Predict Customer Churn

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 65
##  $ 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 ...
```

The raw data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target. We'll use all other columns as features to our model.

We use sapply to check the number if missing values in each columns. We found that there are 11 missing values in "TotalCharges" columns. So, let's remove these rows with missing values.

```
##       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
```

Change "No internet service" to "No" for six columns, they are: "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "streamingTV", "streamingMovies".

Change "No phone service" to "No" for column "MultipleLines"

The minimum tenure is 1 month and maximum tenure is 72 months, we can group them into five tenure groups: "0–12 Month", "12–24 Month", "24–48 Months", "48–60 Month", "> 60 Month".

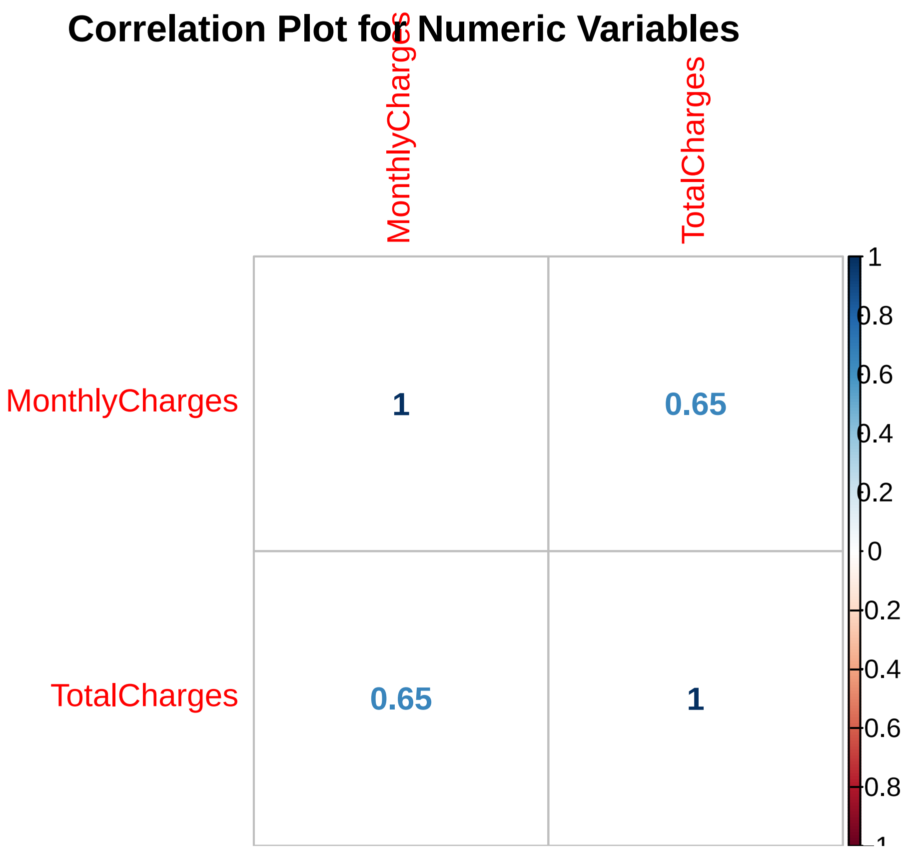
```
## [1] 1
```

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

Change the values in column "SeniorCitizen" from 0 or 1 to "No" or "Yes".

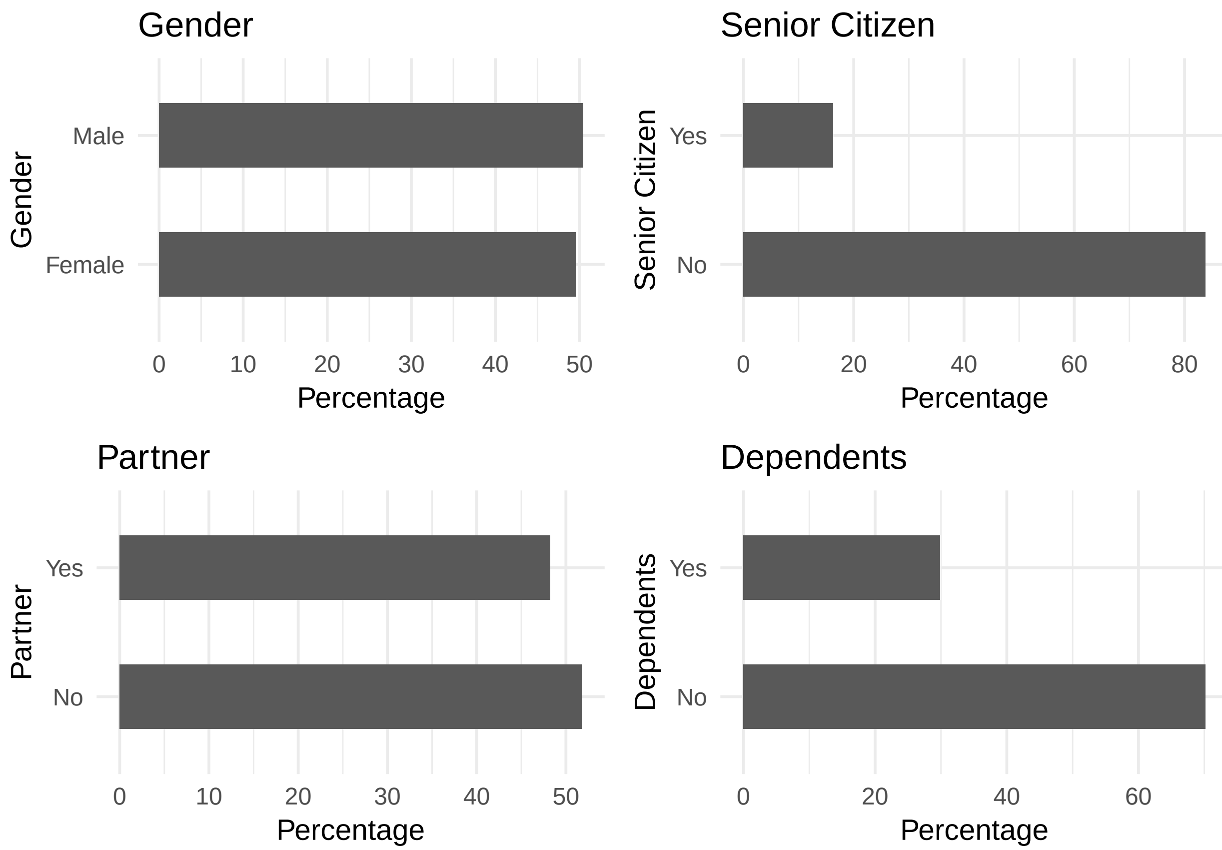
Remove the columns we do not need for the analysis:

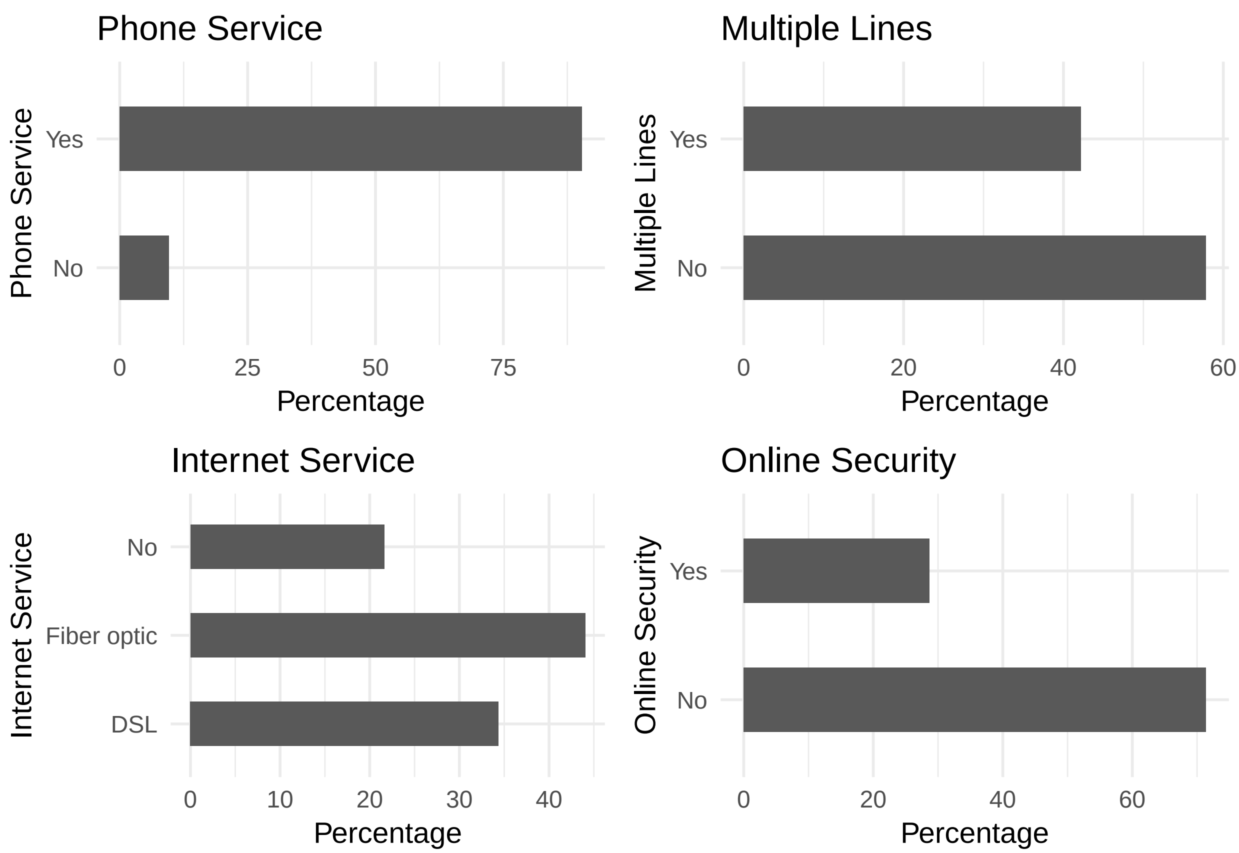
Exploratory data analysis and feature selection

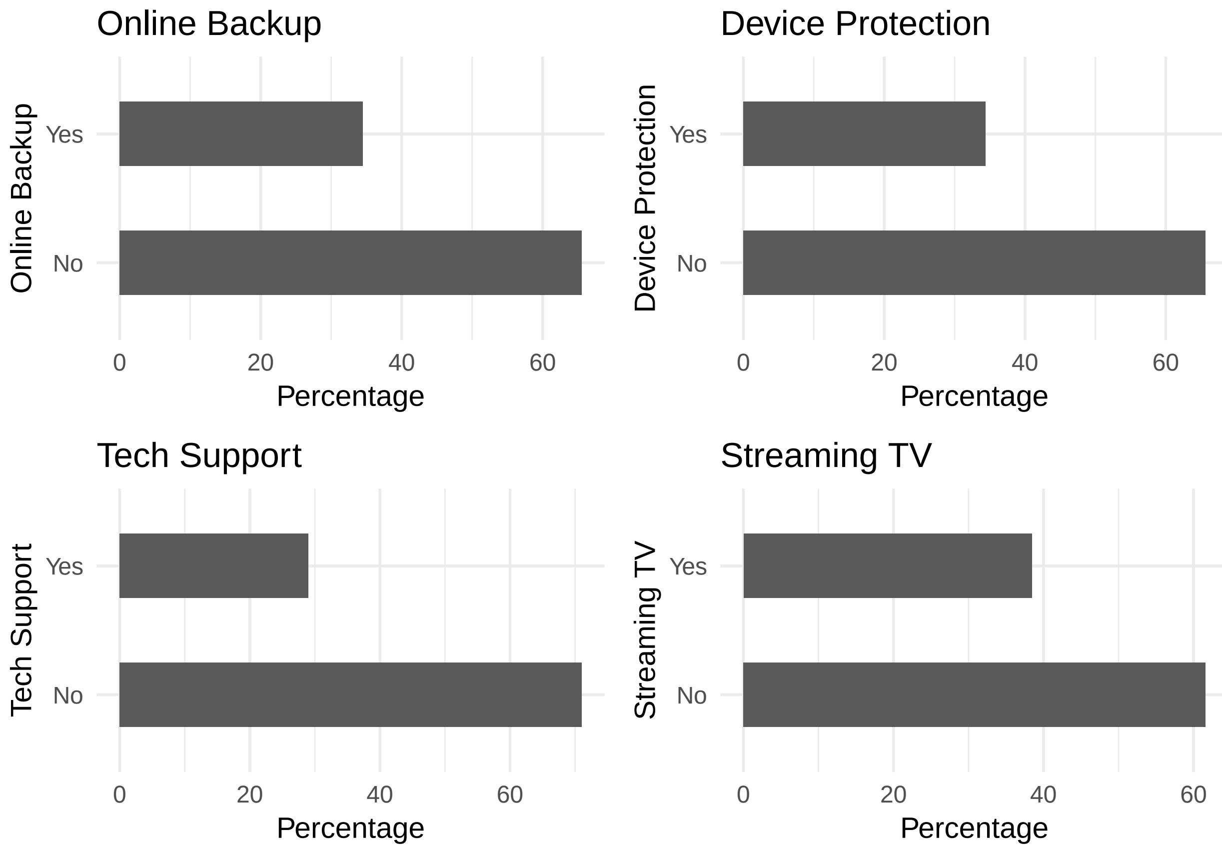


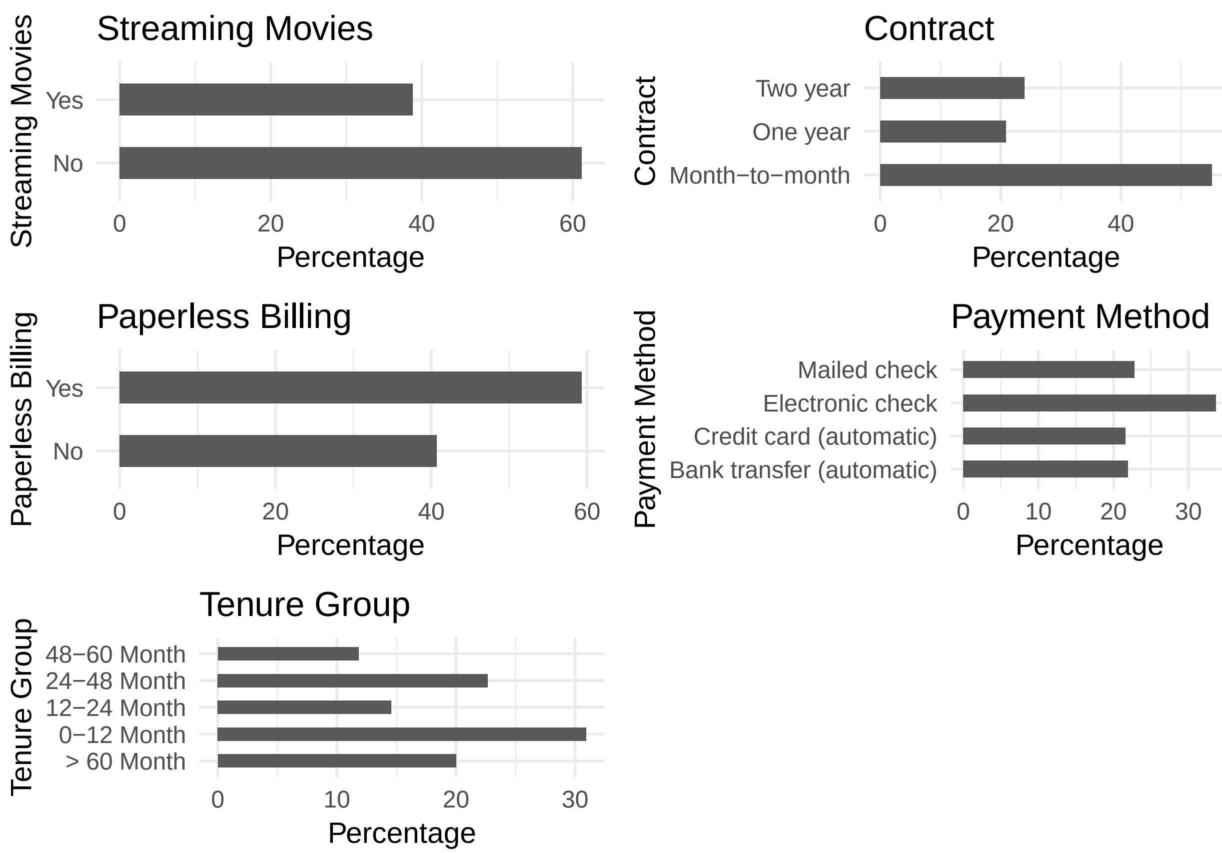
The Monthly Charges and Total Charges are correlated. So one of them will be removed from the model. We remove Total Charges.

Bar plots of categorical variables




Phone Service
Yes
No
0
25
50
75
Percentage
Multiple Lines
Yes
No
0
20
40
60
Percentage
Internet Service
No
Fiber optic
DSL
0
10
20
30
40
Percentage
Online Security
Yes
No
0
20
40
60
Percentage


Online Backup
Yes
No
0
20
40
60
Percentage
Device Protection
Tech Support
Streaming TV



All categorical variables have a reasonable broad distribution, therefore, all of them will be kept for the further analysis.

Logistic Regression Model Fitting

Split the data into training and testing sets.

Confirm the splitting is correct.

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

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

Fitting the Model

```
##
## Call:
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9763  -0.6697  -0.3003   0.6818   3.0648
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -0.92146    0.97265  -0.947 0.343449
## genderMale                            -0.01447    0.07752  -0.187 0.851915
## SeniorCitizenYes                       0.19720    0.10078   1.957 0.050388
```

```
## PartnerYes                          -0.03607    0.09270  -0.389 0.697218
## DependentsYes                       -0.19705    0.10822  -1.821 0.068645
## PhoneServiceYes                      0.45148    0.76932   0.587 0.557297
## MultipleLinesYes                     0.47580    0.20990   2.267 0.023405
## InternetServiceFiber optic           2.08979    0.94669   2.207 0.027281
## InternetServiceNo                   -2.06384    0.95521  -2.161 0.030725
## OnlineSecurityYes                   -0.15728    0.21276  -0.739 0.459756
## OnlineBackupYes                      0.07744    0.20896   0.371 0.710952
## DeviceProtectionYes                  0.21081    0.20784   1.014 0.310458
## TechSupportYes                      -0.17330    0.21408  -0.809 0.418236
## StreamingTVYes                       0.64412    0.38844   1.658 0.097274
## StreamingMoviesYes                   0.75741    0.38720   1.956 0.050452
## ContractOne year                    -0.64979    0.12707  -5.114 3.16e-07
## ContractTwo year                    -1.38085    0.21182  -6.519 7.07e-11
## PaperlessBillingYes                  0.35574    0.08921   3.988 6.67e-05
## PaymentMethodCredit card (automatic) -0.13976    0.13611  -1.027 0.304515
## PaymentMethodElectronic check        0.20280    0.11314   1.793 0.073048
## PaymentMethodMailed check           -0.06792    0.13714  -0.495 0.620424
## MonthlyCharges                      -0.04575    0.03761  -1.217 0.223754
## tenure_group0-12 Month               1.90090    0.20505   9.270  < 2e-16
## tenure_group12-24 Month              0.98695    0.19989   4.938 7.91e-07
## tenure_group24-48 Month              0.66157    0.18431   3.589 0.000331
## tenure_group48-60 Month              0.34234    0.19972   1.714 0.086506
##
## (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: 4112.2  on 4898  degrees of freedom
## AIC: 4164.2
##
## Number of Fisher Scoring iterations: 6
```

Feature analysis:

1. The top three most-relevant features include Contract, Paperless Billing and tenure group, all of which are categorical variables.

```
## 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     0.00      4922     5702.8   0.98079
## SeniorCitizen     1   100.28      4921     5602.5 < 2.2e-16 ***
## Partner           1   120.42      4920     5482.1 < 2.2e-16 ***
## Dependents        1    33.24      4919     5448.8 8.164e-09 ***
## PhoneService      1     1.36      4918     5447.5   0.24304
## MultipleLines     1     4.08      4917     5443.4   0.04336 *
## InternetService   2   506.77      4915     4936.6 < 2.2e-16 ***
## OnlineSecurity    1   168.76      4914     4767.9 < 2.2e-16 ***
## OnlineBackup      1    75.92      4913     4691.9 < 2.2e-16 ***
## DeviceProtection  1    41.93      4912     4650.0 9.460e-11 ***
## TechSupport       1    84.58      4911     4565.4 < 2.2e-16 ***
## StreamingTV       1     0.47      4910     4565.0   0.49444
## StreamingMovies   1     1.37      4909     4563.6   0.24125
## Contract          2   245.85      4907     4317.7 < 2.2e-16 ***
## PaperlessBilling  1    15.40      4906     4302.3 8.680e-05 ***
## PaymentMethod     3    24.88      4903     4277.4 1.634e-05 ***
## MonthlyCharges    1     1.30      4902     4276.1   0.25351
## tenure_group      4   163.95      4898     4112.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analyzing the deviance table we can see the drop in deviance when adding each variable one at a time. Adding InternetService, Contract and tenure_group significantly reduces the residual deviance. The other variables such as PaymentMethod and Dependents seem to improve the model less even though they all have low p-values.

Assessing the predictive ability of the model

```
## [1] "Logistic Regression Accuracy 0.801707779886148"
```

Confusion Matrix

```
## [1] "Confusion Matrix for Logistic Regression"
##
##     FALSE TRUE
##   0  1417  131
##   1   287  273
```

Odds Ratio

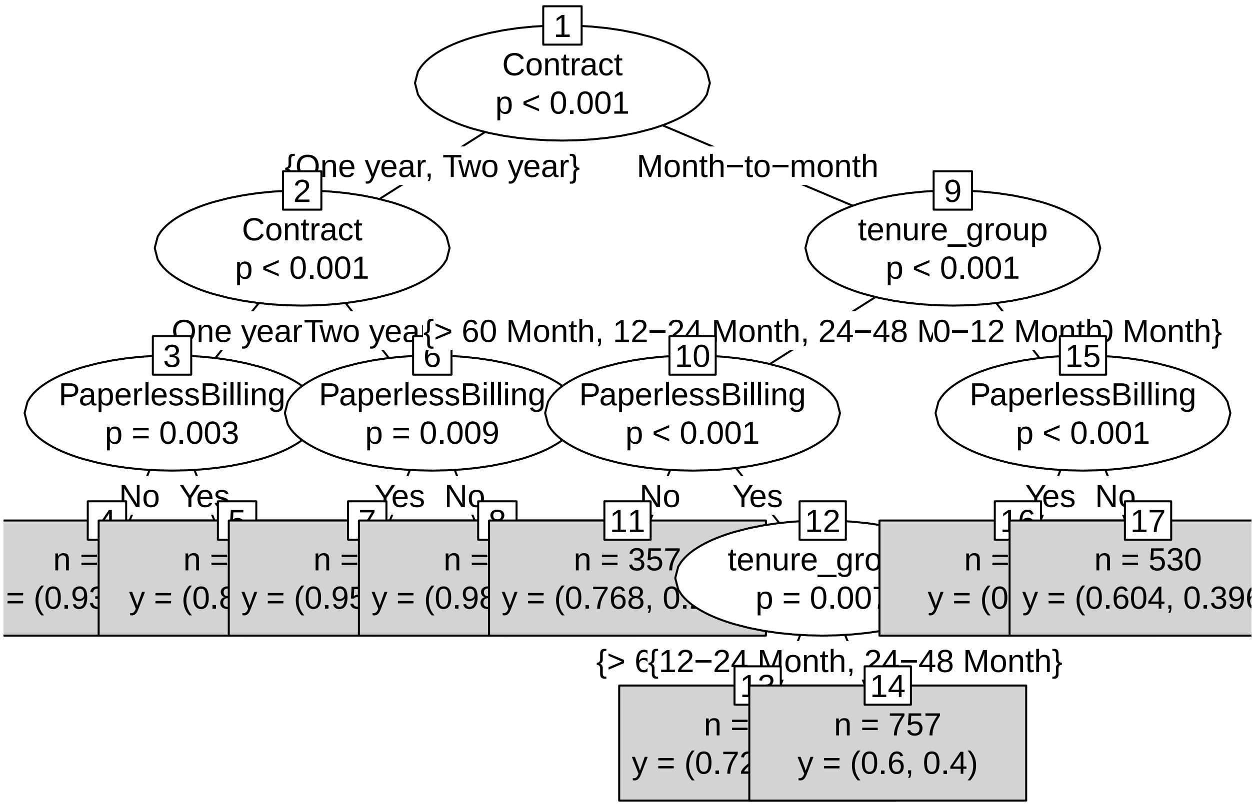
One of the interesting perfomance measurements in logistic regression is Odds Ratio.Basically, Odds retios is what the odds of an event is happening?

```
##                                             OR      2.5 %     97.5 %
## (Intercept)                          0.3979360 0.05910698  2.6791260
## genderMale                           0.9856330 0.84666279  1.1473804
## SeniorCitizenYes                     1.2179860 0.99940452  1.4837530
## PartnerYes                           0.9645762 0.80436773  1.1569171
## DependentsYes                        0.8211506 0.66360551  1.0144251
## PhoneServiceYes                      1.5706371 0.34808974  7.1082463
## MultipleLinesYes                     1.6093037 1.06707750  2.4302103
## InternetServiceFiber optic           8.0832486 1.26797785 51.9089048
## InternetServiceNo                    0.1269651 0.01946711  0.8240131
## OnlineSecurityYes                    0.8544644 0.56292905  1.2964659
## OnlineBackupYes                      1.0805130 0.71745322  1.6279183
## DeviceProtectionYes                  1.2346762 0.82169810  1.8563314
## TechSupportYes                       0.8408890 0.55247563  1.2790204
## StreamingTVYes                       1.9043149 0.89027141  4.0832766
## StreamingMoviesYes                   2.1327367 0.99960545  4.5624149
## ContractOne year                     0.5221562 0.40591083  0.6681851
## ContractTwo year                     0.2513658 0.16389674  0.3766105
## PaperlessBillingYes                  1.4272424 1.19869245  1.7006799
## PaymentMethodCredit card (automatic) 0.8695689 0.66562047  1.1351673
## PaymentMethodElectronic check        1.2248303 0.98182660  1.5301056
## PaymentMethodMailed check            0.9343384 0.71434440  1.2230806
## MonthlyCharges                       0.9552794 0.88728363  1.0282574
## tenure_group0-12 Month               6.6919206 4.49667334 10.0511642
## tenure_group12-24 Month              2.6830404 1.81966903  3.9862991
## tenure_group24-48 Month              1.9378364 1.35533779  2.7936011
## tenure_group48-60 Month              1.4082456 0.95272561  2.0864372
```

For each unit increase in Monthly Charge, there is a 2.4% decrease in the likelihood of a customer's churning.

Decision Tree

For illustration purpose, we are going to use only three variables, they are "Contract", "tenure_group" and "PaperlessBilling".



Out of three variables we use, Contract is the most important variable to predict customer churn or not churn.

If a customer in a one-year contract and not using PapelessBilling, then this customer is unlikely to churn.

On the other hand, if a customer is in a month-to-month contract, and in the tenure group of 0-12 months, and using PaperlessBilling, then this customer is more likely to churn.

```
## [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:
```

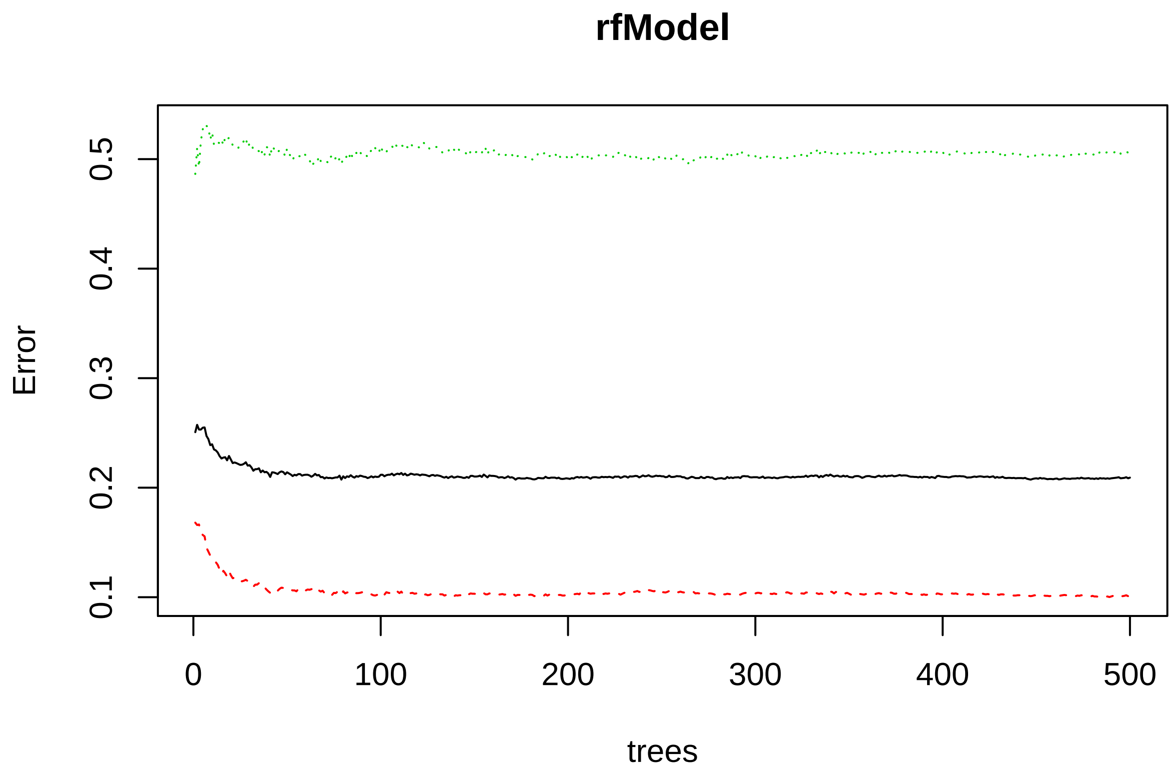
```
##       No Yes class.error
## No  3247 368   0.1017981
## Yes  662 647   0.5057296
```

Prediction is pretty good when predicting "No". Error rate is much higher when predicting "Yes".

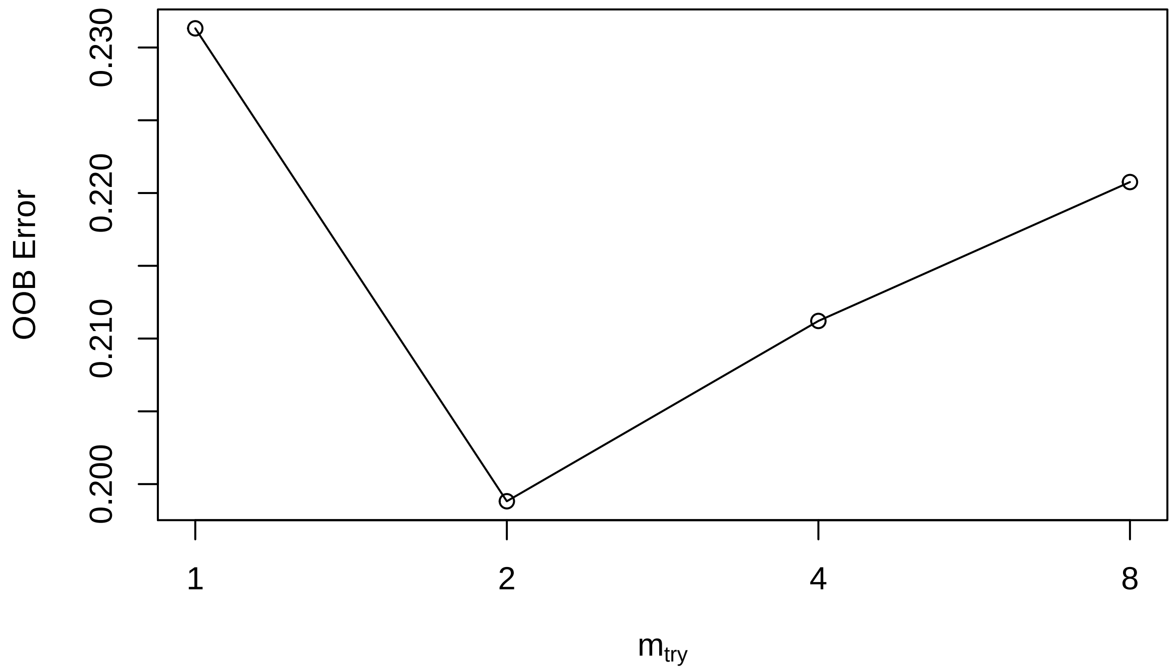
Prediction and confusion matrix

```
## 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
##
```

Error rate for Random Forest Model




```
## 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
```



Fit the Random Forest Model again

```
##
## Call:
##  randomForest(formula = Churn ~ ., data = training, ntree = 200,      mtry = 2, importance = TRUE, p
##                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
```

Make Predictions and Confusion Matrix again

```
## 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