Advance Stats I Project: Predict Customer Churn in the Telco Industry

Luke Philip Ogweno

Hong Shi

Introduction

- Customer retention is one of the primary growth pillars for products with a subscription-based business model.
- Competition is tough in the SaaS market where customers are free to choose from plenty of providers even within one product category.
- Several bad experiences – or even one – and a customer may quit.
- When analyzing customer data from a company many interesting patterns can be observed and further analysis can lead to predictive models for various metrics.
  - Customer churn.
  - Monthly payments.
- Usually, customers want to get quality service for the best possible price. If they don't get it, they may choose another service provider.

What is customer churn?

- Customer churn (or customer attrition) is the tendency of customers to abandon a brand and stop paying clients of a particular business.
- The percentage of customers that discontinue using a company's products or services during a particular time period is called a customer churn (attrition) rate.
- Churn rate is a health indicator for businesses whose customers are subscribers and paying for services on a recurring basis, notes the head of the data analytics department at Science Soft Alex Bekker.

Case for customer churn prediction

MUSIC AND VIDEO STREAMING SERVICES

MEDIA

TELECOM COMPANIES (CABLE OR WIRELESS)

SOFTWARE AS A SERVICE PROVIDERS

- **Music and video streaming services** are probably the most commonly associated with the subscription business model (Netflix, YouTube, Apple Music, Google Play, Spotify, Hulu, Amazon Video, Deezer, etc.).
- **Media**. Digital presence is a must among the press, so news companies offer readers digital subscriptions besides print ones (Bloomberg, The Guardian, Financial Times, The New York Times, Medium etc.).
- **Telecom companies (cable or wireles**s). These companies may provide a full range of products and services, including wireless network, internet, TV, cell phone, and home phone services (AT&T, Sprint, Verizon, Cox Communications, etc.). Some specialize in mobile telecommunications (China Mobile, Vodafone, T-Mobile, etc.).
- **Software as a service providers.** The adoption of cloud-hosted software is growing. According to Gartner, the SaaS market remains the largest segment of the cloud market. Its revenue is expected to grow 17.8 percent and reach $85.1 billion in 2019. The product range of SaaS providers is extensive: graphic and video editing (Adobe Creative Cloud, Canva), accounting (Sage 50cloud, FreshBooks), eCommerce (BigCommerce, Shopify), email marketing (MailChimp, Zoho Campaigns), and many others.

These company types may use churn rate to measure the effectiveness of cross-department operations and product management.

Methodology

Understanding a problem and the final goal: **In our case is the churn prediction**

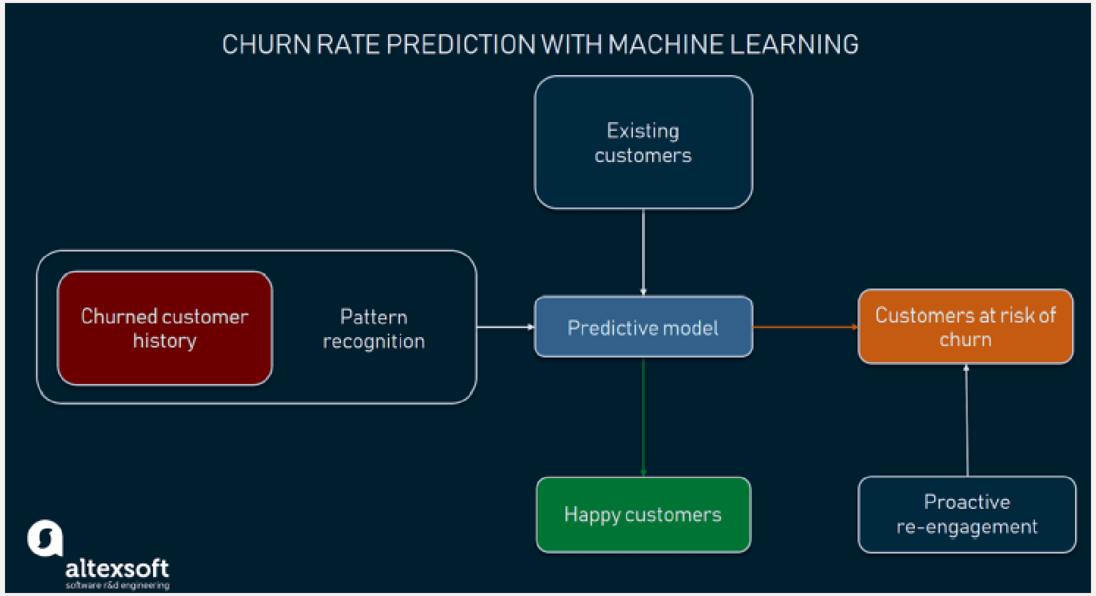
Data collection

Data preparation and preprocessing

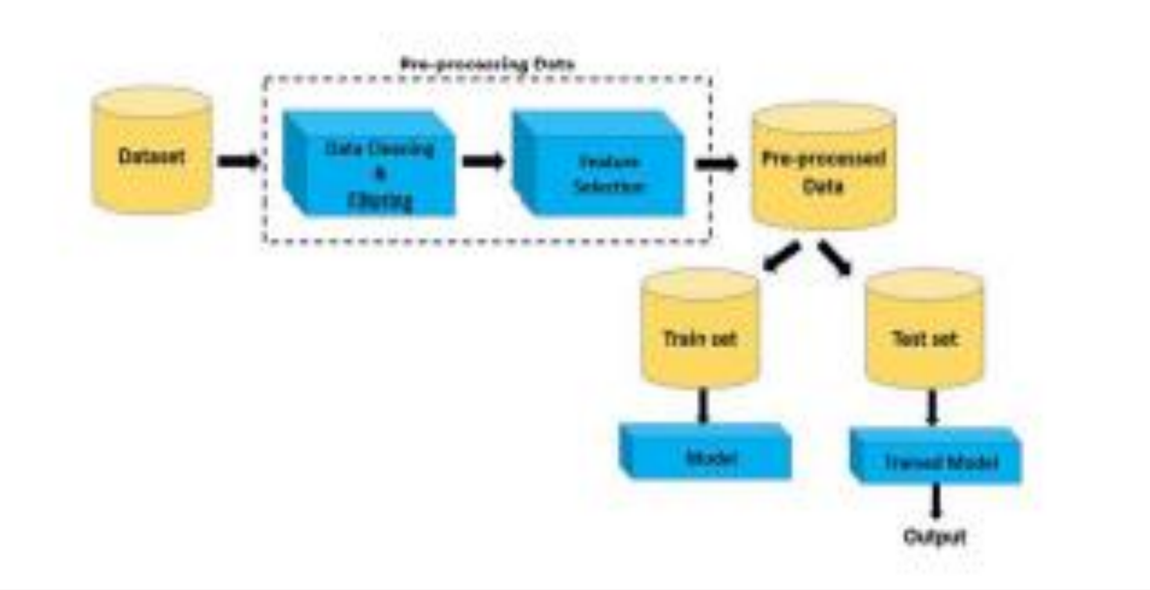
Modeling and testing

Model deployment and monitoring

Churn rate predictive model



Exploratory Data Analysis



Statistical Learning Methods:

DECISION TREE

RANDOM FOREST ANALYSIS

LOGISTIC REGRESSION

Goal of this project

- ❑ This project aims to predict churn behaviors or not churn behaviors to help retain customers

Data Overview

The data was downloaded from IBM Sample Data Sets for customer retention programs. (IBM Sample Data Sets).

Each row represents a customer, each column contains a customer's attribute.

Customers who left within the last month – the column is called Churn.

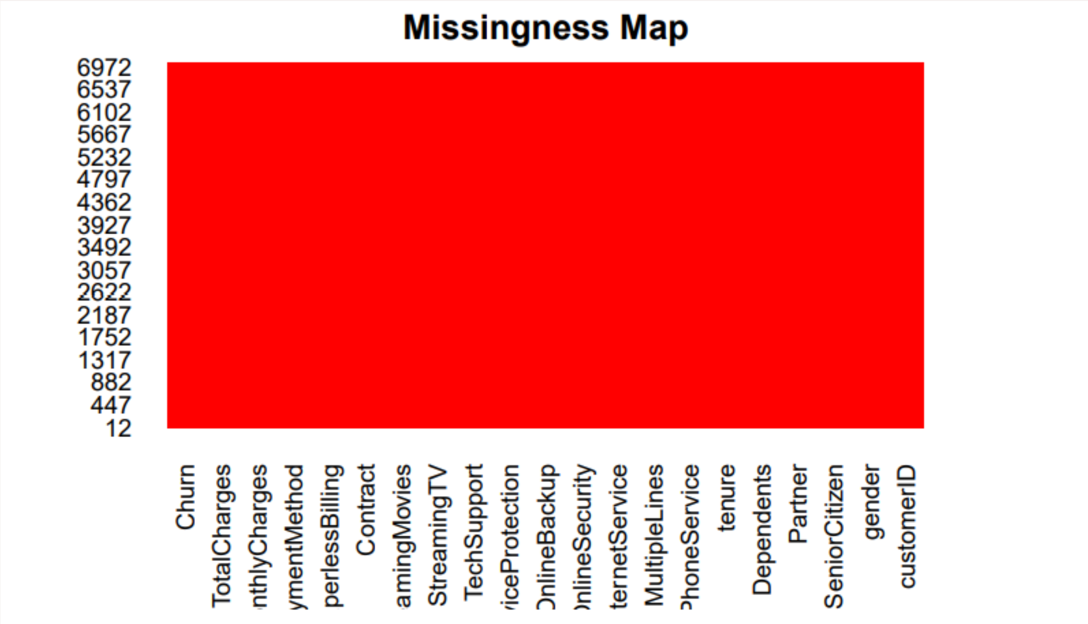
Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.

Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
Demographic info about customers – gender, age range, and if they have partners and dependents.

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

```
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr  "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr  "Yes" "No" "No" "No" ...
 $ Dependents      : chr  "No" "No" "No" "No" ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
 $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
 $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
 $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
 $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
 $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
 $ TechSupport     : chr  "No" "No" "No" "Yes" ...
 $ StreamingTV     : chr  "No" "No" "No" "No" ...
 $ StreamingMovies : chr  "No" "No" "No" "No" ...
 $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
 $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
 $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (a
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : chr  "No" "No" "Yes" "No" ...
```

Exploration and Data Analysis (EDA)




```r
{r}
sapply(churn, function(x) sum(is.na(x)))
```

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

❑ No missing data in this dataset!

❑ Missing values in each column

Data wrangling

Changed "No internet service" to "No" for six columns, they are: "OnlineSecurity",

"OnlineBackup", "DeviceProtection", "TechSupport",

"streaming tv",

"streaming movies".

Change "No phone service" to "No" for column "multiple lines".

Grouping Tenure: Since the minimum tenure is 1 month and maximum tenure is 72 months, we grouped them into five tenure groups:

"0-12 Months",

"12-24 Months",

"24-48 Months",

"48-60 Month",

"> 60 Month".

. Changed the values in column "SeniorCitizen"

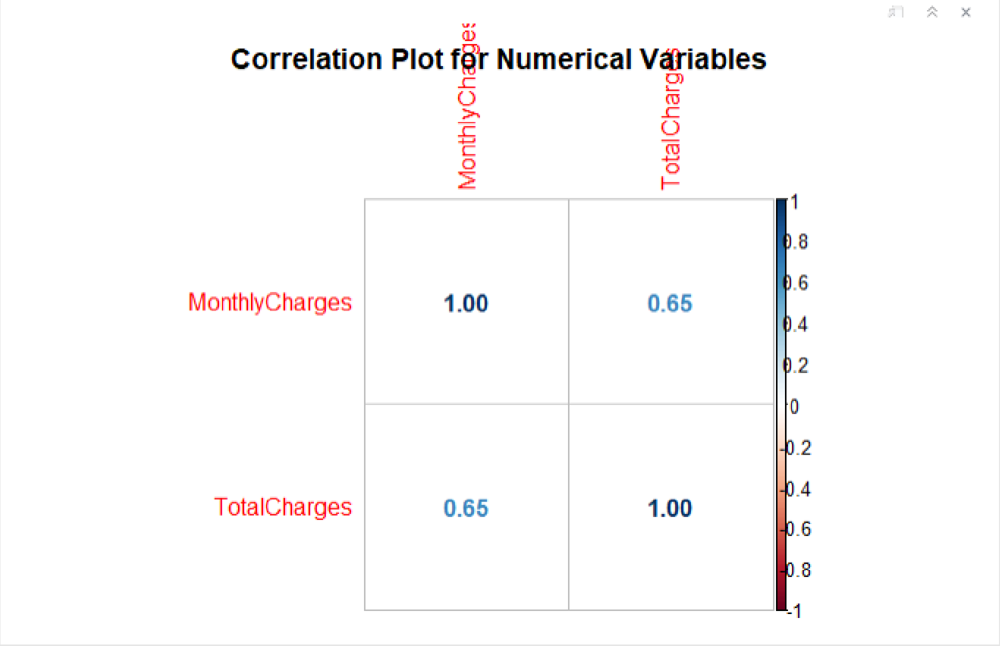
from 0 or 1 to "No" or "Yes".

Removed the columns we do not need for the analysis.

Exploratory data analysis and feature selection

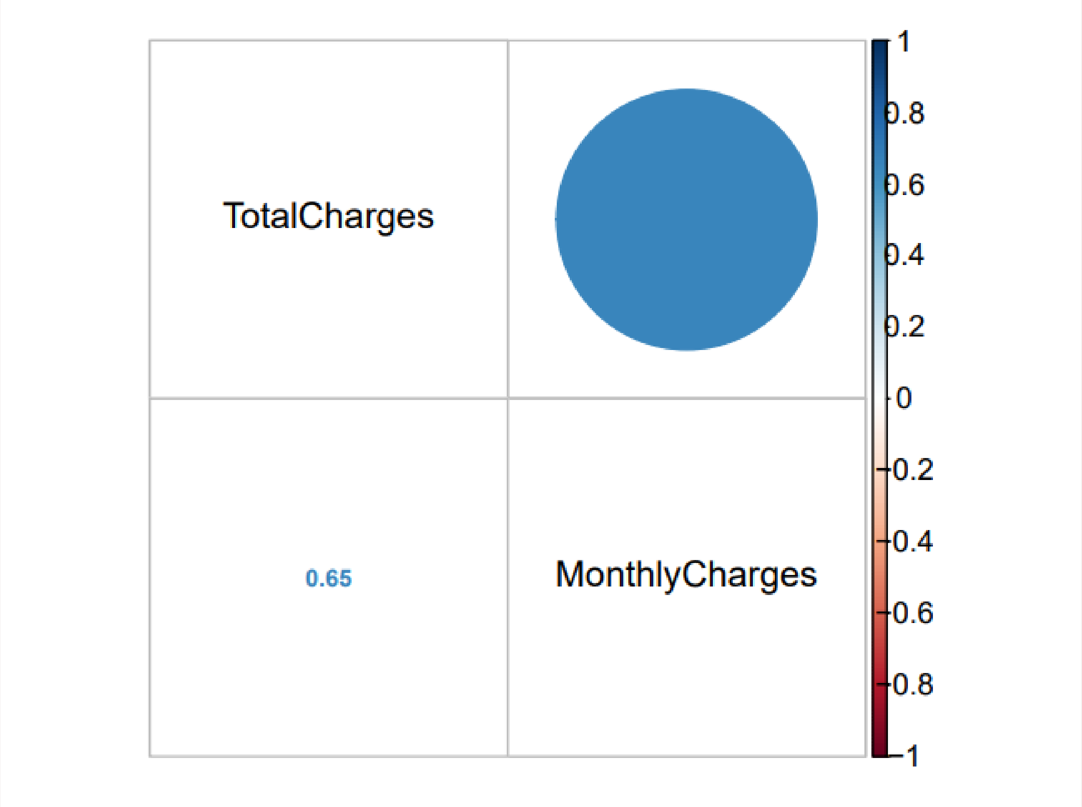
Correlation between numeric variables

```
{r}
numeric.var <- sapply(churn, is.numeric)
corr.matrix <- cor(churn[,numeric.var])
corrplot(corr.matrix, main="\n\nCorrelation Plot for Numerical Variables", method
="number")
```



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

- The plot shows high correlations between Totalcharges and tenure and between TotalCharges and MonthlyCharges. Pay attention to these variables while training models later.
- Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set. But it affects calculations regarding individual predictors.
- The Monthly Charges and Total Charges are correlated. So, one of them will be removed from the model.
- We removed TotalCharges.

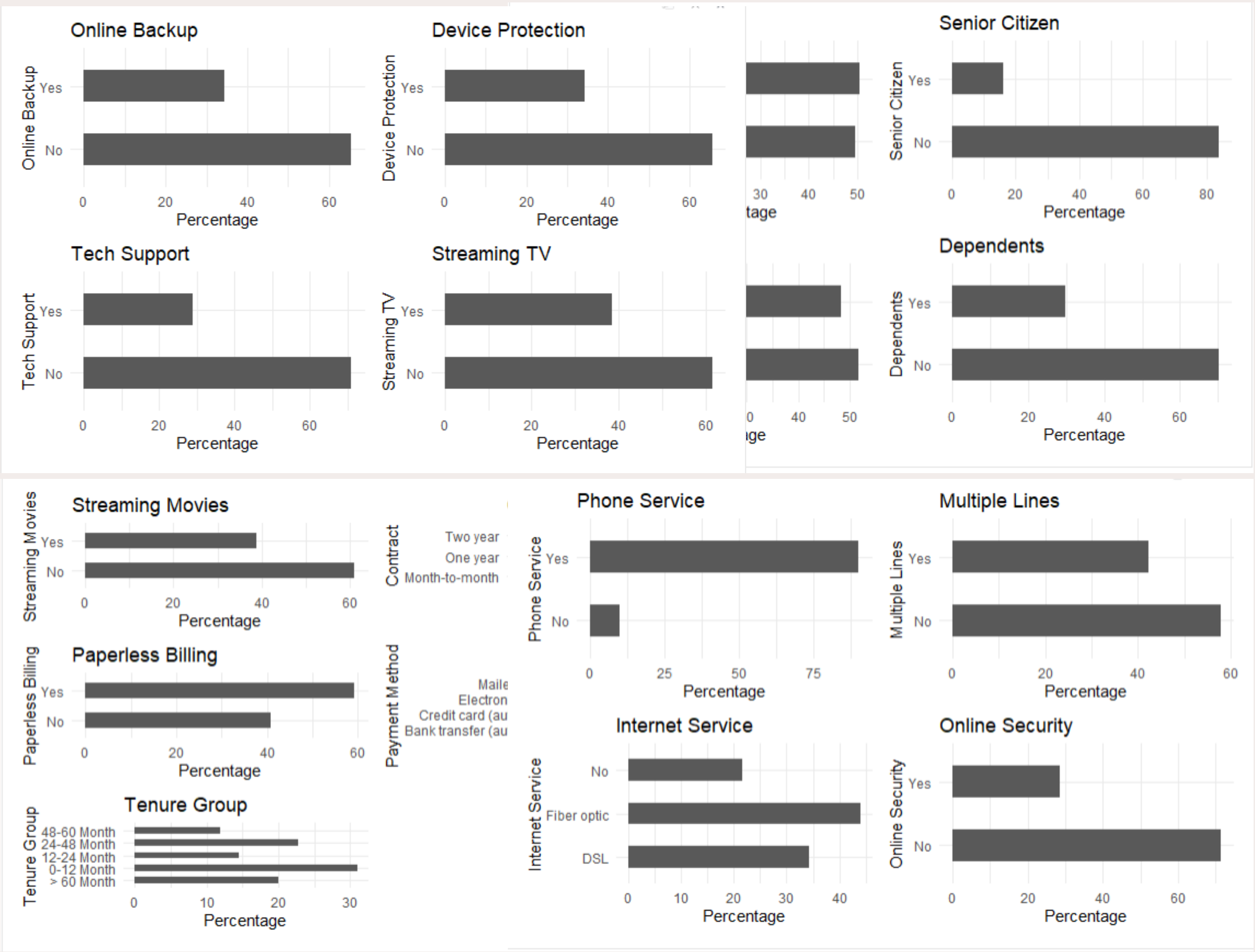


Continuous Variables

- The number of current customers with MonthlyCharges below $25 is extremely high.
- For the customers with Monthlycharges greater than $30, the distributions are similar between those who churned and who did not churn.

Bar plots of categorical variables

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



Logistic Regression

Logistic Regression

First, we split the data into training and testing sets

```
{r}
intrain<- createDataPartition(churn$Churn,p=0.7,list=FALSE)
set.seed(2022)
training<- churn[intrain,]
testing<- churn[-intrain,]
```

Check out the results if correct

```
{r}
dim(training); dim(testing)
```

```
[1] 4924   19
[1] 2108   19
```

```
Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9615  -0.6592  -0.2808   0.6424   3.0950

Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -0.89604    0.99087  -0.904  0.36584
genderMale                           -0.04488    0.07824  -0.574  0.56623
SeniorCitizenYes                      0.22879    0.10249   2.232  0.02559 *
PartnerYes                           -0.08440    0.09329  -0.905  0.36562
DependentsYes                        -0.13605    0.10715  -1.270  0.20418
PhoneServiceYes                       0.31467    0.77908   0.404  0.68629
MultipleLinesYes                      0.50886    0.21235   2.396  0.01656 *
InternetServiceFiber optic            1.98387    0.95974   2.067  0.03873 *
InternetServiceNo                    -1.92803    0.96638  -1.995  0.04603 *
OnlineSecurityYes                    -0.19618    0.21371  -0.918  0.35863
OnlineBackupYes                       0.05084    0.21096   0.241  0.80954
DeviceProtectionYes                   0.14409    0.21180   0.680  0.49632
TechSupportYes                       -0.20182    0.21740  -0.928  0.35322
StreamingTVYes                        0.69892    0.39383   1.775  0.07595 .
StreamingMoviesYes                    0.61696    0.39158   1.576  0.11512
ContractOne year                     -0.79030    0.12886  -6.133 8.62e-10 ***
ContractTwo year                     -1.56447    0.21480  -7.284 3.25e-13 ***
PaperlessBillingYes                   0.32496    0.08966   3.624  0.00029 ***
PaymentMethodCredit card (automatic) -0.14541    0.13521  -1.075  0.28219
PaymentMethodElectronic check         0.18934    0.11315   1.673  0.09425 .
PaymentMethodMailed check            -0.02470    0.13719  -0.180  0.85710
MonthlyCharges                       -0.04015    0.03810  -1.054  0.29195
tenure_group0-12 Month                1.84776    0.20462   9.030  < 2e-16 ***
tenure_group12-24 Month               0.82522    0.20164   4.093 4.27e-05 ***
tenure_group24-48 Month               0.56704    0.18273   3.103  0.00191 **
tenure_group48-60 Month               0.28815    0.19896   1.448  0.14755
---
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: 4045.6  on 4898  degrees of freedom
AIC: 4097.6

Number of Fisher Scoring iterations: 6
```

Feature Analysis

```
{r}
anova(LogModel, 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     0.80      4922     5702.0 0.3705104
SeniorCitizen     1   108.09      4921     5593.9 < 2.2e-16 ***
Partner           1   124.61      4920     5469.2 < 2.2e-16 ***
Dependents        1    30.82      4919     5438.4 2.831e-08 ***
PhoneService      1     0.93      4918     5437.5 0.3352669
MultipleLines     1     5.25      4917     5432.2 0.0219247 *
InternetService   2   496.38      4915     4935.9 < 2.2e-16 ***
OnlineSecurity    1   179.01      4914     4756.9 < 2.2e-16 ***
OnlineBackup      1    76.56      4913     4680.3 < 2.2e-16 ***
DeviceProtection  1    50.54      4912     4629.8 1.167e-12 ***
TechSupport       1    87.64      4911     4542.1 < 2.2e-16 ***
StreamingTV       1     0.44      4910     4541.7 0.5082631
StreamingMovies   1     0.01      4909     4541.7 0.9237943
Contract          2   285.27      4907     4256.4 < 2.2e-16 ***
PaperlessBilling  1    12.28      4906     4244.1 0.0004576 ***
PaymentMethod     3    27.20      4903     4216.9 5.345e-06 ***
MonthlyCharges    1     1.89      4902     4215.0 0.1696467
tenure_group      4   169.41      4898     4045.6 < 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.

Odds Ratio

One of the interesting performance measurements in logistic regression is Odds Ratio.Basically, Odds ratio is what the odds of an event is happening.

```
{r}
exp(cbind(OR=coef(LogModel), confint(LogModel)))
```

```
Waiting for profiling to be done...
                                           OR      2.5 %     97.5 %
(Intercept)                         0.4081848 0.05843796  2.8449400
genderMale                          0.9561102 0.82012875  1.1145815
SeniorCitizenYes                    1.2570751 1.02808565  1.5365642
PartnerYes                          0.9190590 0.76545831  1.1035375
DependentsYes                       0.8727981 0.70695778  1.0761487
PhoneServiceYes                     1.3698005 0.29752252  6.3128509
MultipleLinesYes                    1.6633934 1.09748393  2.5235833
InternetServiceFiber optic          7.2708502 1.11057926 47.8532201
InternetServiceNo                   0.1454344 0.02184231  0.9659622
OnlineSecurityYes                   0.8218634 0.54033958  1.2491036
OnlineBackupYes                     1.0521596 0.69579015  1.5911833
DeviceProtectionYes                 1.1549879 0.76262017  1.7498228
TechSupportYes                      0.8172392 0.53338968  1.2510047
StreamingTVYes                      2.0115860 0.93027585  4.3578916
StreamingMoviesYes                  1.8532883 0.86076872  3.9967881
ContractOne year                    0.4537089 0.35142261  0.5825610
ContractTwo year                    0.2091979 0.13553485  0.3151113
PaperlessBillingYes                 1.3839794 1.16127359  1.6505242
PaymentMethodCredit card (automatic) 0.8646670 0.66303926  1.1267861
PaymentMethodElectronic check       1.2084551 0.96859200  1.5095377
PaymentMethodMailed check           0.9756001 0.74582216  1.2772428
MonthlyCharges                      0.9606414 0.89143729  1.0350842
tenure_group0-12 Month              6.3456204 4.26573995  9.5188296
tenure_group12-24 Month             2.2823815 1.54170235  3.4005803
tenure_group24-48 Month             1.7630372 1.23645724  2.5327105
tenure_group48-60 Month             1.3339511 0.90344611  1.9725979
```

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

```
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -1.672173   0.368645  -4.536 5.73e-06 ***
SeniorCitizenYes                     0.308140   0.100146   3.077 0.002092 **
DependentsYes                       -0.174041   0.096333  -1.807 0.070815 .
MultipleLinesYes                     0.284101   0.103896   2.734 0.006248 **
InternetServiceFiber optic           1.380453   0.227892   6.057 1.38e-09 ***
InternetServiceNo                   -1.366440   0.208605  -6.550 5.74e-11 ***
OnlineSecurityYes                   -0.260398   0.106183  -2.452 0.014192 *
TechSupportYes                      -0.208976   0.107566  -1.943 0.052043 .
StreamingTVYes                       0.484673   0.116841   4.148 3.35e-05 ***
StreamingMoviesYes                   0.464643   0.114093   4.072 4.65e-05 ***
ContractOne year                    -0.713688   0.126484  -5.643 1.68e-08 ***
ContractTwo year                    -1.514632   0.209122  -7.243 4.39e-13 ***
PaperlessBillingYes                  0.301924   0.088557   3.409 0.000651 ***
PaymentMethodCredit card (automatic) -0.065011  0.134630  -0.483 0.629176
PaymentMethodElectronic check        0.284192   0.111468   2.550 0.010787 *
PaymentMethodMailed check            0.030848   0.134309   0.230 0.818341
MonthlyCharges                      -0.018770   0.006568  -2.858 0.004266 **
tenure_group0-12 Month               1.817382   0.198418   9.159  < 2e-16 ***
tenure_group12-24 Month              0.956762   0.198048   4.831 1.36e-06 ***
tenure_group24-48 Month              0.644349   0.181826   3.544 0.000394 ***
tenure_group48-60 Month              0.263736   0.198697   1.327 0.184401
---
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: 4135.6  on 4903  degrees of freedom
AIC: 4177.6

Number of Fisher Scoring iterations: 6
```

Logistic Regression Confusion Matrix

```
{r}
print("Confusion Matrix for Logistic Regression"); table(testing$Churn, fitted
.results > 0.5)
```

```
[1] "Confusion Matrix for Logistic Regression"

    FALSE TRUE
  0  1405  143
  1   272  288
```

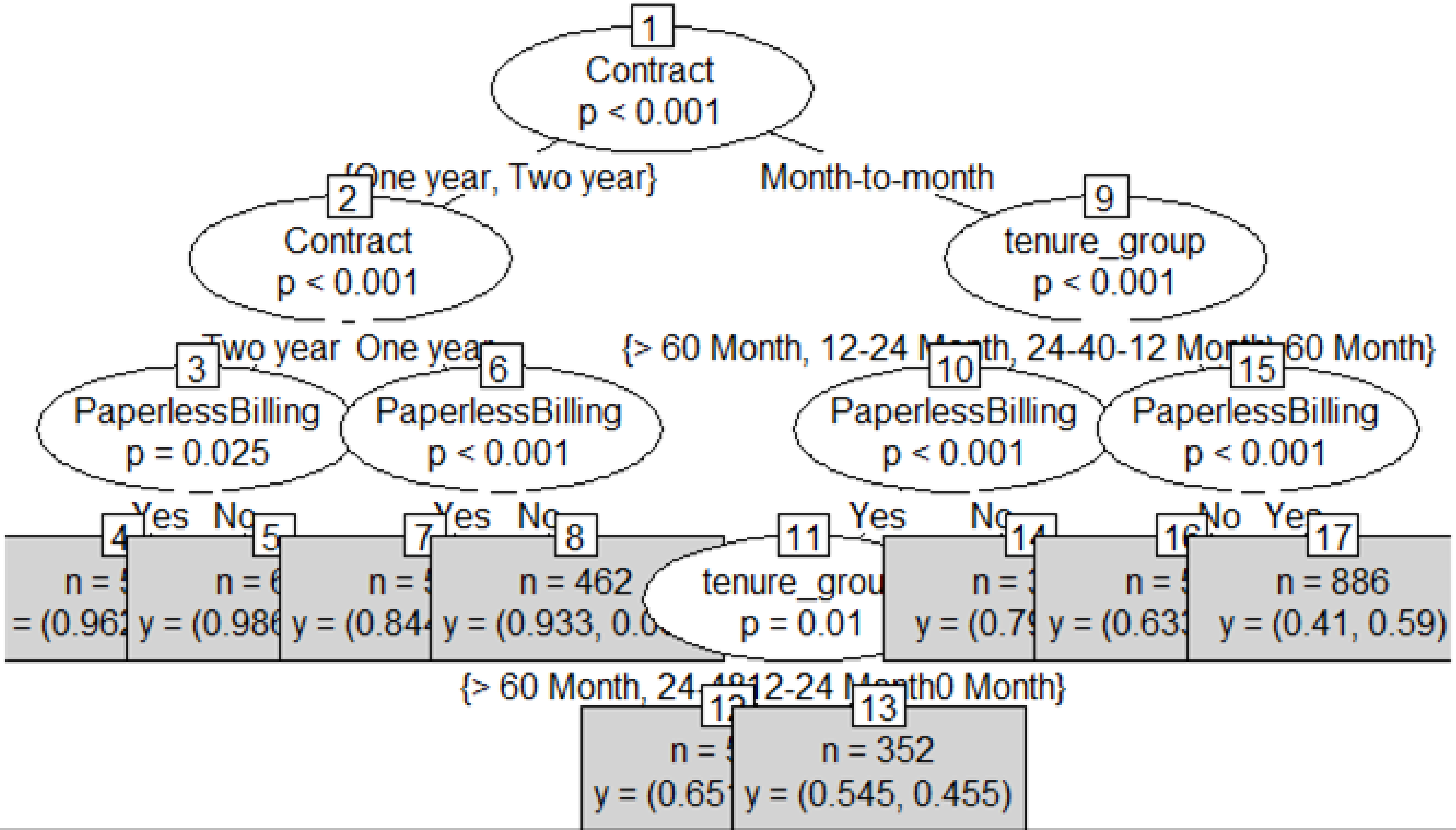
Assessing the predictive ability of the Logistic Regression model

```
{r}
testing$Churn <- as.character(testing$Churn)
testing$Churn[testing$Churn=="No"] <- "0"
testing$Churn[testing$Churn=="Yes"] <- "1"
fitted.results <- predict(LogModel,newdata=testing,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != testing$Churn)
print(paste('Logistic Regression Accuracy',1-misClasificError))
```

```
[1] "Logistic Regression Accuracy 0.803130929791271"
```

```
Call:
glm(formula = Churn ~ SeniorCitizen + Dependents + MultipleLines +
    InternetService + OnlineSecurity + TechSupport + StreamingTV +
    StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
    MonthlyCharges + tenure_group, family = binomial(link = "logit"),
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9293  -0.6763  -0.2949   0.6932   3.1120
```

Decision Tree



1
Contract
p < 0.001
{One year, Two year}
Month-to-month
2
Contract
p < 0.001
9
tenure_group
p < 0.001
Two year
One year
{> 60 Month, 12-24 Month, 24-4
0-12 Month
60 Month}
3
PaperlessBilling
p = 0.025
6
PaperlessBilling
p < 0.001
10
PaperlessBilling
p < 0.001
15
PaperlessBilling
p < 0.001
Yes
No
Yes
No
Yes
No
No
Yes
4
5
7
8
n = 462
y = (0.933, 0.0
11
tenure_grou
p = 0.01
14
16
17
n = 886
y = (0.41, 0.59)
{> 60 Month, 24-4
12-24 Month
0 Month}
12
13
n = 352
y = (0.545, 0.455)

Grouping:

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

Conclusion:

1. Out of the three variables we use, Contract is the most important variable to predict customer churn or not churn.
2. If a customer is in a one-year or two-year contract, no matter whether he (she) has PapelessBilling or not, he (she) is less likely to churn.
3. On the other hand, if a customer is in a month-to-month contract, in the tenure group of 0–12 months, and using PaperlessBilling, then this customer is more likely to churn.

❑ The accuracy of the Decision Tree has hardly improved. Let's see if we can do better using Random Forest.

Decision Tree Confusion Matrix

We are using all the variables to product confusion matrix table and make predictions.

```
{r}
pred_tree <- predict(tree, testing)
print("Confusion Matrix for Decision Tree"); table(Predicted = pred_tree, Actual =
testing$Churn)
```

```
[1] "Confusion Matrix for Decision Tree"
         Actual
Predicted   No  Yes
      No  1412  350
      Yes  136  210
```

Decision Tree Accuracy

```
{r}
p1 <- predict(tree, training)
tab1 <- table(Predicted = p1, Actual = training$Churn)
tab2 <- table(Predicted = pred_tree, Actual = testing$Churn)
print(paste('Decision Tree Accuracy',sum(diag(tab2))/sum(tab2)))
```

```
[1] "Decision Tree Accuracy 0.769449715370019"
```

Random Forest

Random Forest Initial Model

```r
{r}
set.seed(2017)
rfModel <- randomForest(Churn ~., data = training)
print(rfModel)
```

```
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: 21.02%
Confusion matrix:
      No Yes class.error
No  3244 371   0.1026279
Yes  664 645   0.5072574
```

- ❑ The error rate is relatively low when predicting “No”, and
- ❑ The error rate is much higher when predicting “Yes”.

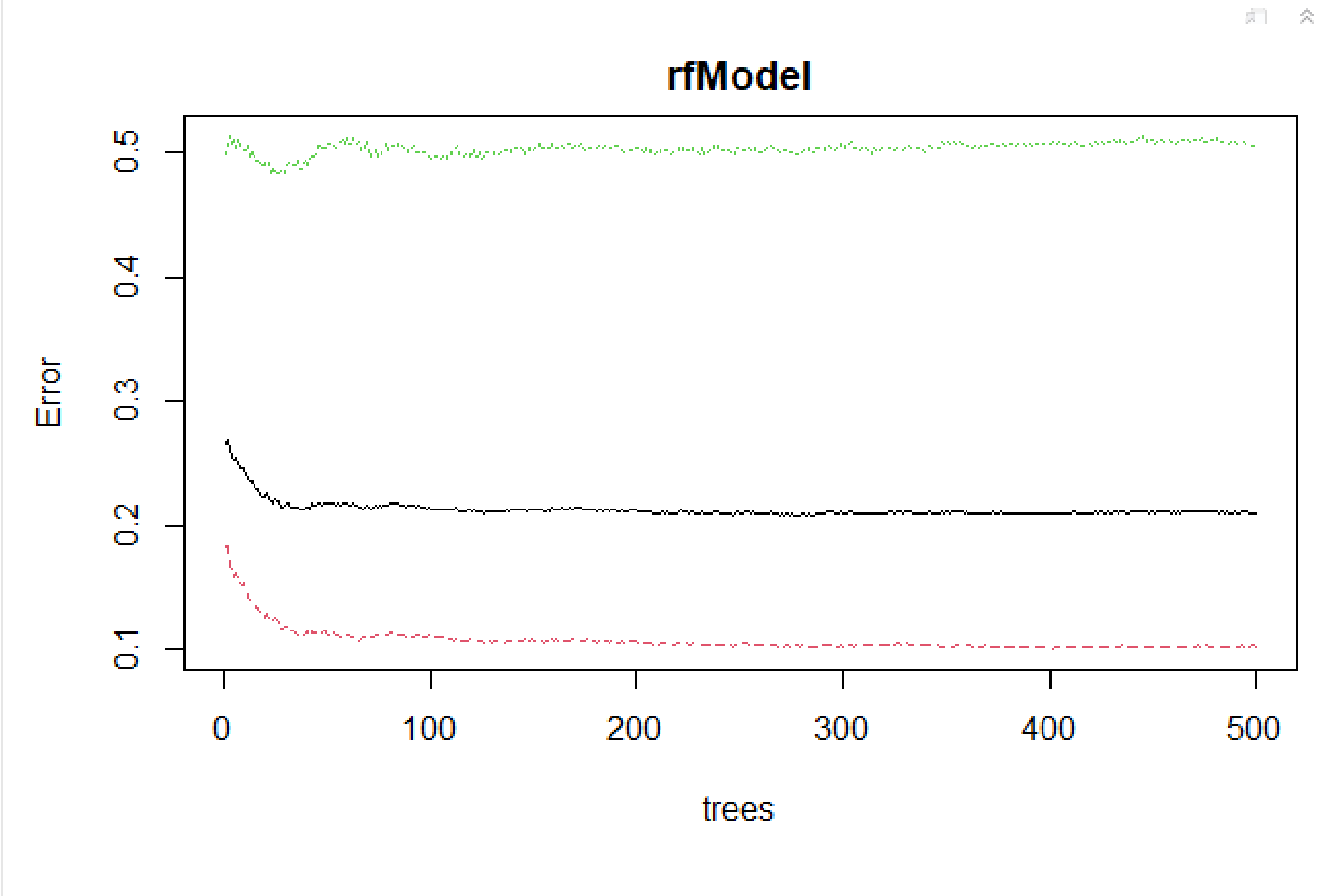
Random Forest Prediction and Confusion Matrix

```
{r}
pred_rf <- predict(rfModel, testing)
caret::confusionMatrix(pred_rf, testing$Churn)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1400  284
       Yes  148  276

               Accuracy : 0.7951
                 95% CI : (0.7772, 0.8121)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 5.339e-11

                  Kappa : 0.4306

 Mcnemar's Test P-Value : 8.293e-11

            Sensitivity : 0.9044
            Specificity : 0.4929
         Pos Pred Value : 0.8314
         Neg Pred Value : 0.6509
             Prevalence : 0.7343
         Detection Rate : 0.6641
   Detection Prevalence : 0.7989
      Balanced Accuracy : 0.6986

       'Positive' Class : No
```

Random Forest Error Rate

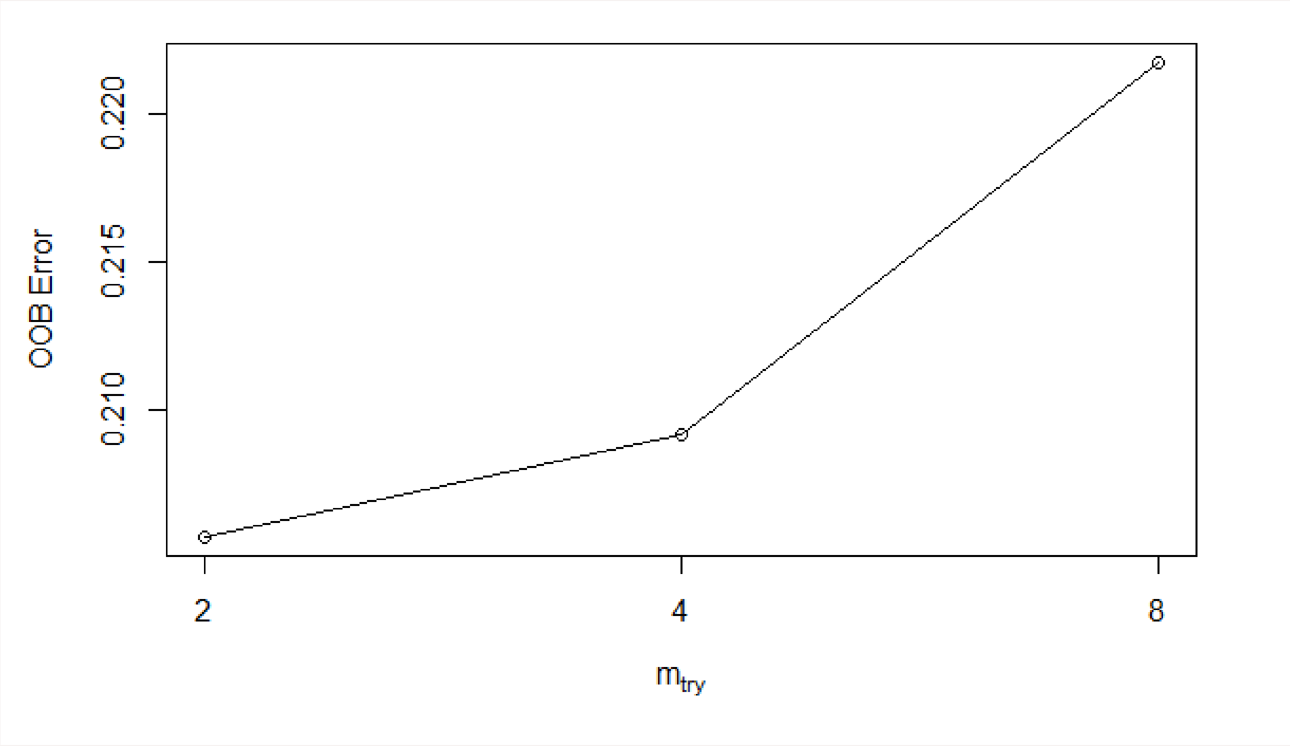
```r
{r}
plot(rfModel)
```



- We use this plot to help us determine the number of trees.
- As the number of trees increases, the OOB error rate decreases and then becomes almost constant.
- We are not able to decrease the OOB error rate after about 100 to 200 trees.

Tune Random Forest Model

```
mtry = 4  OOB error = 20.92%
Searching left ...
mtry = 8        OOB error = 22.18%
-0.06019417 0.05
Searching right ...
mtry = 2        OOB error = 20.57%
0.01650485 0.05
```



- ❑ We use this plot to give us some ideas on the number of mtry to choose.
- ❑ OOB error rate is at its lowest when mtry is 2. Therefore, we choose mtry=2.

Fit the Random Forest Model After Tuning

```
{r}
rfModel_new <- randomForest(Churn ~., data = training, ntree = 200, mtry = 2,
importance = TRUE, proximity = TRUE)
print(rfModel_new)
```

```
Call:
 randomForest(formula = Churn ~ ., data = training, ntree = 200,      mtry = 2,
importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 2

        OOB estimate of  error rate: 20.67%
Confusion matrix:
      No Yes class.error
No  3307 308  0.08520055
Yes  710 599  0.54239878
```

OOB error rate decreased to 19.7% from 20.65% earlier.

Random Forest Predictions and Confusion Matrix After Tuning

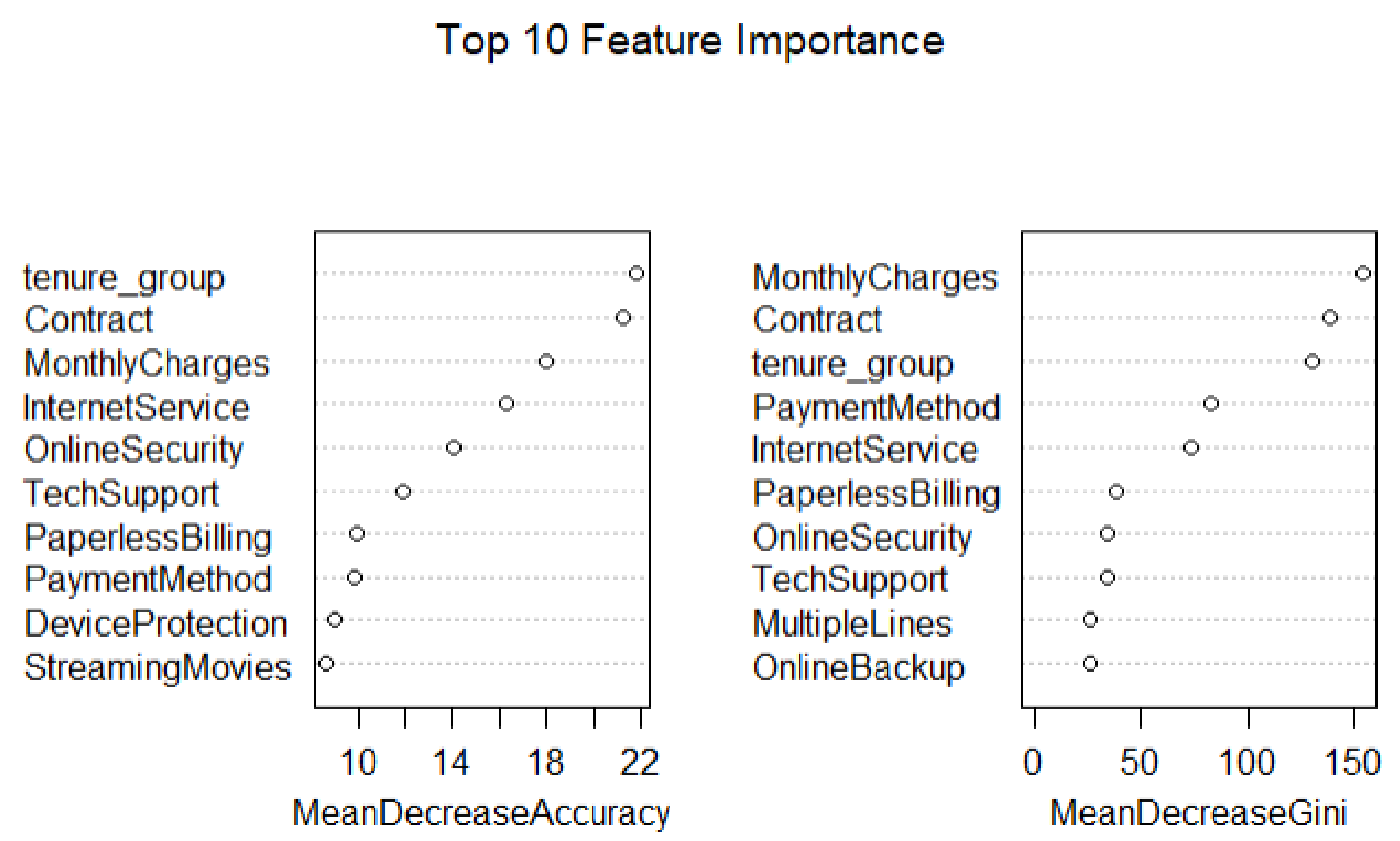
```
{r}
pred_rf_new <- predict(rfModel_new, testing)
caret::confusionMatrix(pred_rf_new, testing$Churn)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1420  301
       Yes  128  259

               Accuracy : 0.7965
                 95% CI : (0.7787, 0.8135)
    No Information Rate : 0.7343
    P-Value [Acc > NIR] : 1.872e-11

                  Kappa : 0.4214

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9173
            Specificity : 0.4625
         Pos Pred Value : 0.8251
         Neg Pred Value : 0.6693
             Prevalence : 0.7343
         Detection Rate : 0.6736
   Detection Prevalence : 0.8164
      Balanced Accuracy : 0.6899

       'Positive' Class : No
```

- The accuracy and sensitivity improved, compared with the initial Random Forest model.

Random Forest Feature Importance

```
{r}
varImpPlot(rfModel_new, sort=T, n.var = 10, main = 'Top 10 Feature Importance')
```



Summary

From the above example, we can see that Logistic Regression and Random Forest performed better than Decision Tree for customer churn analysis for this particular dataset.

Throughout the analysis, we have learned several important things:

- ❑ Feature such as tenure_group, Contract, PaperlessBilling, MonthlyCharges, and internet service appears to play a role in customer churn.
- ❑ There does not seem to be a relationship between gender and churn.
- ❑ Customers in a month-to-month contract, with PaperlessBilling and are within 12 months of tenure, are more likely to churn;
- ❑ On the other hand, customers with one or two-year contracts, with longer than 12 months tenure, that are not using PaperlessBilling, are less likely to churn.


Questions