prediction e confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1385  285
##        Yes  163  275
##
##                Accuracy : 0.7875
##                  95% CI : (0.7694, 0.8048)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 9.284e-09
##
##                   Kappa : 0.4146
##  Mcnemar's Test P-Value : 1.086e-08
##
##             Sensitivity : 0.8947
##             Specificity : 0.4911
##          Pos Pred Value : 0.8293
##          Neg Pred Value : 0.6279
##              Prevalence : 0.7343
##          Detection Rate : 0.6570
##    Detection Prevalence : 0.7922
##       Balanced Accuracy : 0.6929
##
##        'Positive' Class : No
##
```

Taxa de erro para o modelo de floresta aleatório




```
## mtry = 4  OOB error = 21.12%
## Searching left ...
## mtry = 8     OOB error = 22.08%
## -0.04519231 0.05
## Searching right ...
## mtry = 2     OOB error = 19.88%
## 0.05865385 0.05
## mtry = 1     OOB error = 23.13%
## -0.1634321 0.05
```



Ajustar o modelo de floresta aleatório novamente

```
##
## Call:
##  randomForest(formula = Churn ~ ., data = training, ntree = 200,      mtry = 2, importance = TRUE, pr
##                Type of random forest: classification
##                      Number of trees: 200
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 20.06%
## Confusion matrix:
##       No Yes class.error
## No  3300 315  0.08713693
## Yes  673 636  0.51413293
```

Torne as previsões e a matriz de confusão novamente

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  1410  306
##        Yes  138  254
##
##               Accuracy : 0.7894
##                 95% CI : (0.7713, 0.8066)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 2.734e-09
##
##                  Kappa : 0.403
```

```
##  Mcnemar's Test P-Value : 2.273e-15
##
##             Sensitivity : 0.9109
##             Specificity : 0.4536
##          Pos Pred Value : 0.8217
##          Neg Pred Value : 0.6480
##              Prevalence : 0.7343
##          Detection Rate : 0.6689
##    Detection Prevalence : 0.8140
##       Balanced Accuracy : 0.6822
##
##        'Positive' Class : No
##
```

Random Forest Feature Importance