05 MODELLING DATA

Created	
Tags	
Property	

1. Desenvolvimento e metodologia

Para garantir uma boa capacidade de generalização para a modelagem, foi criada uma estratégia de teste. Foram utilizados dados sobre os quais os modelos não foram treinados para este fim. Os modelos serão considerados com base na avaliação do desempenho sobre o conjunto de teste e comparados por meio da métrica de acurácia (AUC), sendo os algoritmos escolhidos os modelos de Machine Learning (ML): Logistic Regression, Decision Tree e Random Forest.

Para os detalhes metodológicos, seguem que:

- A definição do problema consiste em selecionar o alvo (Churn) e o tipo de problema (classificação binária).
- A ingestão de dados analisa cada recurso para maximizar seu potencial de previsão.
- Os algoritmos são treinados, e após isso testados, sendo a maneira de encontrar o melhor deles é baseada na acurácia de previsão dos resultados.
- A estratégia de avaliação e seleção de modelos indica como calcular as métricas que permitem a comparação entre os algoritmos mais bem ajustados.

A última parte da metodologia consiste em comparar o desempenho de cada algoritmo treinado usando a melhor combinação dos parâmetros.

A política de divisão (split dos dados) consiste em :

- A estratégia de teste é separar o conjunto de dados em dois subconjuntos, com proporções de 70/30 por cento do total de registros, em que 70% dos registros para treino e os 30% complementares para o teste.
- Divisão do conjunto de dados, de modo aleatório, separando um conjunto de dados de teste para a avaliação de desempenho, após o conjunto de treino ser aperfeiçoado.
- O uso de uma semente aleatória fixa (seed), no momento do split, permitindo resultados reprodutíveis.
- Nesse caso não foi utilizada a variável de tempo de processamento na avaliação dos modelos testados.

Segue um descritivo resumido sobre cada algoritmo:

Logistic Regression (LR) é um algoritmo de classificação usando um modelo linear, ou seja, calcula o recurso alvo como uma combinação linear de recursos de entrada. A LR minimiza uma função de custo específica, o que a torna apropriada para a classificação, já que envolve a regressão de variáveis preditoras em um resultado binário usando uma função de ligação binomial.

Decision Tree (DT) é um algoritmo de classificação que constrói uma árvore de decisão, em que cada nó da árvore inclui uma condição em uma das características de entrada. Quando "crescer" (ou seja, treinar) a floresta faz com que cada árvore seja

utilizada com uma amostra aleatória do conjunto de treinamento. Assim para cada ponto de decisão na árvore, considera-se um subconjunto aleatório das características de entrada.

Random Forest (RF) é feita de muitas árvores de decisão, em que cada árvore na floresta prevê um registro, e cada árvore vota para a resposta final da floresta, que seleciona a classe com mais votos.

1. Dataset final [clean_churn]

O conjunto de dados está recodificado, limpo e pronto para ser trabalhado na modelagem de predição. Os dados foram tratados para permitir a execução de testes de verificação de qual modelo tem a melhor performance na predição de churners, e com isso evidenciar os insights, e apurar quais desses são os mais exequíveis para a visão de negócio.

1.1 Resumo explicativo do Dataset

Sobre as variáveis associadas aos clientes, foram analisadas as informações demográficas sobre os clientes - qual o sexo, a identificação etária (dividida em duas condições, idosos ou não) e se os clientes têm parceiros ou dependentes.

Consta também o detalhamento dos serviços que cada cliente assinou - de telefonia (com uma ou mais linhas), o tipo de internet e os adicionais online: segurança, backup, proteção de dispositivo, suporte técnico, streaming de TV e streaming de filmes.

Além disso, as informações contratuais dos clientes - há quanto tempo é cliente, o tipo de contrato, a forma de pagamento, o tipo de faturamento e as cobranças mensais.

Para as questões das modelagens, cada modelo de Machine Learning pode demandar um refinamento no conjunto de dados, para compatibilizá-los com as métricas exigidas em cada processamento, assim, em cada caso, as regras de refinamento serão aplicadas quando necessárias.
Ao final, será feita a comparação dos modelos para aferir acurácia e eficiência, apontando por fim qual modelo deve ser seguido e, posteriormente, o levantamento de possíveis ações que possam reverter ou, ao menos, reduzir o percentual de churners.

Reiterando que o modelo é uma versão inicial, e que necessita de ajustes com o uso de mais critérios, atributos e testagem de diversos modelos para uma evolução.

```
# Dataset padrão [churn_clean].
# Verificação da estrutura do novo dataset, tratado e limpo.
str(churn_clean)
>
tibble [7,032 × 21] (S3: tbl_df/tbl/data.frame)
 $ Gender          : chr [1:7032] "Male" "Female" "Female" "Female" ...
 $ SeniorCitizen   : chr [1:7032] "No" "No" "No" "No" ...
 $ Partner         : chr [1:7032] "No" "No" "No" "Yes" ...
 $ Dependents      : chr [1:7032] "No" "No" "No" "No" ...
 $ PhoneService    : chr [1:7032] "Yes" "Yes" "Yes" "Yes" ...
 $ MultipleLines   : chr [1:7032] "No" "No" "No" "Yes" ...
 $ InternetService : chr [1:7032] "DSL" "Fiber optic" "Fiber optic" "Fiber optic" ...
 $ OnlineSecurity  : chr [1:7032] "Yes" "No" "No" "No" ...
 $ OnlineBackup    : chr [1:7032] "Yes" "No" "No" "No" ...
 $ DeviceProtection: chr [1:7032] "No" "No" "Yes" "Yes" ...
 $ TechSupport     : chr [1:7032] "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr [1:7032] "No" "No" "Yes" "Yes" ...
 $ StreamingMovies : chr [1:7032] "No" "No" "Yes" "Yes" ...
 $ Contract        : chr [1:7032] "Month-to-month" "Month-to-month" "Month-to-month" "Month-to-month" ...
 $ PaperlessBilling: chr [1:7032] "Yes" "Yes" "Yes" "Yes" ...
 $ PaymentMethod   : chr [1:7032] "Mailed check" "Electronic check" "Electronic check" "Electronic check" ...
 $ Tenure          : num [1:7032] 2 2 8 28 49 10 1 1 47 1 ...
 $ TenureYear      : chr [1:7032] "0-1 ano" "0-1 ano" "0-1 ano" "2-3 anos" ...
 $ MonthlyCharges  : num [1:7032] 53.9 70.7 99.7 104.8 103.7 ...
 $ TotalCharges    : num [1:7032] 108 152 820 3046 5036 ...
 $ Churn           : chr [1:7032] "Yes" "Yes" "Yes" "Yes" ...

# Dimensão do conjunto de dados.
dim(churn_clean)
>
[1] 7032   21
```

Efetuada transformação inicial nos dados para tipo fator (factor), já que o fator em R é uma variável que armazena os dados como um vetor de valores. A vantagem é que as variáveis categóricas são carregadas, nos modelos estatísticos, de forma diferente das

variáveis contínuas, assim, armazenar dados como fatores garante que as funções de modelagem tratem todos os dados de forma igual.

Após as demonstrações realizadas na EDA o atributo TenureYear não será mais utilizado nos modelos de ML. Assim sendo, a variável será removida do dataset.

```
# Remoção da variável pelo nome: TenureYear.
churn_clean$TenureYear  <- NULL


# Confirmando as variáveis em tipo factor.
churn_clean <- as.data.frame(unclass(churn_clean),
                      stringsAsFactors = TRUE)


# Conferindo as classes das variáveis.
sapply(churn_clean, class)
>
Gender    SeniorCitizen    Partner    Dependents    PhoneService    MultipleLines
"factor"       "factor"   "factor"      "factor"        "factor"         "factor"

InternetService    OnlineSecurity    OnlineBackup    DeviceProtection    TechSupport
       "factor"          "factor"        "factor"            "factor"       "factor"

StreamingTV    StreamingMovies
   "factor"           "factor"

Contract    PaperlessBilling    PaymentMethod
"factor"            "factor"         "factor"

Tenure    MonthlyCharges    TotalCharges
"numeric"      "numeric"       "numeric"

Churn
"factor"
```

A transformação inicial nos dados para tipo fator (factor), já que o fator em R é uma variável que armazena os dados como um vetor de valores. A vantagem é que as variáveis categóricas são carregadas, nos modelos estatísticos, de forma diferente das variáveis contínuas, assim, armazenar dados como fatores garante que as funções de modelagem tratem todos os dados de forma igual.

```
# Usando função glimpse para demonstrar os tipos de dados e das variáveis, e uma amostra dos 5 primeiros valores (editado manualmente).
glimpse(churn_clean)
>
Rows: 7,032 Columns: 20
$ Gender           <fct> Male, Female, Female, Female, Male
$ SeniorCitizen    <fct> No, No, No, No, No
$ Partner          <fct> No, No, No, Yes, No
$ Dependents       <fct> No, No, No, No, No
$ PhoneService     <fct> Yes, Yes, Yes, Yes, Yes
$ MultipleLines    <fct> No, No, Yes, Yes, Yes
$ InternetService  <fct> DSL, Fiber optic, Fiber optic, Fiber optic, Fiber optic
$ OnlineSecurity   <fct> Yes, No, No, No, No
$ OnlineBackup     <fct> Yes, No, No, No, Yes
$ DeviceProtection <fct> No, No, Yes, Yes, Yes
$ TechSupport      <fct> No, No, No, Yes, No
$ StreamingTV      <fct> No, No, Yes, Yes, Yes
$ StreamingMovies  <fct> No, No, Yes, Yes, Yes
$ Contract         <fct> Month-to-month, Month-to-month, Month-to-month, Month-to-month, Month-to-month
$ PaperlessBilling <fct> Yes, Yes, Yes, Yes, Yes
$ PaymentMethod    <fct> Mailed check, Electronic check, Electronic check, Electronic check, Bank transfer (automatic)
$ Tenure           <dbl> 2, 2, 8, 28, 49
$ MonthlyCharges   <dbl> 53.85, 70.70, 99.65, 104.80, 103.70
$ TotalCharges     <dbl> 108.15, 151.65, 820.50, 3046.05, 5036.30
$ Churn            <fct> Yes, Yes, Yes, Yes, Yes

# Temos 17 variáveis em formato factor e 3 double (numéricas).
```

```
# Conferência de existência de algum valor NA na base após as alterações.
any(is.na(churn_clean))
>
```

```
[1] FALSE

# Conferência de existência valores estranhos, zerados, nulos, após as alterações.
# Essa validação também demonstra os tipos de cada variável e quantos níveis de valores existem em cada fator.
df_status(churn_clean)
>
           variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
1            Gender       0       0    0    0     0     0  factor      2
2     SeniorCitizen       0       0    0    0     0     0  factor      2
3           Partner       0       0    0    0     0     0  factor      2
4        Dependents       0       0    0    0     0     0  factor      2
5      PhoneService       0       0    0    0     0     0  factor      2
6     MultipleLines       0       0    0    0     0     0  factor      2
7   InternetService       0       0    0    0     0     0  factor      3
8    OnlineSecurity       0       0    0    0     0     0  factor      2
9      OnlineBackup       0       0    0    0     0     0  factor      2
10 DeviceProtection       0       0    0    0     0     0  factor      2
11      TechSupport       0       0    0    0     0     0  factor      2
12      StreamingTV       0       0    0    0     0     0  factor      2
13  StreamingMovies       0       0    0    0     0     0  factor      2
14         Contract       0       0    0    0     0     0  factor      3
15 PaperlessBilling       0       0    0    0     0     0  factor      2
16    PaymentMethod       0       0    0    0     0     0  factor      4
17           Tenure       0       0    0    0     0     0 numeric     72
18   MonthlyCharges       0       0    0    0     0     0 numeric   1584
19     TotalCharges       0       0    0    0     0     0 numeric   6530
20            Churn       0       0    0    0     0     0  factor      2
```

1.1 Dataset Split

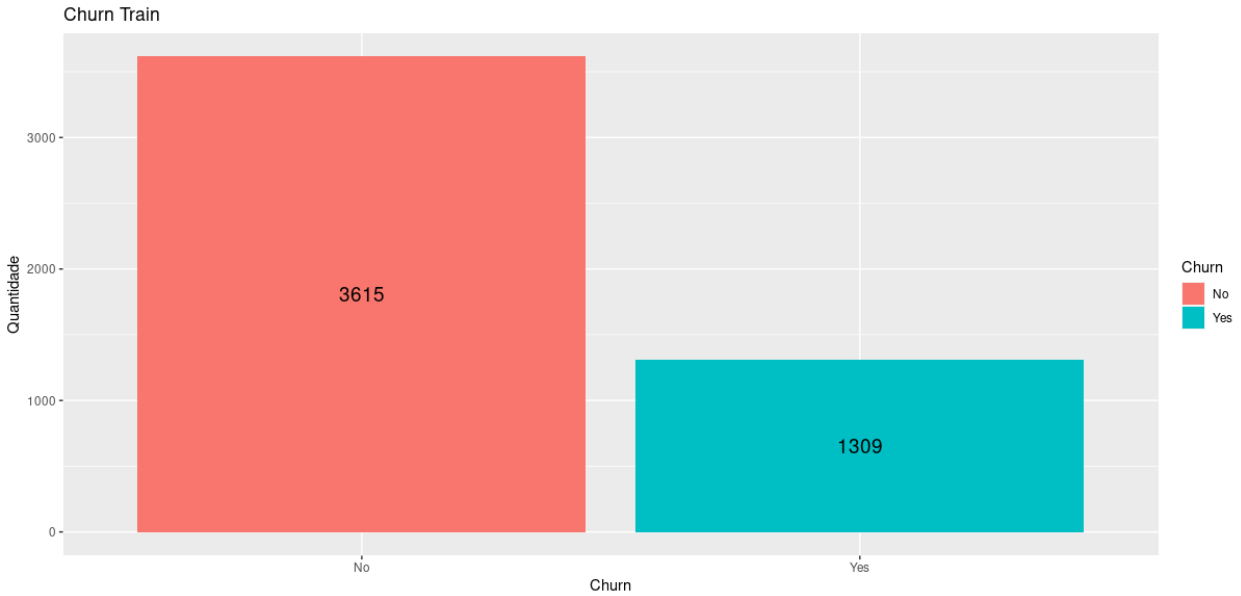
A fim de gerar condições idênticas a todos os modelos testados, o dataset [churn_clean] será dividido em dois subsets , o primeiro será o conjunto de treino com proporção de 70% dos dados, e o segundo o conjunto de teste com os 30% restantes.

Todos os modelos trabalharão com os mesmos dados, pois após o split inicial cada subset será copiado para atender a cada modelo, com isso garantir-se-á a preservação dos dados, para a configuração que cada modelagem exigir.

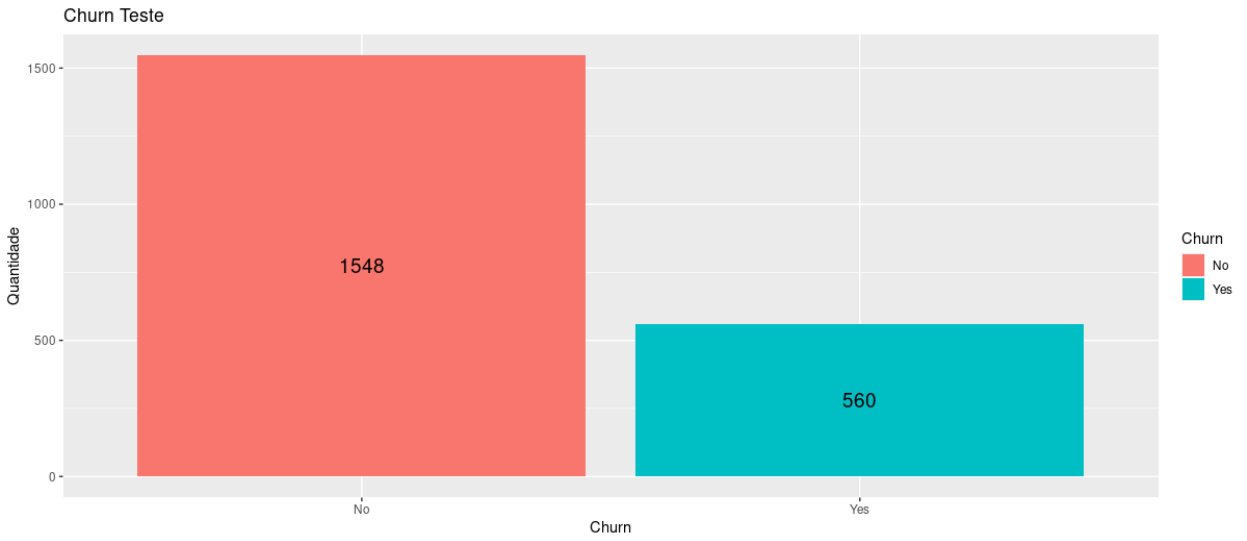
```
# Dataset Split: cteste (teste) e ctrain (treino).
set.seed(1971)

data_split  <- createDataPartition(churn_clean$Churn, p=0.7,list=FALSE)

# Geração dos subsets de treino e teste.
ctrain <- churn_clean[data_split, ]
cteste <- churn_clean[-data_split, ]


# Verificando que a proporção de valores da variável alvo manteve-se após refinamento.
prop.table(table(ctrain$Churn))
>
       No       Yes
0.7341592 0.2658408

prop.table(table(cteste$Churn))
>
       No       Yes
0.7343454 0.2656546

# Checagem das proporções cross com o quantitativo de cada subset, apenas para Churn.
CrossTable(ctrain$Churn, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)
>
Total Observations in Table:  4924

          |        No |       Yes |
          |-----------|-----------|
          |      3615 |      1309 |
          |     0.734 |     0.266 |
          |-----------|-----------|

CrossTable(cteste$Churn, prop.t=TRUE, prop.r=TRUE, prop.c=TRUE)
>
Total Observations in Table:  2108

          |        No |       Yes |
          |-----------|-----------|
          |      1548 |       560 |
```

```
          |     0.734 |     0.266 |
          |-----------|-----------|
```

```
# Plotagem do quantitativo de registros da variável Churn para o subset de treino.
ggplot(ctrain, aes(x = Churn)) +
  geom_bar(aes(fill = Churn)) +
  geom_text(aes(label = ..count..),
            stat = "count",
            position=position_stack(vjust=0.5),
            size = 5) +
  labs(title = "Churn Train",
       x = "Churn",
       y = "Quantidade")
```




```
# Plotagem do quantitativo de registros Churn para o subset de teste.
ggplot(cteste, aes(x = Churn)) +
  geom_bar(aes(fill = Churn)) +
  geom_text(aes(label = ..count..),
            stat = "count",
            position=position_stack(vjust=0.5),
            size = 5) +
  labs(title = "Churn Test",
       x = "Churn",
       y = "Quantidade")
```



1.2 Refinamento do Dataset

Como é possível, para a execução ideal do processamento, que cada algoritmo exija um refinamento específico dos dados, que ainda não tenha sido aplicado, os substes serão copiados e renomeados para identificar o tipo de algoritmo utilizado.

Assim, em caso de exigências de refinamentos não aplicados ao dataset [churn_clean] se darão nas cópias dos subsets de cada modelagem, a fim de conciliar as diferentes demandas no trato dos dados (fator, numérico, binomial, categórico) sem impactar os demais testes.

2. Modelagem

2.1 Modelo Logistic Regression (LRM)

Análises para o algoritmo Logistic Regression (LR).

```
# Criando os substes renomeados para LR Model a partir dos subsets iniciais.
ctrainLRM <- ctrain
ctesteLRM <- cteste

# Não há necessidade de recoding específico.
```

```
# Verificando a estrutura do subset de treino.
str(ctrainLRM)
>
'data.frame': 4924 obs. of  20 variables:
 $ Gender          : Factor w/ 2 levels "Female","Male": 1 1 1 2 2 2 2 1 1 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 2 2 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 1 1 1 1 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 1 2 1 2 2 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 2 2 2 3 2 1 2 2 2 2 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 2 2 1 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 2 1 1 1 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 1 2 2 1 2 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 1 2 2 1 2 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 2 2 1 2 1 2 2 2 2 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 3 3 4 3 3 3 2 2 3 ...
 $ Tenure          : num  2 8 28 1 47 1 34 15 8 18 ...
 $ MonthlyCharges  : num  70.7 99.7 104.8 20.1 99.3 ...
 $ TotalCharges    : num  151.7 820.5 3046.1 20.1 4749.1 ...
```

```
 $ Churn           : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...


# Verificando a estrutura do subset de teste.
str(ctesteLRM)
>
'data.frame': 2108 obs. of  20 variables:
 $ Gender          : Factor w/ 2 levels "Female","Male": 2 2 1 2 1 2 1 1 1 1 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 2 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 1 1 1 2 2 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 1 2 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 2 1 1 1 2 2 2 2 2 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 1 2 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 1 1 1 1 1 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 2 1 1 1 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 2 2 2 2 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 1 2 1 1 2 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 2 2 2 2 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 4 1 2 3 4 3 1 3 3 3 ...
 $ Tenure          : num  2 49 10 1 17 5 11 2 1 10 ...
 $ MonthlyCharges  : num  53.9 103.7 55.2 39.6 64.7 ...
 $ TotalCharges    : num  108.2 5036.3 528.4 39.6 1093.1 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...Análises para modelo Logistic Regression (LRM).
```

```
# Treinando o modelo LR.

# O ajuste do modelo será dado usando a função de modelagem linear geral [glm].
churnLRM <- glm(Churn ~.,
              data = ctrainLRM,
              family = binomial(link = 'logit'))


# Exibindo o resultado.
print(summary(churnLRM))
>
Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = ctrainLRM)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8537  -0.6691  -0.2812   0.7309   3.4188

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                          8.031e-02  9.680e-01   0.083 0.933875
GenderMale                          -1.521e-02  7.762e-02  -0.196 0.844675
SeniorCitizenYes                     1.926e-01  1.017e-01   1.894 0.058286 .
PartnerYes                          -5.407e-02  9.271e-02  -0.583 0.559725
DependentsYes                       -1.480e-01  1.069e-01  -1.384 0.166484
PhoneServiceYes                     -1.479e-01  7.721e-01  -0.192 0.848119
MultipleLinesYes                     2.676e-01  2.100e-01   1.274 0.202540
InternetServiceFiber optic           9.359e-01  9.486e-01   0.987 0.323867
InternetServiceNo                   -1.131e+00  9.594e-01  -1.179 0.238354
OnlineSecurityYes                   -2.514e-01  2.136e-01  -1.177 0.239284
OnlineBackupYes                     -1.376e-01  2.096e-01  -0.657 0.511355
DeviceProtectionYes                  6.521e-02  2.102e-01   0.310 0.756346
TechSupportYes                      -3.065e-01  2.149e-01  -1.426 0.153796
StreamingTVYes                       2.406e-01  3.879e-01   0.620 0.535122
StreamingMoviesYes                   3.341e-01  3.875e-01   0.862 0.388571
ContractOne year                    -7.178e-01  1.293e-01  -5.551 2.84e-08 ***
ContractTwo year                    -1.504e+00  2.181e-01  -6.895 5.38e-12 ***
PaperlessBillingYes                  2.633e-01  8.915e-02   2.954 0.003141 **
PaymentMethodCredit card (automatic)  5.808e-02  1.361e-01   0.427 0.669685
PaymentMethodElectronic check        4.300e-01  1.128e-01   3.811 0.000139 ***
PaymentMethodMailed check            4.655e-02  1.375e-01   0.338 0.735048
Tenure                              -5.596e-02  7.523e-03  -7.438 1.02e-13 ***
MonthlyCharges                      -9.234e-03  3.773e-02  -0.245 0.806664
TotalCharges                         2.822e-04  8.517e-05   3.313 0.000922 ***
---
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: 4065.5  on 4900  degrees of freedom
AIC: 4113.5

Number of Fisher Scoring iterations: 6

# Segundo o autor do livro Estatistica aplicada a experimentacao animal, a função pR2 realiza análises de regressão com tratamentos quantit
# variância, teste t para coeficientes. Realiza análises para falta de ajustes.
pR2(churnLRM)
>
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden          r2ML          r2CU
-2032.7576748 -2851.3798674  1637.2443852     0.2870969     0.2828734     0.4123909

# Segundo os autores do livro Numerical Ecology with R o VIF é uma medida da proporção em que a variância de um coeficiente de regressão
# é inflacionado pela presença de outra variável explicativa.
# Calcular os fatores de variação de inflação de todos os preditores em modelos de regressão.
vif(churnLRM)
>
                         GenderMale                    SeniorCitizenYes                         PartnerYes
                           1.004585                            1.135426                           1.366982
                      DependentsYes                     PhoneServiceYes                   MultipleLinesYes
                           1.285590                           33.451341                           7.204160
          InternetServiceFiber optic                  InternetServiceNo                 OnlineSecurityYes
                         143.221978                           51.124625                           5.166196
                     OnlineBackupYes                 DeviceProtectionYes                     TechSupportYes
                           6.425506                            6.389615                           5.163285
                      StreamingTVYes                  StreamingMoviesYes                   ContractOne year
                          24.327282                           24.362797                           1.352881
                   ContractTwo year                 PaperlessBillingYes PaymentMethodCredit card (automatic)
                           1.325253                            1.130976                           1.633510
       PaymentMethodElectronic check           PaymentMethodMailed check                             Tenure
                           2.117799                            2.009229                          15.961541
                     MonthlyCharges                       TotalCharges
                         668.144647                           20.601899

# Examinar o resultado da aplicação da função ANOVA ao VIF, a fim de avaliar a significância dos atributos.
```

```
# Análise de Variância - ANOVA.
anova(churnLRM, test="Chisq")
>
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     0.94      4922     5701.8 0.3325439
SeniorCitizen    1    96.68      4921     5605.1 < 2.2e-16 ***
Partner          1   131.28      4920     5473.9 < 2.2e-16 ***
Dependents       1    31.09      4919     5442.8 2.468e-08 ***
PhoneService     1     4.99      4918     5437.8 0.0254938 *
MultipleLines    1     3.95      4917     5433.8 0.0469263 *
InternetService  2   496.62      4915     4937.2 < 2.2e-16 ***
OnlineSecurity   1   155.14      4914     4782.1 < 2.2e-16 ***
OnlineBackup     1    75.91      4913     4706.2 < 2.2e-16 ***
DeviceProtection 1    38.76      4912     4667.4 4.804e-10 ***
TechSupport      1    76.32      4911     4591.1 < 2.2e-16 ***
StreamingTV      1     0.05      4910     4591.0 0.8265832
StreamingMovies  1     0.55      4909     4590.5 0.4569645
Contract         2   318.51      4907     4272.0 < 2.2e-16 ***
PaperlessBilling 1     9.16      4906     4262.8 0.0024794 **
PaymentMethod    3    41.56      4903     4221.3 4.966e-09 ***
Tenure           1   144.29      4902     4077.0 < 2.2e-16 ***
MonthlyCharges   1     0.00      4901     4077.0 0.9577336
TotalCharges     1    11.44      4900     4065.5 0.0007175 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Ao examinar os valores de significância, vemos variáveis preditoras de importância semelhantes.
# Os valores 'p' mais baixos podem ser identificados como os melhores preditores de rotatividade de clientes.

# Todas as variáveis com valor p < 0,001 (***) denotam maior significância.

# Analisando a tabela de variância, podemos ver a queda no desvio ao adicionar as variáveis abaixo.
                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
SeniorCitizen    1    96.68      4921     5605.1 < 2.2e-16 ***
Partner          1   131.28      4920     5473.9 < 2.2e-16 ***
```

```
InternetService   2   496.62      4915     4937.2 < 2.2e-16 ***
OnlineSecurity    1   155.14      4914     4782.1 < 2.2e-16 ***
OnlineBackup      1    75.91      4913     4706.2 < 2.2e-16 ***
TechSupport       1    76.32      4911     4591.1 < 2.2e-16 ***
Contract          2   318.51      4907     4272.0 < 2.2e-16 ***
Tenure            1   144.29      4902     4077.0 < 2.2e-16 ***
---
# Isso nos informa que ao adicionar esses atributos há redução significativamente do desvio residual.
```

```
# Aplicando o modelo de predição ao subset de teste.
LRM_prob1 <- predict(churnLRM,
                     ctesteLRM,
                     type="response")


LRM_pred1 <- ifelse(LRM_prob1 > 0.5,
                    "Yes",
                    "No")

# Gerando dados da acurácia inicial após primeira predição.
misClassficError <- mean(LRM_pred1 != ctesteLRM$Churn)

print(paste("Acurácia ",
            1-misClassficError))
>
[1] "Accuracy 0.804079696394687"

# Informando que a taxa de precisão do modelo de regressão logística é de 80,40%.

# Verificando o resultado da Matriz de Confusão para primeira predição.
table(Predicted = LRM_pred1,
      Actual = ctesteLRM$Churn)
>
           Actual
Predicted   No  Yes
      No  1385  250
      Yes  163  310

# As entradas diagonais fornecem nossas previsões corretas:
  # o canto superior esquerdo sendo TN e
  # o inferior direito sendo TP.
  # o canto superior direito fornece o FN e
  # o canto inferior esquerdo fornece o FP.
```

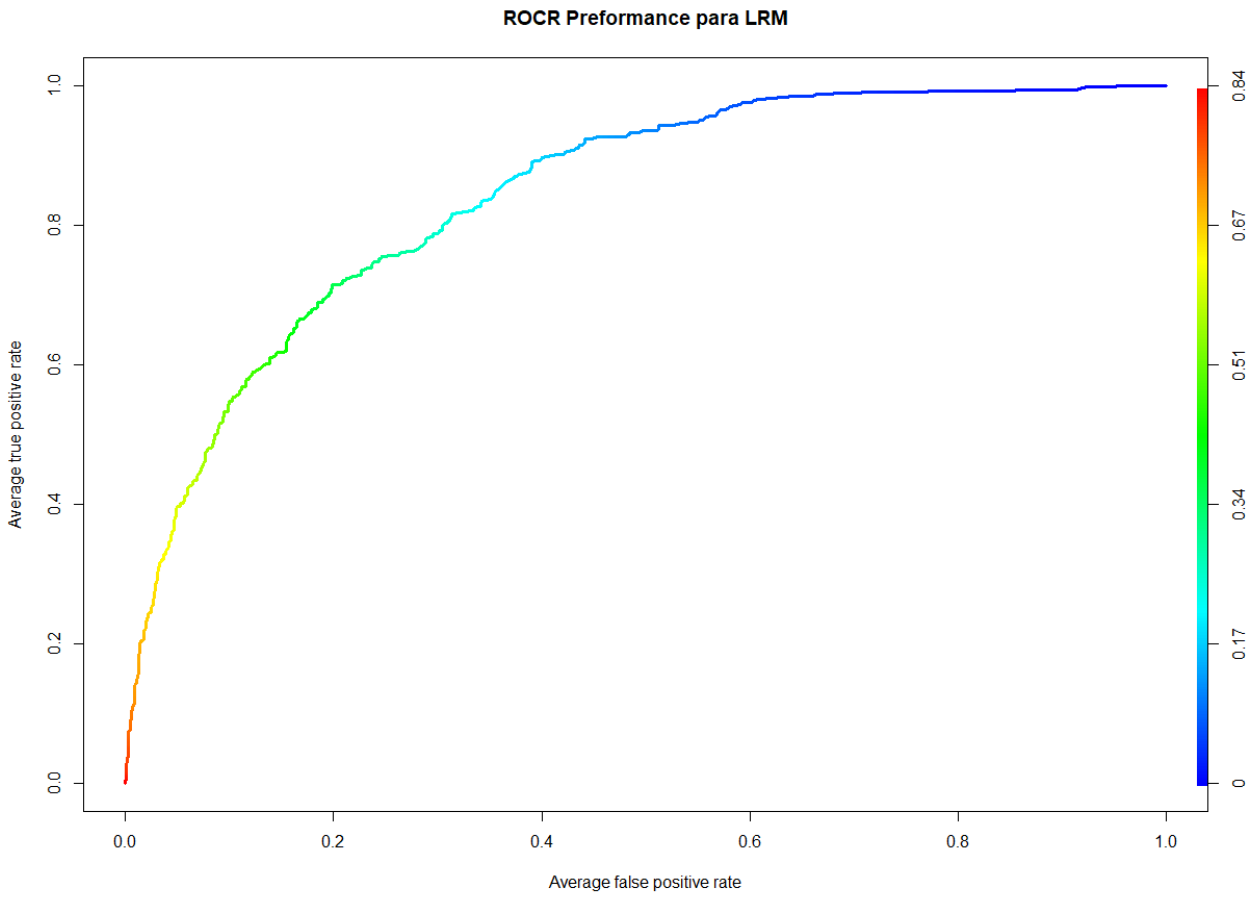
```
# Reaplicando o modelo de predição ao subset de treino e de teste, segunda predição do modelo.
LRM_prob2 <- predict(churnLRM,
                     ctrainLRM,
                     type="response")

LRM_pred2 <- ifelse(LRM_prob2 > 0.5,
                    "Yes",
                    "No")


# Gerando dados para Matriz de Cnfusão após reaplicação da predição.
LRM_tab1 <- table(Predicted = LRM_pred2,
                  Actual = ctrainLRM$Churn)

LRM_tab2 <- table(Predicted = LRM_pred1,
                  Actual = ctesteLRM$Churn)
```

```
# Matriz de Confusão para LRM após predição.

# Train.
caret::confusionMatrix(as.factor(LRM_pred2),
                       as.factor(ctrainLRM$Churn),
                       positive = "Yes" )
>
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3244  577
```

```
       Yes  371  732

               Accuracy : 0.8075
                 95% CI : (0.7962, 0.8184)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.4807
 Mcnemar's Test P-Value : 2.774e-11

            Sensitivity : 0.5592
            Specificity : 0.8974
         Pos Pred Value : 0.6636
         Neg Pred Value : 0.8490
             Prevalence : 0.2658
         Detection Rate : 0.1487
   Detection Prevalence : 0.2240
      Balanced Accuracy : 0.7283

       'Positive' Class : Yes

# Test.
caret::confusionMatrix(as.factor(LRM_pred1),
                       as.factor(ctesteLRM$Churn),
                       positive = "Yes" )
>
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1385  250
       Yes  163  310

               Accuracy : 0.8041
                 95% CI : (0.7865, 0.8208)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 4.508e-14
                  Kappa : 0.4717
 Mcnemar's Test P-Value : 2.318e-05

            Sensitivity : 0.5536
            Specificity : 0.8947
         Pos Pred Value : 0.6554
         Neg Pred Value : 0.8471
             Prevalence : 0.2657
         Detection Rate : 0.1471
   Detection Prevalence : 0.2244
      Balanced Accuracy : 0.7241

       'Positive' Class : Yes
```

```
# Gerando resultado da Acurácia.
LRM_acc <- sum(diag(LRM_tab2))/sum(LRM_tab2)

print(paste("Acurácia",LRM_acc))
>
[1] "Acurácia 0.804079696394687"  # Predição 2
```

```
# Gerando métrica de performance com base nas predições.
LRM_pred1 <- predict(churnLRM,
                     ctesteLRM,
                     type="response")

LRM_pred2 <- prediction(LRM_pred1,
                        ctesteLRM$Churn)

LRM_perf <- performance(LRM_pred2,
                        measure = "tpr",
                        x.measure = "fpr")

# Exibindo resultado graficamente.
plot(LRM_perf,
     avg= "threshold",
     colorize=TRUE,
     lwd= 3,
     main= "ROCR Preformance para LRM")
```




```
# Gerando valor de Acurácia.
LRM_auc <- performance(LRM_pred2,
                   measure = "auc")

# Carregando valor de Acurácia .
LRM_auc <- LRM_auc@y.values[[1]]

# Exibindo valor de Acurácia (AUC).
print(paste("Acurácia Final", LRM_auc))

[1] "Acurácia Final 0.84341546696198"

# A taxa de precisão do modelo é de 84,34%.
```

Análises para o algoritmo LRM apresentou taxa de precisão do modelo é de 84,34%.

Em demais testes, fazendo a redução de variáveis, mesmo as de menor significância do teste ANOVA, o valor final da acurácia não foi melhor do que os 84,34% obtidos com essa linha de execução.

Dos demais testes, o resultado anterior mais próximo foi LRM_auc = 0.8384575, com a retirada dos atributos:
Gender, Dependents, PhoneService, MultipleLines, StreamingTV, StreamingMovies, PaperlessBilling e MonthlyCharges.

2.2 Modelo Decision Tree (DTM)

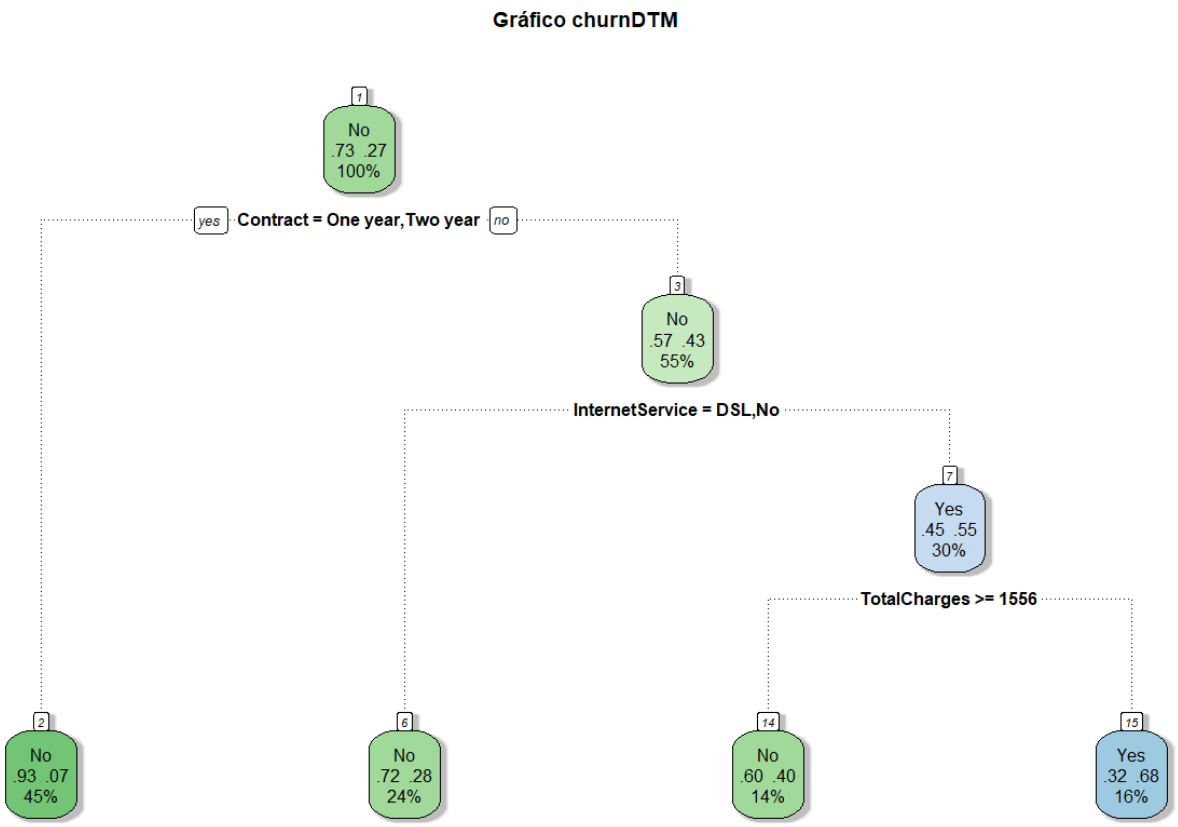
Análises para o algoritmo Decision Tree (DTM).

```
# Criando os substes renomeados para DT Model a partir dos subsets iniciais.
ctrainDTM <- ctrain
ctesteDTM <- cteste

# Não há necessidade de recoding específico.
```

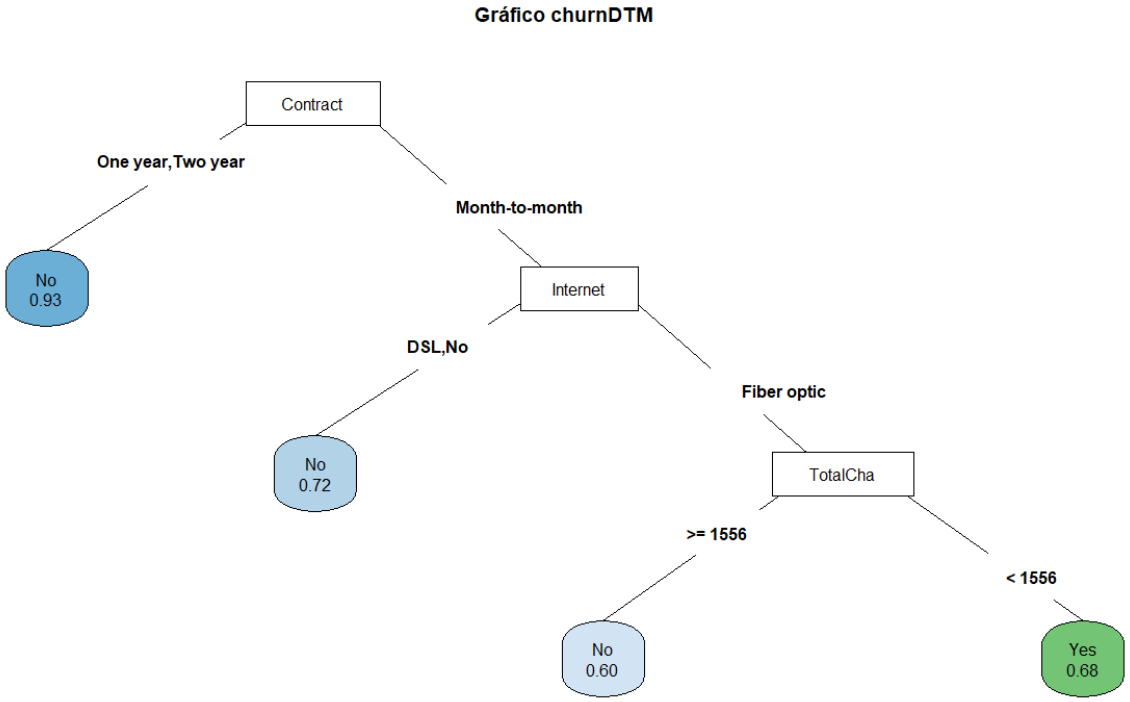
```
# Verificando a estrutura do subset de treino.
str(ctrainDTM)
>
'data.frame': 4924 obs. of  20 variables:
 $ Gender          : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 2 1 2 2 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 2 1 2 2 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 2 1 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 2 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 1 1 2 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 2 2 2 2 1 1 2 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 1 2 2 1 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 1 2 1 1 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 2 1 2 1 2 1 1 2 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 2 1 1 2 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 1 1 2 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 2 1 1 1 1 2 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 1 3 3 2 3 1 4 2 ...
 $ Tenure          : num  1 34 45 2 8 22 28 62 13 58 ...
 $ MonthlyCharges  : num  29.9 57 42.3 70.7 99.7 ...
 $ TotalCharges    : num  29.9 1889.5 1840.8 151.7 820.5 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 2 1 1 1 ...


# Verificando a estrutura do subset de teste.
str(ctesteDTM)
>
'data.frame': 2108 obs. of  20 variables:
 $ Gender          : Factor w/ 2 levels "Female","Male": 2 1 2 2 2 1 2 1 2 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 2 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 1 1 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 1 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 3 2 2 1 1 1 1 2 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 1 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 1 1 1 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 1 1 2 2 2 2 1 1 1 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 1 1 1 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 1 2 1 1 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 2 2 1 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 1 3 1 1 1 1 1 1 1 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 2 2 1 2 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 4 4 2 1 3 2 3 4 1 3 ...
 $ Tenure          : num  2 10 16 49 25 10 1 17 1 5 ...
 $ MonthlyCharges  : num  53.9 29.8 18.9 103.7 105.5 ...
 $ TotalCharges    : num  108 302 327 5036 2686 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 2 2 2 1 2 ...
```

```
# Treinando o modelo DT.
churnDTM <- rpart(Churn ~.,
                  data = ctrainDTM,
                  method="class")
```

```
# Exibindo o resultado detalhado graficamente.
fancyRpartPlot(churnDTM,
               main = 'Gráfico churnDTM')

# Gráfico com nível de detalhamento nos percentuais para Churn mais relevantes para o modelo.
```




```
# Exibindo o resultado resumido graficamente .
prp(churnDTM,
    type = 4,
    extra = 6,
    box.palette = "auto",
    faclen = 0,
    main = "Gráfico churnDTM")

# Gráfico com nível de detalhamento nos atributos para Churn mais relevantes para o modelo.
```



A partir desta árvore de decisão, podemos interpretar o seguinte:

A variável Contract com 1 ano ou 2 anos tem menos Taxa de Churn no modelo. Clientes com contratos mensais são mais propensos ao churn.
Os clientes sem o serviço de Internet ou com serviço DSL são menos propensos ao churn. Clientes com serviço de Fibra Ótica são mais propensos ao churn.
Os clientes com valor total de gastos igual ou superior a $1556 são menos propensos a churn. Clientes com valor total de gastos inferior a $1556 são mais propensos ao churn.

Aplicação da predição no subset produtivo.

```
# Aplicando o modelo de predição ao subset de teste.
DTM_prob1 <- predict(churnDTM,
                     newdata = ctesteDTM)

DTM_pred1 <- ifelse(DTM_prob1[ , 2] > 0.5,
                    "Yes",
                    "No")

# Gerando a Matriz de Confusão.
DTM_confMat <- table(Actual = ctesteDTM$Churn,
                     Predicted = DTM_pred1)

# Verificando o resultado da Matriz de Confusão.
DTM_confMat
>
       Predicted
Actual   No  Yes
   No  1439  109
   Yes  352  208

# As entradas diagonais fornecem nossas previsões corretas:
  # o canto superior esquerdo sendo TN;
  # o canto inferior direito sendo TP;
  # o canto superior direito fornece o FN;
  # o canto inferior esquerdo fornece o FP.
```

A partir dessa matriz de confusão, podemos ver que o modelo tem um bom desempenho na previsão de clientes que não desistem (1.439 corretos versus 109 incorretos), mas não tem um desempenho tão bom na previsão de clientes que desistem (352 incorretos versus 208 corretos).

Finalizando a execução do DTM.

```
# Efetuando nova rodada de predição para Matriz de Confusão e Acurácia.
DTM_prob2 <- predict(churnDTM,
                     data = ctrainDTM)

DTM_pred2 <- ifelse(DTM_prob2[,2] > 0.5,
                    "Yes",
                    "No")

DTM_tab1 <- table(Predicted = DTM_pred2,
                  Actual = ctrainDTM$Churn)

DTM_tab2 <- table(Predicted = DTM_pred1,
                  Actual = ctesteDTM$Churn)
```

```
# Gerando Matriz de Confusão para cada subset.

# Train.
caret::confusionMatrix(as.factor(DTM_pred2),
                       as.factor(ctrainDTM$Churn),
                       positive = "Yes")
>
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3359  753
       Yes  256  556

               Accuracy : 0.7951
                 95% CI : (0.7835, 0.8063)
    No Information Rate : 0.7342
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.4027
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4248
            Specificity : 0.9292
         Pos Pred Value : 0.6847
         Neg Pred Value : 0.8169
             Prevalence : 0.2658
         Detection Rate : 0.1129
   Detection Prevalence : 0.1649
      Balanced Accuracy : 0.6770

       'Positive' Class : Yes


# Teste.
caret::confusionMatrix(as.factor(DTM_pred1),
                       as.factor(ctesteDTM$Churn),
                       positive = "Yes")
>
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1439  352
       Yes  109  208

               Accuracy : 0.7813
                 95% CI : (0.763, 0.7988)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 3.641e-07
                  Kappa : 0.3494
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.37143
```

```
            Specificity : 0.92959
         Pos Pred Value : 0.65615
         Neg Pred Value : 0.80346
             Prevalence : 0.26565
         Detection Rate : 0.09867
   Detection Prevalence : 0.15038
      Balanced Accuracy : 0.65051

       'Positive' Class : Yes
```

```
# Carregando valor de Acurácia.
DTM_acc <- sum(diag(DTM_tab2))/sum(DTM_tab2)

# Demonstrando o valor de Acurácia.
print(paste("Acurácia Final", DTM_acc))
>
[1] "Acurácia Final 0.7813092979127133"

# A taxa de precisão do modelo é de 78,13%.
```

O algoritmo para DTM apresenta taxa de precisão de 78,13%. Esse percentual de acurácia é menor do que a modelagem de LRM.

2.3 Modelo Random Forest (RFM)

Análises para o algoritmo Random Forest (RFM).

```
# Criando os substes renomeados para DT Model a partir dos subsets iniciais.
ctrainRFM <- ctrain
ctesteRFM <- cteste

# Não há necessidade de recoding específico.
```

```
# Verificando a estrutura do subset de treino.
str(ctrainRFM)
>
'data.frame': 4924 obs. of  20 variables:
 $ Gender          : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 2 1 2 2 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 2 1 2 2 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 2 2 1 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 2 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 2 1 1 2 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 2 2 2 2 1 1 2 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 1 2 2 1 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 1 2 1 1 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 2 1 2 1 2 1 1 2 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 2 1 1 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 2 1 1 2 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 1 1 2 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 2 1 1 1 1 2 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 1 3 3 2 3 1 4 2 ...
 $ Tenure          : num  1 34 45 2 8 22 28 62 13 58 ...
 $ MonthlyCharges  : num  29.9 57 42.3 70.7 99.7 ...
 $ TotalCharges    : num  29.9 1889.5 1840.8 151.7 820.5 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 2 1 1 1 ...


# Verificando a estrutura do subset de teste.
str(ctesteRFM)
>
'data.frame': 2108 obs. of  20 variables:
 $ Gender          : Factor w/ 2 levels "Female","Male": 2 1 2 2 2 1 2 1 2 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 2 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 2 1 1 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 1 2 2 2 ...
 $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 3 2 2 1 1 1 2 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 1 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 1 1 1 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 1 1 2 2 2 2 1 1 1 ...
```

```
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 1 1 1 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 1 2 1 1 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 2 2 1 2 2 1 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 1 3 1 1 1 1 1 1 1 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 2 2 1 2 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 4 4 2 1 3 2 3 4 1 3 ...
 $ Tenure          : num  2 10 16 49 25 10 1 17 1 5 ...
 $ MonthlyCharges  : num  53.9 29.8 18.9 103.7 105.5 ...
 $ TotalCharges    : num  108 302 327 5036 2686 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 2 2 2 1 2 ...
```

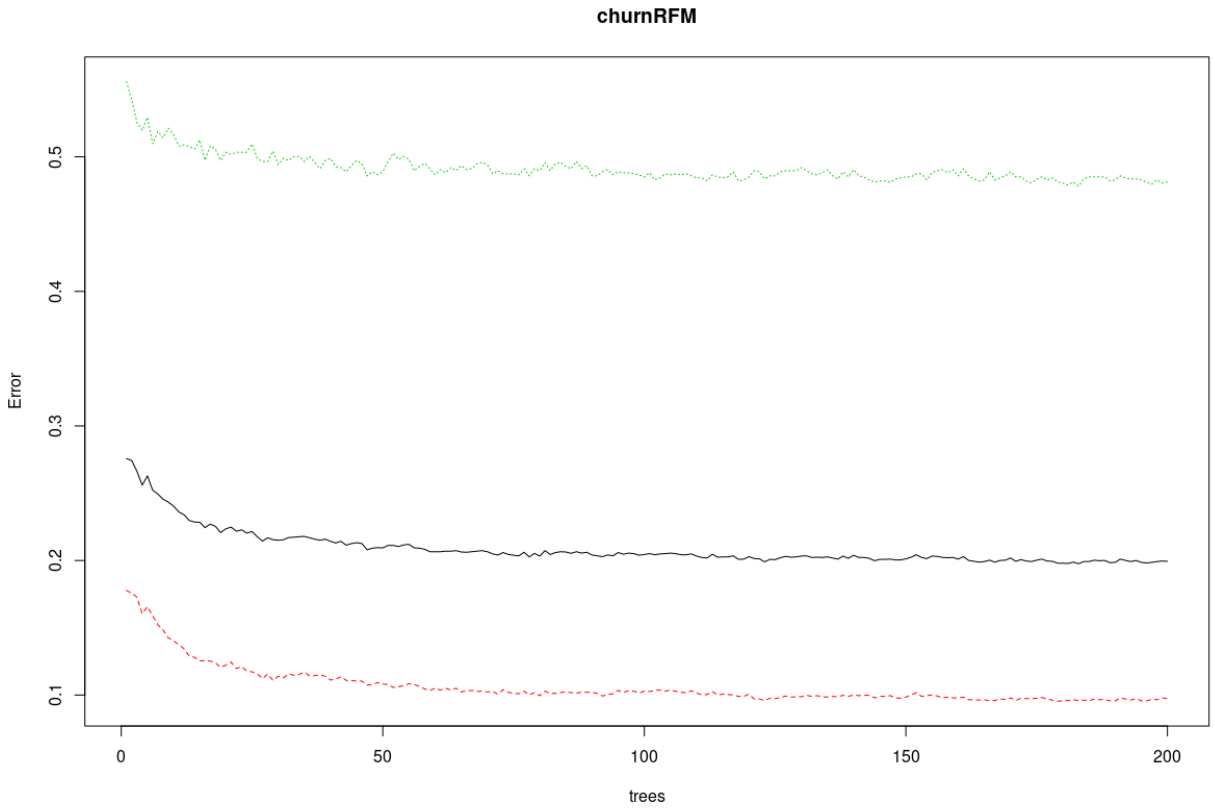
```
# Treinando o modelo inicial de RF.
churnRFM <- randomForest(Churn ~.,
                         data = ctrainRFM,
                         ntree=200,
                         type="classification")

# Exibindo o resultado do modelo sumarizado.
print(churnRFM)
>
Call:
 randomForest(formula = Churn ~ ., data = ctrainRFM, ntree = 200,      type = "classification")
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 4

        OOB estimate of  error rate: 20.53%
Confusion matrix:
      No Yes class.error
No  3250 365   0.1009682
Yes  646 663   0.4935065
```

```
# Aplicando o modelo de predição ao subset de teste.
RFM_Pred <- predict(churnRFM,
                    ctesteRFM)

# Verificando a precisão da classificação.
mean(RFM_Pred == ctesteRFM$Churn)
>
[1] 0.788425

# Verificando o resultado da Matriz de Confusão.
caret::confusionMatrix(RFM_Pred,
                       ctesteRFM$Churn)
>
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1389  287
       Yes  159  273

               Accuracy : 0.7884
                 95% CI : (0.7704, 0.8057)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 5.067e-09
                  Kappa : 0.4151
 Mcnemar's Test P-Value : 1.814e-09

            Sensitivity : 0.8973
            Specificity : 0.4875
         Pos Pred Value : 0.8288
         Neg Pred Value : 0.6319
             Prevalence : 0.7343
         Detection Rate : 0.6589
   Detection Prevalence : 0.7951
      Balanced Accuracy : 0.6924

       'Positive' Class : No
```

```
# Exibindo resultado da Matriz de Confusão.
print("Confusion Matrix Para Random Forest"); table(Actual = ctesteRFM$Churn,
                                                    Predicted = RFM_Pred)
```

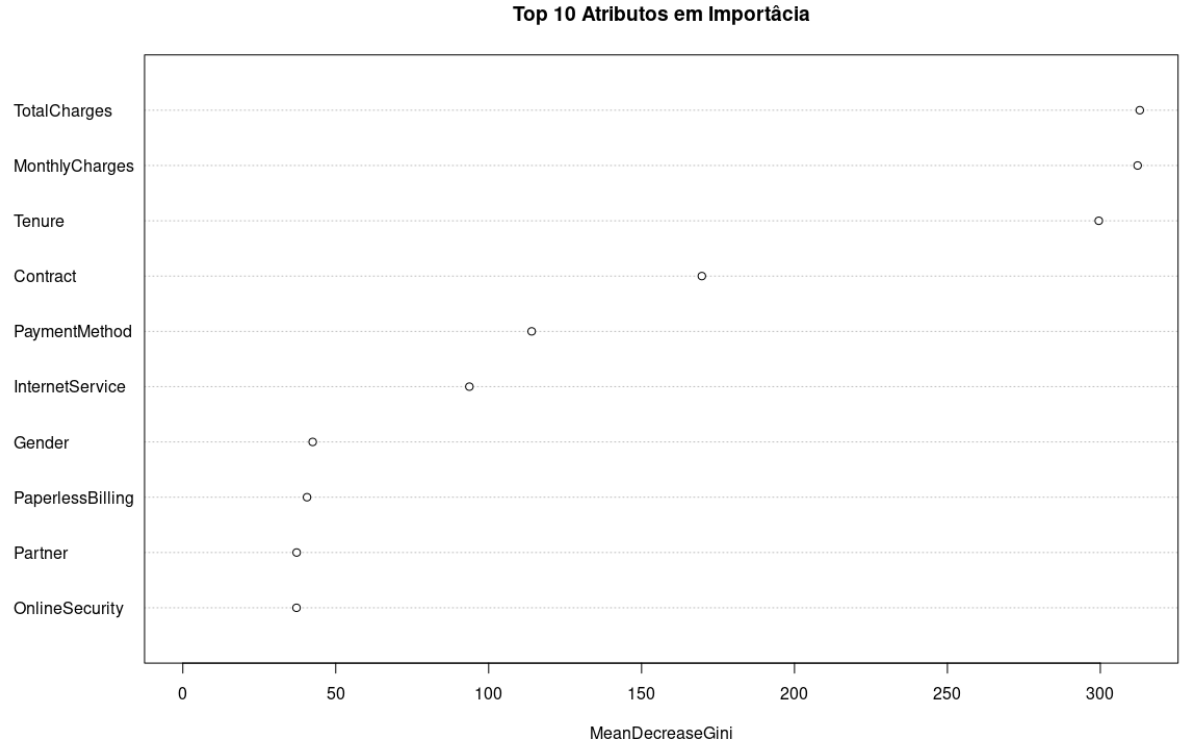
```
>
       Predicted
Actual   No  Yes
   No  1389  159
   Yes  287  273
```

```
# Visualizando graficamente a redução da taxa de erros para o modelo inicial.
plot(churnRFM)
```



Aparentemente há pouco ganho além de aproximadamente 200 trees.

```
# Exibir graficamente a importância dos atributos do modelo RF.
varImpPlot(churnRFM,
           sort=T,
           n.var = 10,
           main = 'Top 10 Feature Importance')
```



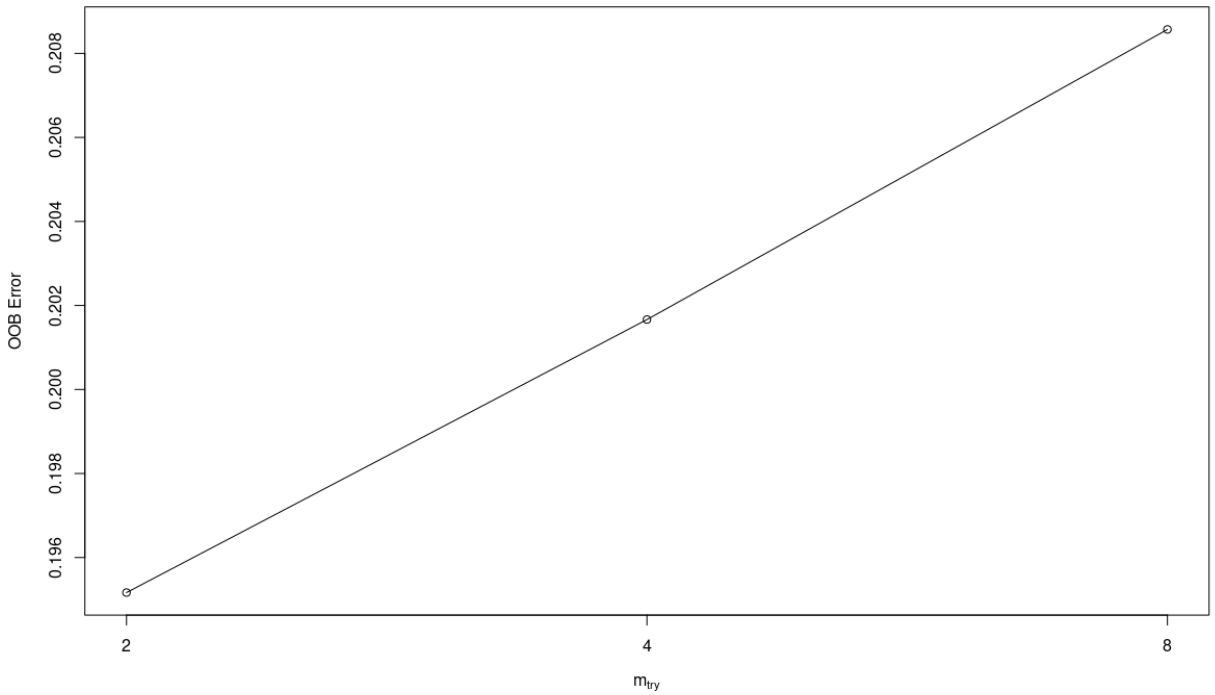

```
# Demonstrando numericamente a importância da variável do modelo RF.
importance(churnRFM)
>
                  MeanDecreaseGini
Gender                    42.45913
SeniorCitizen             32.23243
Partner                   37.22015
Dependents                33.30007
PhoneService              11.13870
MultipleLines             31.99530
InternetService           93.69749
OnlineSecurity            37.17037
OnlineBackup              34.30162
DeviceProtection          30.94710
TechSupport               35.15199
StreamingTV               28.56597
StreamingMovies           29.30604
Contract                 169.70688
PaperlessBilling          40.60528
PaymentMethod            114.06006
Tenure                   299.49777
MonthlyCharges           312.19396
TotalCharges             312.90178
```

Apenas confirmando os atributos com mais importância para o modelo.

ORDEM CLASSIFICAÇÃO	
Gender	42,45913
SeniorCitizen	32,23243
Partner	37,22015
Dependents	33,30007
PhoneService	11,1387
MultipleLines	31,9953
InternetService	93,69749
OnlineSecurity	37,17037
OnlineBackup	34,30162
DeviceProtection	30,9471
TechSupport	35,15199
StreamingTV	28,56597
StreamingMovies	29,30604
Contract	169,70688
PaperlessBilling	40,60528
PaymentMethod	114,06006
Tenure	299,49777
MonthlyCharges	312,19396
TotalCharges	312,90178

ORDEM NUMÉRICA DECRESCENTE	
TotalCharges	312,90178
MonthlyCharges	312,19396
Tenure	299,49777
Contract	169,70688
PaymentMethod	114,06006
InternetService	93,69749
Gender	42,45913
PaperlessBilling	40,60528
Partner	37,22015
OnlineSecurity	37,17037
TechSupport	35,15199
OnlineBackup	34,30162
Dependents	33,30007
SeniorCitizen	32,23243
MultipleLines	31,9953
DeviceProtection	30,9471
StreamingMovies	29,30604
StreamingTV	28,56597
PhoneService	11,1387

```
# Demonstrando numericamente a importância da variável do modelo RF.
# Afinando Modelo2 mtry com tuneRF.
churnRFM3 <- tuneRF(x = subset(ctrainRFM, select = -Churn),
                    y = ctrainRFM$Churn,
                    ntreeTry = 100,
                    doBest = TRUE)

# Para ntreeTry = 100 quando mtry = 4, OOB diminuí em 20.63%.
# Para ntreeTry = 100 quando mtry = 2, OOB diminuí em 20.04%.
```



O **OOBError** é um método de medição do erro de predição para Random Forest. É uma estimativa para subamostras de dados não tratadas no treinamento.

```
# # Exibindo o resultado sumarizado.
print(churnRFM3)
>
Call:
 randomForest(x = x, y = y, mtry = res[which.min(res[, 2]), 1])
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 19.72%
Confusion matrix:
      No Yes class.error
No  3301 314   0.0868603
Yes  657 652   0.5019099
```

```
# Performando a predição com o subset de teste.
RFM_Pred3 <- predict(churnRFM3,
                     ctesteRFM)

# Gerando a Matriz de Confusão.
caret::confusionMatrix(RFM_Pred3,
                       ctesteRFM$Churn)
>
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1413  307
       Yes  135  253

# O desempenho é um pouco semelhante ao modelo de árvore de decisão.
# A taxa de falsos negativos é baixa (1413 corretos versus 307 incorretos),
# mas a taxa de falsos positivos é alta (253 corretos versus 135 incorretos).

               Accuracy : 0.7903
                 95% CI : (0.7723, 0.8075)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 1.459e-09
                  Kappa : 0.4042
 Mcnemar's Test P-Value : 4.166e-16

            Sensitivity : 0.9128
            Specificity : 0.4518
         Pos Pred Value : 0.8215
         Neg Pred Value : 0.6521
             Prevalence : 0.7343
         Detection Rate : 0.6703
   Detection Prevalence : 0.8159
      Balanced Accuracy : 0.6823

       'Positive' Class : No
```

```
# Gerando o valor de Acurácia.
RFM_binomial <- tibble("target" = ctesteRFM$Churn,
                       "prediction" = RFM_Pred3)

RFM_basic_table <- table(RFM_binomial)

# Carregando valor de Acurácia.
RFM_acc = (RFM_basic_table[1,1] + RFM_basic_table[2,2])/sum(RFM_basic_table)

# Demonstrando o valor de Acurácia.
print(paste("Acurácia Final", RFM_acc))
>
[1] "Acurácia Final 0.790322580645161"

# A taxa de precisão do modelo é de 79,03%.
```

O algoritmo para RFM apresenta acurácia de 79,03%. Valor próximo do algoritmo DTM.

Em outro teste, com a retirada da variável TotalCharges, por estar fortemente correlacionada com a variável MonthlyCharges, o valor da acurácia 78,97% não foi melhor do que o valor obtido com essa linha de execução (testes para **churnRFM2** e **RFM_Pred2**).

Semelhante ao modelo Decision Tree, o modelo Random Forest identificou o status do contrato e da duração da posse (Tenure) como preditoras importantes para o churn. O status do serviço de Internet não aparece tão importante neste modelo, e a variável total de cobranças agora é altamente enfatizada.

O modelo Random Forest é um pouco mais preciso que o modelo de Decision Tree, sendo capaz de prever corretamente o status de churn de um cliente no subconjunto de teste com 79,03% de precisão.

Ambos os modelos envolvendo algoritmos de decisão foram semelhantes na acurácia, porém abaixo do valor da modelagem logística. Por isso o modelo com melhor performance deve ser o implantado e utilizado, e com isso aperfeiçoado para trazer mais benefícios sobre a predição de churn sobre a base de clientes. E ainda assim, os percentuais de acurácia, para os modelos RFM e DTM, ficaram abaixo do valor da modelagem LRM com 84,34%.

3. Conclusão

Em razão do algoritmo Logistic Regression ter apresentado a maior acurácia, dentre os três testados, indicando assim que esse modelo tem melhor interpretabilidade na avaliação da significância do preditor, as conclusões finais serão baseadas nos dados apurados desse modelo, especificamente.

Dentro do conjunto de dados algumas considerações foram constatadas sobre o Churn, e confirmadas com a modelagem. Os testes detalharam e corroboram análises iniciais e demonstram a relação de cada atributo com a Taxa de Churn, classificando como clientes mais ou menos propensos ao churn (churners). As avaliações estão separadas por grupo de características de cada atributo, podendo ser combinadas, com a inferência sobre o cenário e seguidas de possíveis recomendações sobre cada aspecto ou cenário, a fim de propor insights sobre os resultados, que é o objetivo final da modelagem.

Dados Demográficos

Não há diferença significativa na Taxa de Churn para o atributo gênero. O percentual de churn é maior no caso de clientes não idosos.
Clientes com Parceiros e Dependentes são menos propensos ao churn, em comparação com aqueles que não têm Parceiros e Dependentes.

Inferências com relação aos dados demográficos:

Os homens são mais propensos a desistir do que as mulheres?

▼ Explorando o gênero não há indicações de que algum dos gêneros seja mais propenso a churn do que outro.

Os idosos são mais propensos a churn?

▼ Aproximadamente 16% dos clientes são idosos. Dos clientes idosos 42% são churners (~7% do total).
Por outro lado, dos 84% de clientes que não são idosos, apenas 24% são churners.
Esses resultados tendem para que clientes não idosos são mais propensos a churn (~20% do total).

Indivíduos com parceiro mudam mais do que aqueles sem parceiro?

▼ Aproximadamente 50% dos clientes têm parceiros. Dos clientes com parceiros, 20% de churners (~10% do total).
Para pessoas sem parceiros, cerca de 33% de churners (~15,5% do total).
Esses resultados tendem para que clientes sem parceiros são mais propensos a churn.

As pessoas com dependentes se desconectam mais do que as pessoas que não têm dependentes?

▼ Aproximadamente 30% dos clientes têm dependentes. Dos clientes com dependentes, 15% de churners (~4,5% do total).
Para os 70% que não têm dependentes, 31% de churners(~22% do total).
Esses resultados tendem para que clientes sem dependentes são mais propensos a churn.

Baseadas nas inferências dos resultados dessas análises, tem-se identificados os subconjuntos de clientes em cenários de maior probabilidade de churn, para cada segmento. Uma comparação das análises com os valores de encargos totais permite recomendar em qual nicho de clientes a Telco pode concentrar esforços, com o intuito de minimizar o churn.

```
# Soma do total de encargos pagos agrupados por Churn.
aggregate(TotalCharges ~ Churn + Churn,
          data = churn_clean,
          FUN = sum)
>
  Churn TotalCharges
1    No     13193242
2   Yes      2862927

# Editado e utilizado o dataset [churn_clean] para obter o total de registros e sem valores nulos.
```

Verificando o valor total gasto por cada subconjunto de clientes.

```
# Valores agregados por Soma do TotalCharges e Churn X SeniorCitizen.
aggregate(TotalCharges ~ Churn + SeniorCitizen,
          data = churn_clean,
          FUN = sum)
>
Churn SeniorCitizen TotalCharges
1    No            No   10866095.7
2   Yes            No    1980521.8
3    No           Yes    2327146.1
4   Yes           Yes     882405.2

# Indivíduos não idosos são mais propensos a churn, e com tiveram gasto total com cobranças quase 4 vezes maior do que clientes idosos.
# O segmento de clientes churners não idosos tem uma participação de 12,33% da composição dos valores totais.

#---

# Valores agregados por Soma do TotalCharges e Churn X Partner.
aggregate(TotalCharges ~ Churn + Partner,
          data = churn_clean,
          FUN = sum)
>
Churn Partner TotalCharges
1    No      No      4460895
2   Yes      No      1306776
3    No     Yes      8732347
4   Yes     Yes      1556151

# Indivíduos sem parceiros são mais propensos a churn, e tiveram um gasto total com cobranças quase 2 vezes menor do que clientes com parce
# O segmento de clientes churners sem parceiros tem uma participação de 8,14% da composição de valores totais gastos.

#---

# Valores agregados por Soma do TotalCharges e Churn X Dependents.
aggregate(TotalCharges ~ Churn + Dependents,
          data = churn_clean,
          FUN = sum)
>
Churn Dependents TotalCharges
1    No         No    8530129.8
2   Yes         No    2261840.0
3    No        Yes    4663112.0
4   Yes        Yes     601086.9

# Indivíduos sem dependentes são mais propensos a churn, e tiveram um gasto total com cobranças 2 vezes maior do que clientes com parceiros
# O segmento de clientes churners sem dependentes tem uma participação de 14,09% da composição de valores totais gastos.
```

Com esse detalhamento a recomendação de esforços para retenção deve ser concentrada no segmento de clientes sem dependentes, pois tiveram quase $ 2,3 milhões em cobranças, em comparação ao $ 1,3 milhões de cobranças para pessoas sem parceiros e dos $ 900 mil para pessoas idosas.

Serviços Telefônicos

As Taxas de Churn apresentam grande diferença entre os clientes que possuem ou não o serviço de telefonia com mais de uma linha.

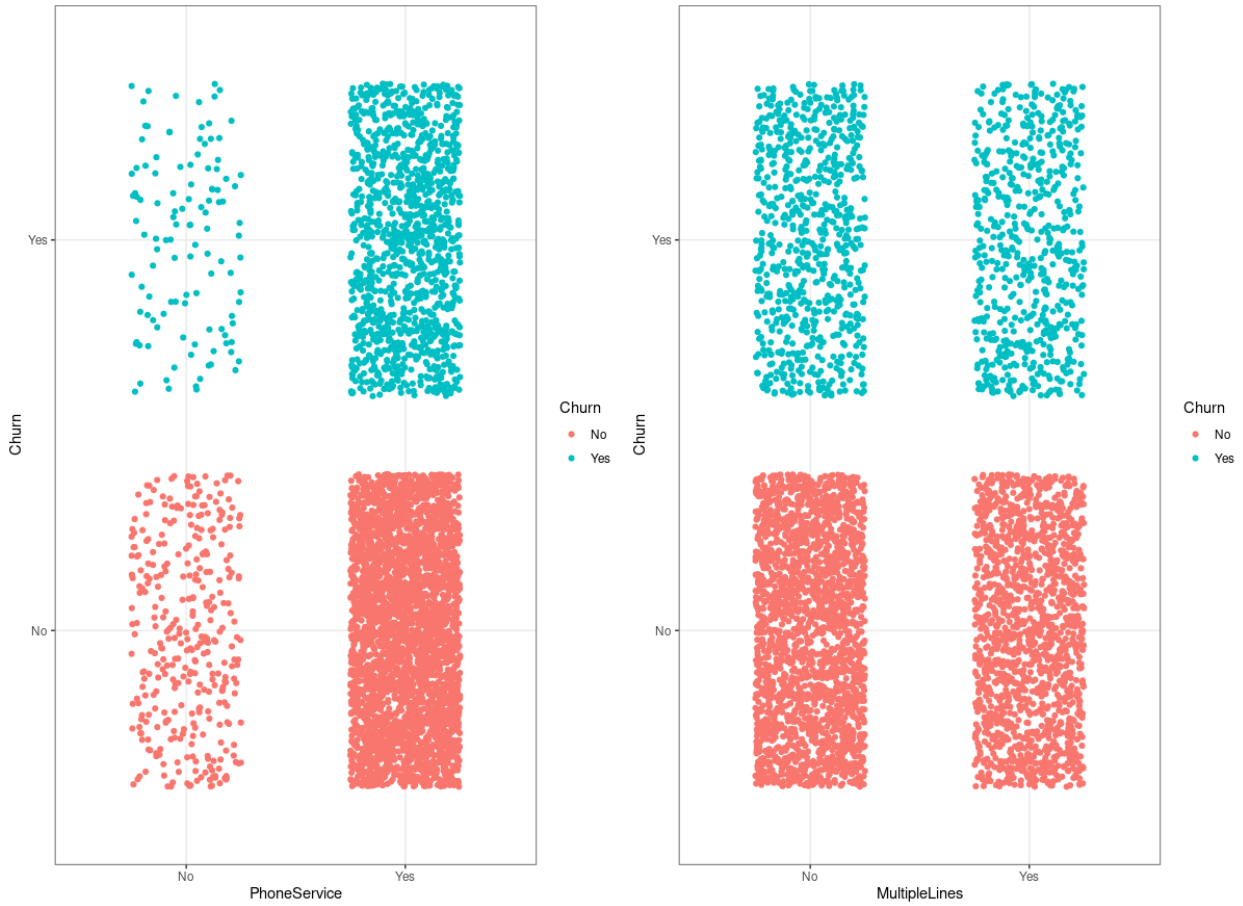
Inferências com relação aos dados de serviço de telefonia:

Clientes com uma linha, ou com duas linhas, são mais propensos a churn?

▼ Cerca de 24% de clientes com uma única linha são chuners. A quantidade de churners reduz para 12% quando associada a uma segunda linha.

A redução de 50% dos churners é indicação suficiente para uma recomendação de oferta de uma linha adicional para minimizar o churn.

```
# Plotagem com representação de churners para o PhoneService X MultipleLines usando apenas como referência o subset de treino.
p1 <- ggplot(ctrain, aes(x = PhoneService, y = Churn, color = Churn)) +
   geom_jitter(width = .25)
p2 <- ggplot(ctrain, aes(x = MultipleLines, y = Churn, color = Churn)) +
   geom_jitter(width = .25)

ggarrange(p1, p2, widths = c(2,2))

#---
# Valores agregados pela Média do tempo de fidelização.
aggregate(Tenure ~ MultipleLines + PhoneService,
          data = churn_clean,
          FUN = mean)
>
MultipleLines PhoneService   Tenure
1            No           No 31.83088
2            No          Yes 24.17046
3           Yes          Yes 41.97101

# A média de fidelização é maior para clientes possuidores de dmais de uma linha.
# Demonstrando que proporcionalmente há menos churners quando o cliente tem mais de uma linha.
```



Porém, com uma simulação do tempo médio de fidelização por tipo de contrato, pode-se agregar mais uma recomendação.

```
# Valores agregados pela Média do tempo de fidelização associada ao tipo de contrato.

# Para PhoneService.
aggregate(Tenure ~ PhoneService + Contract,
          data = churn_clean,
          FUN = mean)
>
  PhoneService       Contract   Tenure
1           No Month-to-month 18.65426
2          Yes Month-to-month 17.97028
3           No       One year 39.67586
4          Yes       One year 42.33534
5           No       Two year 55.83648
6          Yes       Two year 57.20052

# Clientes com contratos mensais tem redução de cerca de 43% na média de fidelização em comparação aos contratos de 1 ano (linhas 2 e 4)
# e entre os contratos de 2 anos, a redução da média é da ordem de 70% (linhas 2 e 6).

# Para MultipleLines.
aggregate(Tenure ~ MultipleLines + Contract,
          data = churn_clean,
          FUN = mean)
>
  MultipleLines       Contract   Tenure
1            No Month-to-month 13.06853
2           Yes Month-to-month 26.05870
3            No       One year 35.63668
4           Yes       One year 51.01786
5            No       Two year 51.08333
6           Yes       Two year 62.69505

# Clientes com contratos mensais tem redução de cerca de 50% na média de fidelização em comparação aos contratos de 1 ano (linhas 2 e 4)
# e entre os contratos de 2 anos, a redução da média é de 58% (linhas 2 e 6).
```

Com esse detalhamento a recomendação pode ser concentrada em clientes com contratos mensais, com uma linha, ofertando mais uma linha na troca por um contrato de 1 ano, visando o aumento na média de permanência, um excelente plano de retenção, buscando incremento de 100% de receita nesse nicho de churners. Para clientes com contrato mensal e possuidor de mais de uma linha, a oferta é apenas de upgrade do contrato. Nesse caso a verificação deveria ser refinada com o acréscimo de informações do consumo das múltiplas linhas para viabilizar uma recomendação atrativa.

Serviços de Internet

Com base nas análises da modelagem, o serviço de Internet por meio de Fibra Óptica é o produto que contém maior Taxa de Churn.

Inferências com relação aos dados de serviços de Internet:

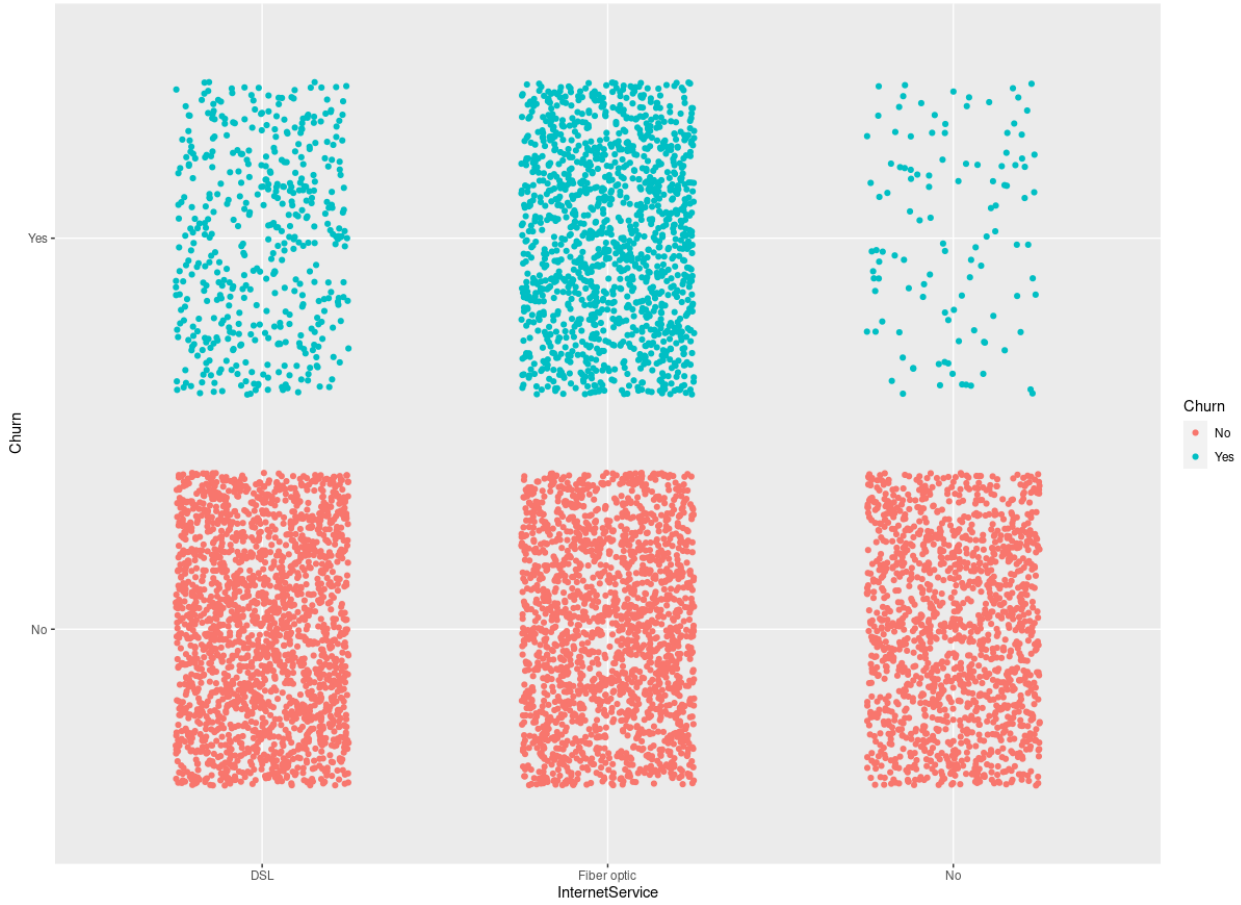
Clientes com Fibra Ótica têm cerca de 3 vezes mais propensão ao churn do que clientes com produto DSL.

- ▼ Do ponto de vista do produto Internet por meio de Fibra Ótica, contratos mensais elevam a Taxa de Churn acima dos 50%, porém há a consideração de que contratos mensais representam 85% do total.

É compreensível que os gastos mensais e totais sejam maiores com a escolha da Fibra Óptica em relação aos serviços de DSL, e seria de se supor uma Taxa de Churn menor, pois nas demais simulações cobranças maiores implicaram em maior permanência, porém não é o que demonstram as análises nesse caso.

```
# Demonstração agregada de churners para os serviços de Internet DSL X Fibra Óptica
# confrontando comm dados de contrato, forma de pagamento, média e máximo de gastos e média de tempo de fidelização.

summary(filter(churn_clean, InternetService == "DSL" & Churn == "Yes"))
>
   Churn    InternetService  OnlineSecurity  OnlineBackup  DeviceProtection  TechSupport  StreamingTV   StreamingMovies
 No :  0    DSL        :459         No :347       No :343           No :342      No :345      No :332           No :322
Yes :459    Fiber optic:  0        Yes :112      Yes :116          Yes :117     Yes :114     Yes :127          Yes :137
            No         :  0


   Churn    InternetService            Contract                  PaymentMethod      TotalCharges        Tenure
 No :  0    DSL        :459   Month to month:394    Bank transfer (automatic): 53   Min.   :  23.45    Min : 1.00
Yes :459    Fiber optic:  0   One year      : 53    Credit card (automatic)  : 72   Mean   : 784.35    Mean:14.11
            No         :  0   Two year      : 12    Electronic check         :207   Max.   :6440.25
                                                     Mailed check             :127

#---

summary(filter(churn_clean, InternetService == "Fiber optic" & Churn == "Yes"))
>
    Churn   InternetService   OnlineSecurity  OnlineBackup  DeviceProtection  TechSupport  StreamingTV  StreamingMovies
 No :   0   DSL        :   0        No :1114        No :890           No :869     No :1101      No :610          No :616
Yes :1297   Fiber optic:1297        Yes :183       Yes :407          Yes :428    Yes : 196     Yes :687         Yes :681
            No         :   0



    Churn   InternetService              Contract                  PaymentMethod      TotalCharges        Tenure
 No :   0   DSL        :   0    Month to month:1162    Bank transfer (automatic):187   Min.   :  68.5    Min : 1.00
Yes :1297   Fiber optic:1297    One year      : 104    Credit card (automatic)  :151   Mean   : 194.6    Mean:20.14
            No         :   0    Two year      :  31    Electronic check         :849   Max.   :8684.8
                                                       Mailed check             :110

# Demonstrando que proporcionalmente há menos churners quando o cliente tem o serviços DSL.
```

```
# Plotagem com representação de churners para o InternetService.

Internet_churners <- ggplot(churn_clean, aes(x = InternetService,
                            y = Churn,
                            color = Churn)) +
  geom_jitter(width = .25)

ggarrange(Internet_churners, widths = c(1,1))
```



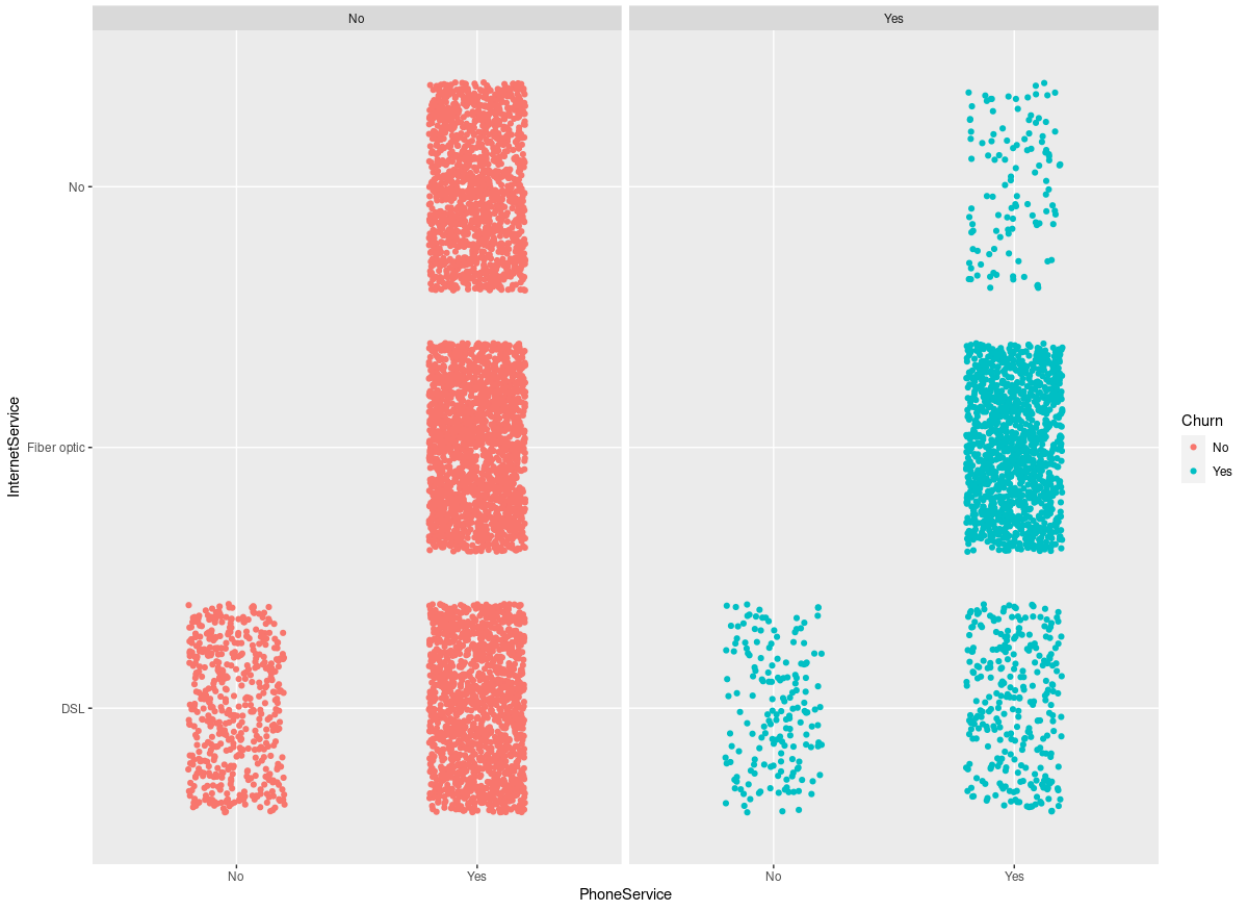
Sobre Fibra Óptica a problemática da elevada Taxa de Churn parece ser com a distribuição do serviço ou da qualidade ou do atendimento, não há uma condição de recomendação com base em insights do modelo. Precisa ser feito um outro levantamento, com mais dados, sobre o serviço para adequar a modelagem e verificar quais situações motivam os churners.

Clientes com internet de Fibra Óptica gastaram mais do que pessoas com DSL, uma possível recomendação pode ser algum tipo de redução de preço no plano de Fibra Óptica como uma promoção, em troca de fidelização, pois os clientes aparentam mais insatisfação com o serviço de Fibra Óptica (internet superior) do que com o de DSL (internet inferior). .

Um caso aparente de regra de negócio da Telco é que a empresa exige um pacote de telefonia associada ao serviço de Internet por Fibra Óptica, uma recomendação seria dissociar esses dois produtos, evitando assim sua obrigatoriedade na aquisição da Internet. Essa simulação não pode ser avaliada devida aos dados estarem atrelados, mas seria um fator de evolução para novos insights.

```
# Valores agregados por quantidade de Contrato e PhoneService X InternetService.
aggregate(Churn ~ PhoneService + InternetService,
          data = churn_clean,
          FUN = length)
>
PhoneService InternetService Churn
1           No             DSL   680
2          Yes             DSL  1736
3          Yes     Fiber optic  3096
4          Yes              No  1520

# Todos os clientes de Fibra possuem serviço telefônico.

#---

# Plotagem com representação de churners para o InternetService X PhoneService.
```

```
ggplot(churn_clean, aes(x = PhoneService, y = InternetService, color = Churn)) +
  geom_jitter(width = .2) +
  facet_wrap(~ Churn)
```



Serviços Adicionais de Internet

Mesmo sendo o serviço de Internet por meio de Fibra Óptica a Taxa de Churn mais alta, para os clientes com serviços adicionais atrelados aos seus contratos de Internet a rotatividade foi um pouco menor.

Inferências com relação aos serviços adicionais de Internet:

Clientes com serviços adicionais associados ao plano de Internet tem redução na Taxa de Churn.

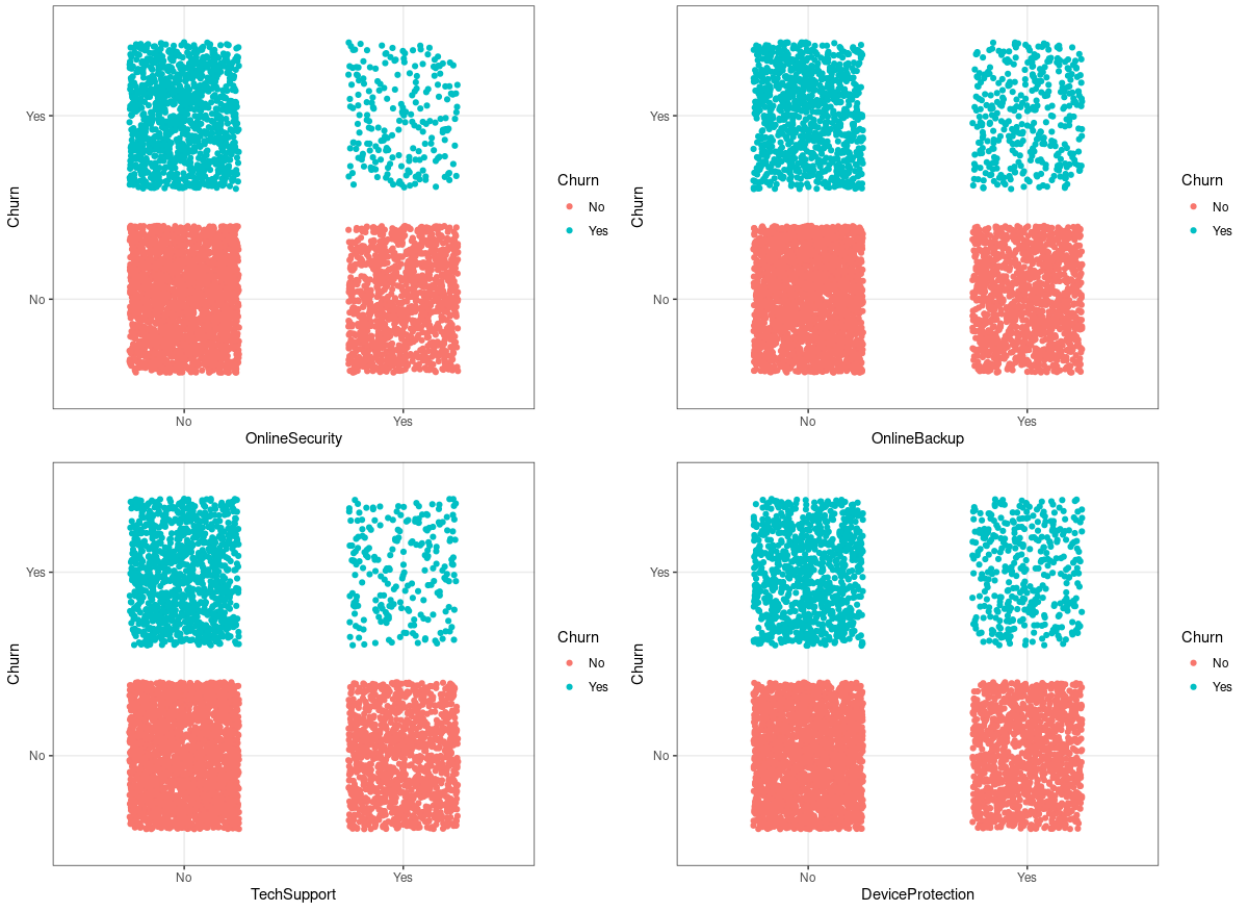
Os clientes que assinam alguns dos serviços DeviceProtection, OnlineBackup, OnlineSecurity e TechSupport têm menor taxa de rotatividade em comparação com os clientes que não assinaram, dentre os serviços adicionais a Taxa de Churn é mais baixa para os clientes com TechSupport.

Para os serviços de streaming, TV e filmes, não houve verificação substancial na variação de churn.

Clientes em contratos mensais têm maior Taxa de Churn, mesmo para os com serviços adicionais ativados.

```
# Plotagem com representação de churners para os serviços adicionais de Internet.
p1 <- ggplot(ctrain, aes(x = OnlineSecurity, y = Churn, color = Churn)) +
  geom_jitter(width = .25)
p2 <- ggplot(ctrain, aes(x = OnlineBackup, y = Churn, color = Churn)) +
  geom_jitter(width = .25)
p3 <- ggplot(ctrain, aes(x = TechSupport, y = Churn, color = Churn)) +
  geom_jitter(width = .25)
p4 <- ggplot(ctrain, aes(x = DeviceProtection, y = Churn, color = Churn)) +
  geom_jitter(width = .25)
```

```
ggarrange(p1, p2, p3, p4, widths = c(4,4))

# Demonstrando que proporcionalmente há menos churners quando o cliente tem os serviços adicionais
# e em proporção ainda menor quando os serviços são de segurança online e de suporte técnico.
```



Como recomendação possível tentar mover os clientes que não possuam esse serviços de ajuda na proteção dos seus sistemas, pode minimizar o churn.

Em uma verificação da base de clientes, apenas para os serviços com menores Taxas de Churn, segurança online ou suporte técnico, verifica-se que 70% dos clientes não possuem um dos serviços e que 57% clientes não possuem ambos os serviços.

Levando-se em conta que a média do total de encargos para clientes que não possuem ambos os serviços é de aproximadamente $ 1.182,37 com fidelização de 14 meses, e que para clientes com os dois serviços ativados, e retidos na empresa, a média de gastos sobe para $ 4.488,15 e a de fidelização para 53 meses com . Há grande chance de ao propor upgrading do plano ativando os serviços seja possível um incremento de receita.

```
# Valores agregados por Média de TotalCharges e TechSupport  X OnlineSecurity.
aggregate(TotalCharges ~ Churn + TechSupport + OnlineSecurity,
          data = churn_clean,
          FUN = mean)
>
  Churn TechSupport OnlineSecurity TotalCharges
1    No          No             No     1434.937
2   Yes          No             No     1182.370
3    No         Yes             No     3403.495
4   Yes         Yes             No     2306.578
5    No          No            Yes     3229.751
```

```
6   Yes          No            Yes     2203.379
7    No         Yes            Yes     4488.155
8   Yes         Yes            Yes     3361.682

# Valores agregados por Média de TotalCharges e TechSupport  X OnlineSecurity.
aggregate(Tenure ~ Churn + TechSupport + OnlineSecurity,
          data = churn_clean,
          FUN = mean)
>
  Churn TechSupport OnlineSecurity   Tenure
1    No          No             No 30.35609
2   Yes          No             No 14.47836
3    No         Yes             No 40.57708
4   Yes         Yes             No 25.11374
5    No          No            Yes 40.63900
6   Yes          No            Yes 25.46429
7    No         Yes            Yes 53.15647
8   Yes         Yes            Yes 36.15152

# Verificações sobre clientes sem os serviços que saíram confornтando com clientes que ativaram
# ambos os sderviços e mantiveram-se na empresa.
```

Sabendo que upgrading de planos tem um alto índice de rejeição, por grande parte de clientes, e a abordagem é mais difícil de conseguir êxito do que propor bônus e 'prêmios', uma forma de minimizar o churn e tentar não elevar gastos seria oferecer aos clientes de Internet um serviço adicional gratuitamente em troca de um contrato de pelo menos um ano. Em uma simulação deste cenário, onde os clientes contam com suporte técnico e contratos de um ou dois anos, foi verificada uma redução maior da Taxa de Churn.

Durante a modelagem observou-se que algumas variáveis, quando combinadas, apresentam uma grande melhora na redução do churn mesmo para contratos mensais. A recomendação nesses casos é criar combos com preços reduzidos para os clientes, em simulações os combos de streaming não trouxeram nenhuma melhora significativa na Taxa de Churn, assim os combos devem considerar os serviços adicionais relacionados à segurança (backup online, proteção de dispositivos e segurança online) e principalmente de suporte técnico.

De acordo com as análises das Survive Trees, após o tipo de contrato, o suporte técnico é a variável mais relevante para reduzir o risco de churn. Através de uma simulação dos contratos mensais tendo em conta o suporte técnico verificamos uma redução na probabilidade de churn, o que no primeiro ano representa 80% de sobrevivência para o cliente com suporte técnico contra 65% para os restantes clientes. A sugestão é criar uma promoção que ofereça vantagens de preço no suporte técnico, em que o desconto seja recuperado por rendimentos por maiores tempos de sobrevivência.

Contratos

As análises referentes aos contratos foram mescladas com as demais variáveis e em todos os modelos, ficou constatada a maior Taxa de Churn para contratos do tipo "month to month" (mensais), conjuntamente com qualquer modalidade de plano, ou atributo.

Os clientes com contratos do tipo mensal, compondo 55% do total, são os mais propensos a se desligar, e justamente por isso é o atributo considerado o mais relevante para definir uma estratégia de retenção de clientes, sendo que o mandato médio dos clientes que deixaram a empresa é de 10 meses (menos de 1 ano).

Claramente os contratos mensais são ponto de atenção para a empresa, que demonstram uma Taxa de Churn de 43%, com média de tempo de fidelização menor do que 12 meses, em relação aos clientes em contratos de duração anual.

Portanto, os cenários para redução da Taxa de Churn recaem em condicionar o cliente a um upgrade para planos anuais, acrescido de serviços adicionais com preços irrisórios, que cativem o cliente a manter-se na empresa, a fim de reduzir os níveis de churn.

Sendo então possível analisar em evolução do modelo se a troca de plano com ativação de serviços, com preços bem menores, tem inflluência na média de permanência do cliente por mais tempo do que os atuais 10 meses. Associada a essa recomendação, pode ser também que o consumo com valores reduzidos ou mesmo gratuitos, tenha uma validade, a fim de efetuar degustações possam atrair o cliente para outros serviços com alguma rentabilidade maior, como por exemplo os serviços de Streaming.

Método de pagamento

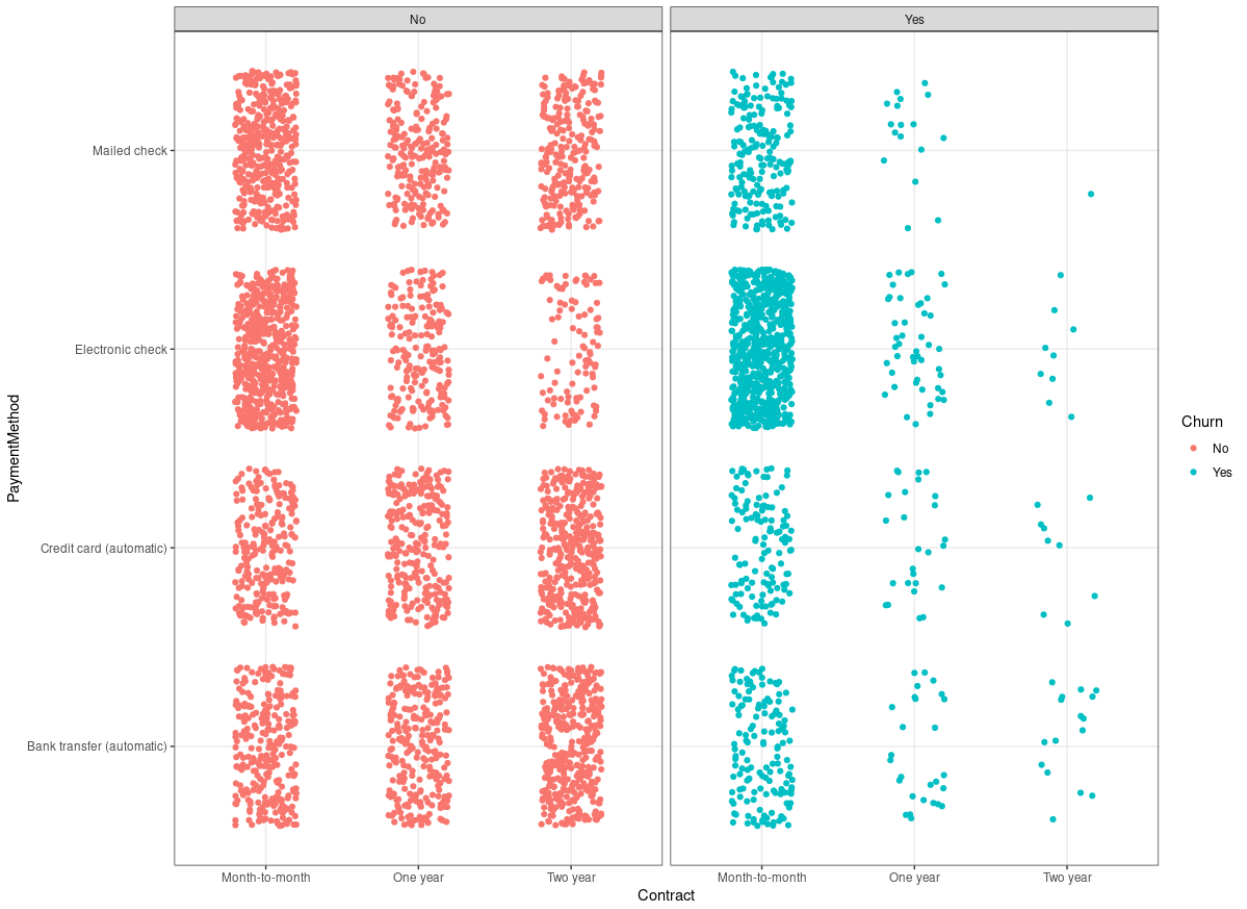
Os clientes possuem variadas formas de efetuar o pagamento dos serviços à empresa. Pensamento comum de que um maior quantidade de formas ofertadas pode favorecer os recebimentos.

Nesse caso foi verificado um problema que, aparentemente, implica em deserção, ao associar alguns atributos com as formas de pagamento disponíveis.

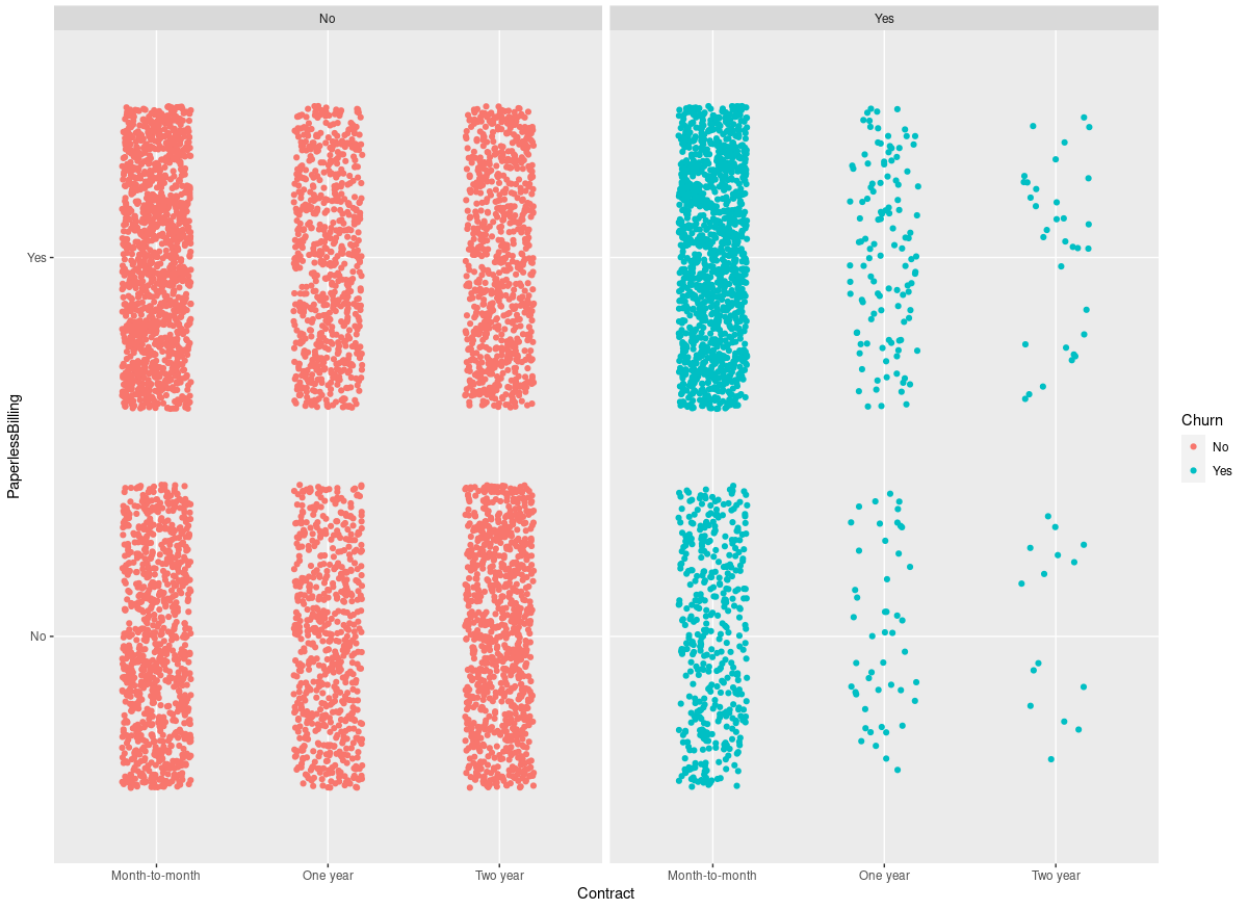
Inferências com relação aos métodos de pagamento:

Os clientes que são obrigados a interagir mensalmente para pagar as contas têm maior propensão ao churn.
A Taxa de Churn é maior no caso de clientes com opção de faturamento sem papel.
Clientes que possuem o método de pagamento “eletronic check” tendem a uma Taxa de Churn maior.

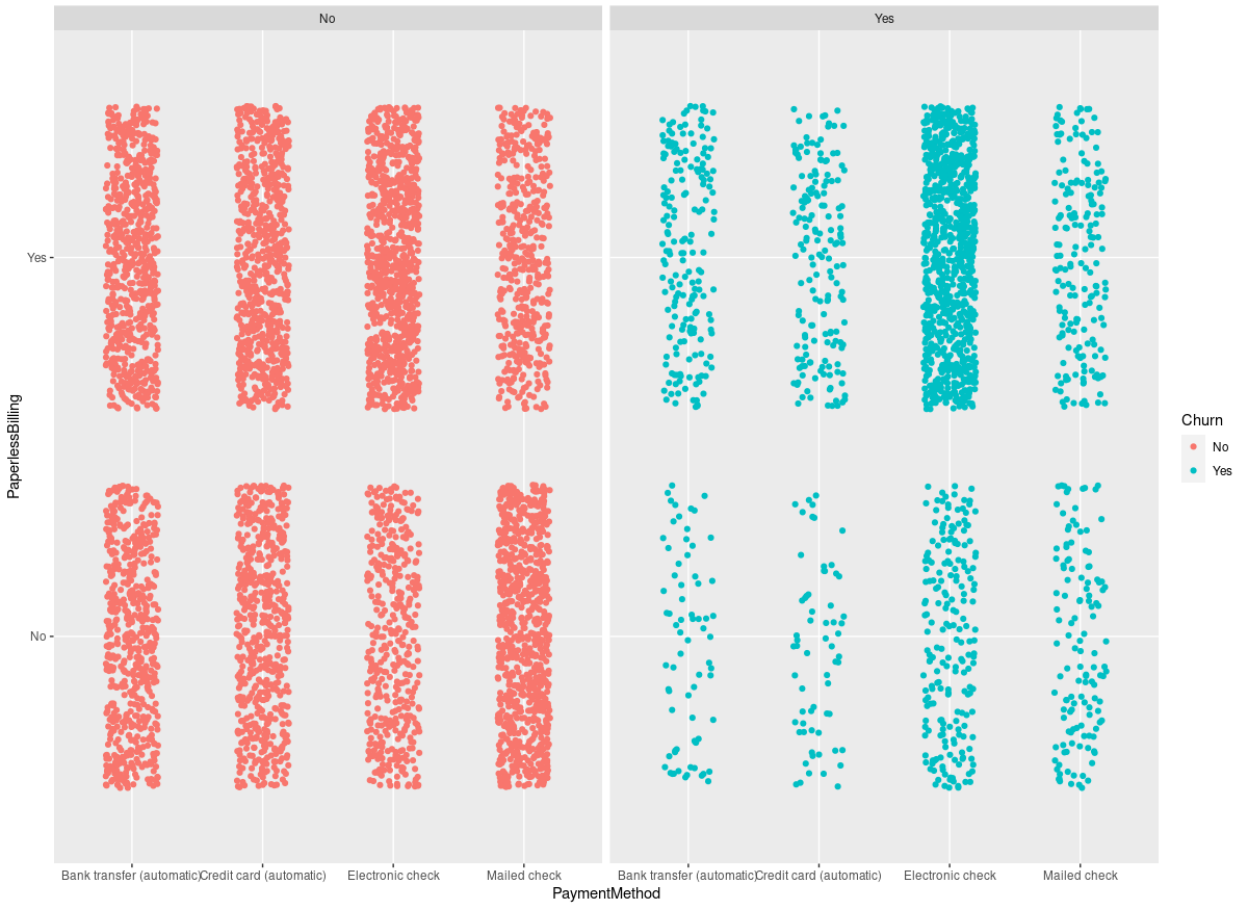
```
# Plotagem com representação de churners para contratos x métodos de pagamentos.
ggplot(ctrain, aes(x = Contract, y = PaymentMethod, color = Churn)) +
  geom_jitter(width = .2) +
  facet_wrap(~ Churn)

# Demonstrando que proporcionalmente há mais churners quando o cliente tem contrato mensal (em contrapartida aos anuais)
# e muito mais quando o pagamento é por cheque eletrônico.
```



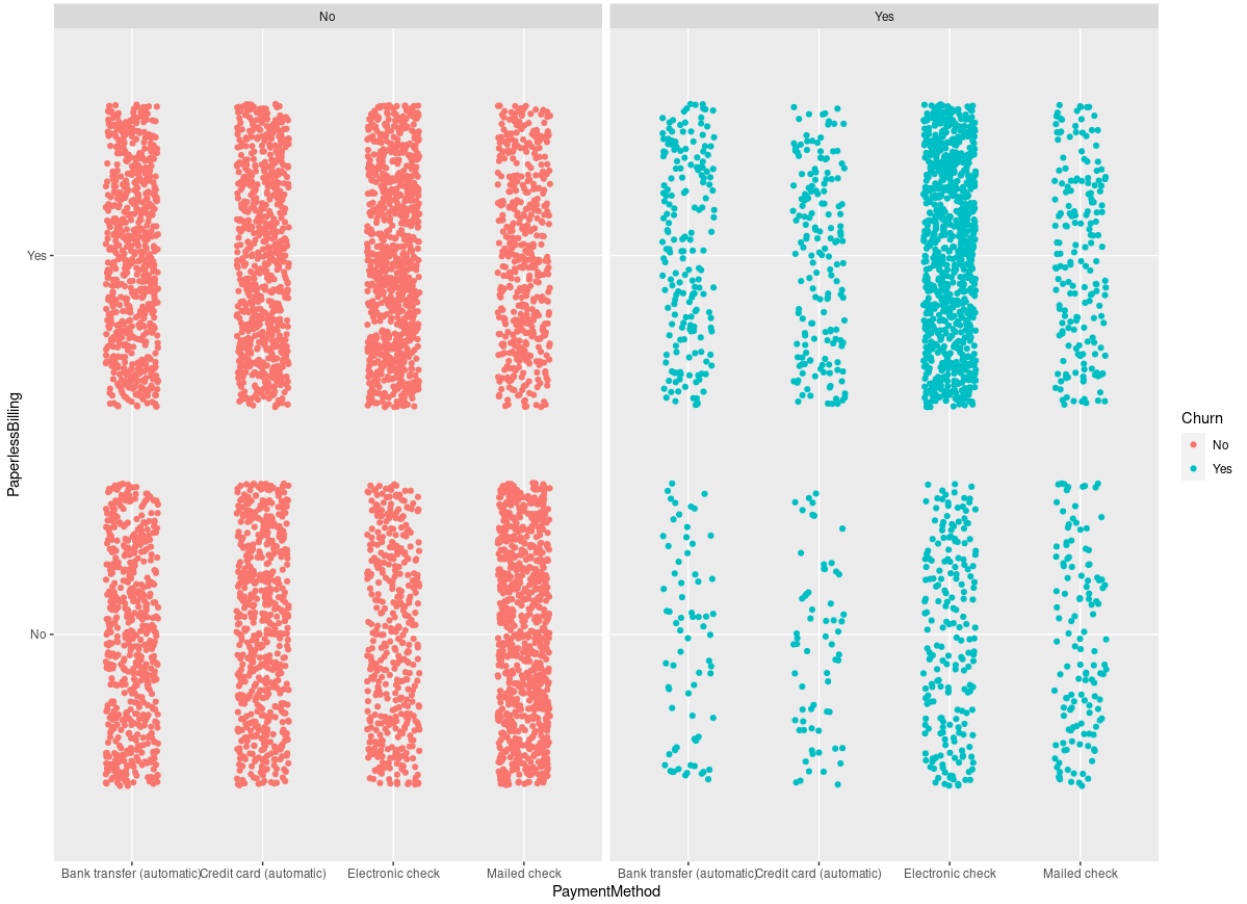

```
# Plotagem com representação de churners para contratos x tipo de fatura.
ggplot(churn_clean, aes(x = Contract, y = PaperlessBilling, color = Churn)) +
  geom_jitter(width = .2) +
  facet_wrap(~ Churn)

# Demonstrando que proporcionalmente há mais churners quando o cliente tem contrato mensal (em contrapartida aos anuais)
# e maior quando o tipo de fatura é sem papel (eletrônica).
# Nesse caso o ofensor do churn parece estar alinhado com a subscrição mensal.
```




```
# Plotagem com representação de churners para métodos de pagamentos x tipo de fatura.
ggplot(churn_clean, aes(x = PaymentMethod, y = PaperlessBilling, color = Churn)) +
  geom_jitter(width = .2) +
  facet_wrap(~ Churn)

# Demonstrando que proporcionalmente há mais churners quando o cliente tem pagamento é por cheque eletrônico
# e maior quando o tipo de fatura é sem papel (eletrônica).
```



Os pagamentos por cheques eletrônicos constituem 33% de todas as formas de pagamentos recebidos pela empresa, e é a maior Taxa de Churn, por isso as recomendações são tentar mover os clientes para o uso de meios de pagamento automático (transferência bancária 20% ou cartão de crédito 20%), mesmo para contratos mensais.



Para o pagamento em cheque eletrônico, aparentemente há uma insatisfação, mais de 60% de Taxa de Churn, e possivelmente seja um problema entre a empresa e as instituições de recebimento. Nesse caso a empresa deve tentar encontrar o problema, ou situações impeditivas, que possam existir neste serviço, avaliar junto às instituições credenciadas sobre as dificuldades e meios de contornar os problemas.

Uma ação conjunta deve ser tomada para identificar se houve alterações ou evoluções sistêmicas, que tenham avariado ou dificultado os recebimentos das faturas da Telco. Essa recomendação não pode ser avaliada com maior profundidade por falta de levantamentos informativos adequados, e deve ser tratada em evolução do modelo.

Ao contrário do senso comum, associar qualquer forma de contrato e de tipo de pagamento ao planos com tipo de fatura sem papel, que correspondem a 59% dos tipos de faturas, aumenta a Taxa de Churn. Não há recomendação nesse sentido, pois precisa ser considerado uma correlação entre os problemas de pagamento.

Essas opções eletrônicas de recebimento de fatura, em geral, geram algum desconto para o caso de clientes aceitem o recebimento da fatura por meio eletrônico, eximindo assim as cobranças de custos de papel, e sua adoção incorre em diversas economias, e outras ações de apoio socioambiental.

Consideração final

Em todos os casos, as recomendações carecem de estudo aprofundado de custos reais, como a troca de serviços adicionais por maior permanência, bônus para tempo de fidelização ou associação de serviços, e uma gama de possíveis combos de serviços ou de experiências diferenciadas (tickets de shows patrocinados pela empresa, participação em eventos exclusivos, brindes personalizados, degustação de serviços não lançados no mercado para testes de receptividade, etc).

Ações de melhoria também em levantamentos mais abrangente sobre os clientes de todos os planos ofertados, para aferir níveis de satisfação e seguir um diagnóstico capaz de propor soluções efetivas e angariar dados para gerar um perfil mais detalhado da base de clientes. Buscar atender às expectativa dos clientes é forma de conseguir diagnosticar os padrões relativos ao churn.

Evolução da modelagem

O setor de telecomunicações sempre sofre com taxas de cancelamento muito altas quando há diversos palyers ofertando planos melhores, há uma grande possibilidade de o cliente sair da atual empresa devido a uma gama de atrativos, esse cenário é muito difícil de prever e, consequentemente, de evitar perdas, mas através de modelagens de previsão, pode-se tentar manter a rotatividade em um nível aceitável, que não cause prejuízos que afetem o fluxo de receitas e mantenha a empresa em constante crescimento.

É satisfatória a performance preditiva do LRM com 84% de acurácia, obtendo corretamente praticamente 5 de cada 6 casos de churn verdadeiros, o que é ótimo. Porém, ainda incorre em elevado índice de falsos positivos, ou falsos negativos, que são riscos para implantação do modelo com confiança, é necessário melhorar a precisão.

É necessário promover ajustes e consequentes evoluções no modelo para mover o limite de decisão para uma probabilidade de churn que alcance precisão acima de 80%. Um enriquecimento da base de dados consiste em aumentar o databank por meio da adição e retificação de registros ausentes ou defasados. O processo inclui a correção, atualização, organização e higienização das bases, além do acréscimo de novas informações que ajuda a montar e entender o perfil de cada usuário. O enriquecimento da base com dados sobre os comportamentos dos clientes churners é ainda mais importante para o incremento do modelo, e para o aumento da precisão. Necessários diversos dados complementares, como pessoais (localização, profissão, idade, escolaridade); empresariais (tipo de emprego, tempo de permanência na empresa atual) e financeiros (renda mensal familiar, imóvel próprio, automóvel); sociais (redes sociais, perfil, amigos e familiares), afinidades e hobbies (atividades, gostos e preferências); comportamentais (interação na plataforma, cliques, pesquisas, cadastros e compartilhamentos) e afinidade com o produto/serviço (interação e engajamento).
Assim, conhecendo melhor o perfil dos clientes, o poder e os hábitos de consumo, suas preferências em cada segmento de produtos, a empresa pode ser muito mais assertiva na aproximação e relacionamento, facilitando os processos de predição e regulando constantemente os modelos de conquista, negociação, venda, upgrading e retenção.