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

Luke Philip Ogweno and Hong Shi

Introduction

Customer churn is one of the major concerns for large companies due to its direct effect on the company's revenue, especially in the telecom field. Companies are seeking to develop the customer churn prediction model to predict the risk of customer churns(Malikireddy and Kasa 2021).

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. One such interesting metric is customer churn. Another interesting metric is the monthly payments. Usually, customers want to get quality service for the best possible price. If they don't get it, then they may end up choosing another service provider.

What is customer churn?

Customer churn (or customer attrition) is a tendency of customers to abandon a brand and stop being a paying client 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 head of data analytics department at ScienceSoft Alex Bekker. Furthermore, it is common knowledge that retaining a customer is about five times less expensive than acquiring a new one ("Marketing Metrics: The Definitive Guide to Measuring

Marketing Performance" 2010), this creates pressure to have better and more effective churn campaigns.

Use cases for customer churn prediction

Among modern service providers that we can find churn prediction includes:

- **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 wireless).** 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.

Identifying at-risk customers with machine learning: problem-solving at a glance

The main trait of machine learning is building systems capable of finding patterns in data, learning from it without explicit programming. In the context of customer churn prediction, these are online behavior characteristics that indicate decreasing customer satisfaction from using company services/products.

The advancement of machine learning and artificial intelligence tends to increases the possibilities to predict customer churns with high performance. The Support system and consumer service dissatisfaction is the main reason to the customer churn. Forecasting the customer churning risk helps the companies to deal with the customer churn problem [(Lalwani et al. 2021), (Al-Mashraie, Chung, and Jeon 2020)].

Generally, machine learning techniques analyze the customer characteristics by using the datasets like call details, account and billing information, the future behavior of customers with personal demographics. Initially, data mining techniques are primarily applied to the churn prediction which is predicted by the telecom churners. For instance, neural networks and decision trees are applied to develop accurate churn prediction systems [(Idris, Iftikhar,

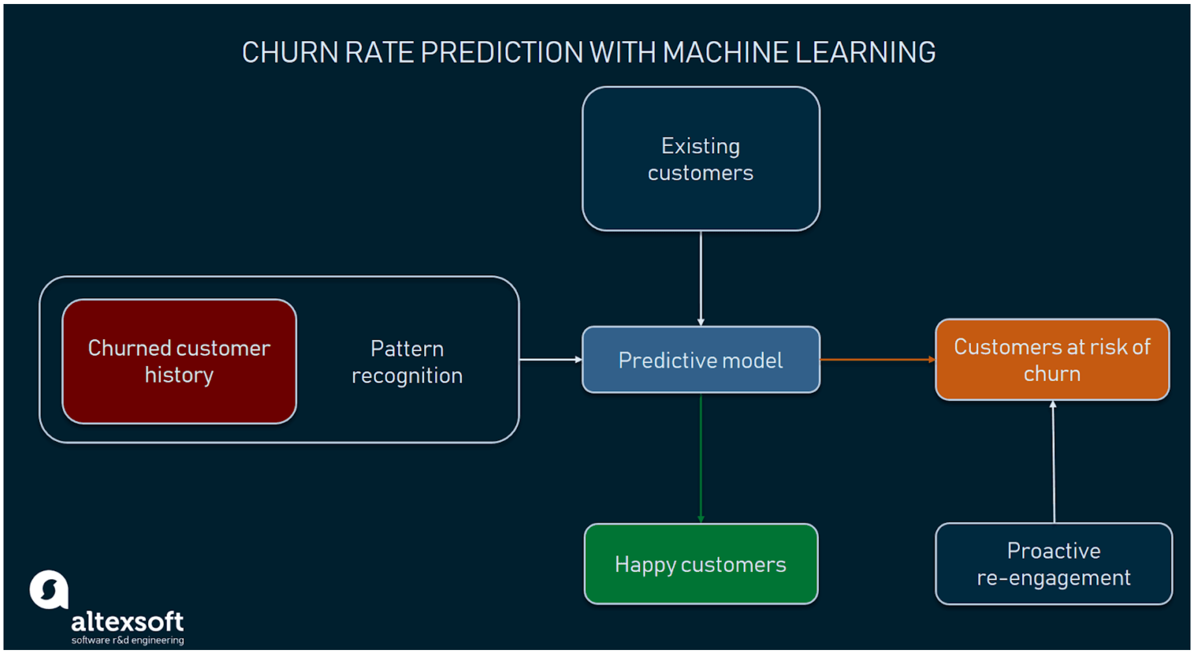


Figure 1: Figure 1: Churn rate predictive model

and Rehman 2017), (Vo et al. 2021)]. Various machine learning algorithm was applied to analyze the churning task like artificial neural networks, random forest, the statistical classifier (KNN), logistic regression, decision tree, support vector machines, and Naïve Bayes. The hybrid classification of more than one method was applied in the churn prediction which outperforms the single algorithm [(Vijaya and Sivasankar 2018), (De Caigny, Coussement, and De Bock 2018)]. Various feature selection and classifier methods are applied in the existing customer churn prediction model [(Alboukaey, Joukhadar, and Ghneim 2020)].

Methodology

The overall scope of work data scientists carry out to build ML-powered systems capable to forecast customer attrition may include the following:

- Understanding a problem and final goal
- Data collection
- Data preparation and preprocessing
- Modeling and testing
- Model deployment and monitoring

Figure 2 shows the method adopted for this project

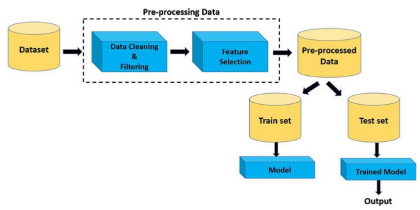


Figure 2: Figure 2: Exploratory Data Analysis

Objectives:

The objectives of this work is to analyse the churn in the telco industry using machine leaning algorithms such as logistic regresion, decision tree and random forest.

Classification

From a machine learning perspective, a churn model is a classification algorithm. In the sense that using historical information, a prediction of which current customers are more like ly to defect, is made. This model is normally created using one of a number of well establish algorithms (Logistic regression, neural networks, random forests, among others)[(KhakAbi, Gholamian, and Namvar 2010), (Ngai, Xiu, and Chau 2009)]

The goal of classification is to determine to which class or category a data point (customer in our case) belongs to. For classification problems, data scientists would use historical data with predefined target variables AKA labels (churner/non-churner) – answers that need to be predicted – to train an algorithm. With classification, businesses can answer the following questions:

- Will this customer churn or not?
- Will a customer renew their subscription?
- Will a user downgrade a pricing plan?
- Are there any signs of unusual customer behavior?

Regression

Customer churn prediction can be also formulated as a regression task. Regression analysis is a statistical technique to estimate the relationship between a target variable and other data values that influence the target variable, expressed in continuous values.

Data Overview

The data was downloaded from IBM Sample Data Sets for customer retention programs. (**IBM Sample Data Sets**). The goal of this project is to predict behaviors of churn or not churn to help retain customers. 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

Load libraries

Load dataset:

```
'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" ...
```

The raw data contains 7043 rows (customers) and 21 columns (features). The "Churn" column is our target. We used all other columns as features to our model. In the dataset only 1869 customers are churners, leading to a churn ratio of 26.54%.

Exploration and Data Analysis (EDA)

Missing values in each columns

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 all 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
```

Check missingness in the variables

Based on the summary, there is no missing data in this dataset!

Data wrangling

Look at the variables, we can see that we have some wranglings to do.

1. We will change "No internet service" to "No" for six columns, they are: "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "streamingTV", "streamingMovies".

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

3. Grouping Tenure

Since 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
```

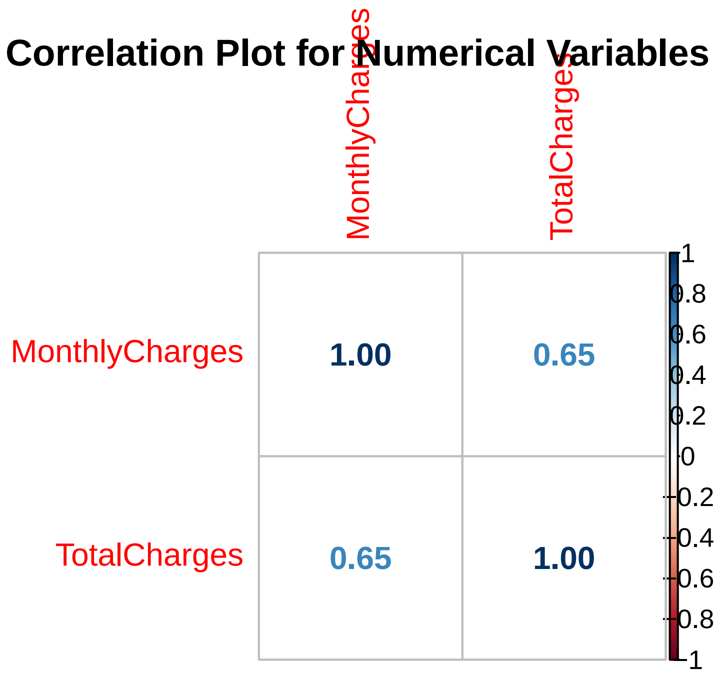
Grouping as shown below

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

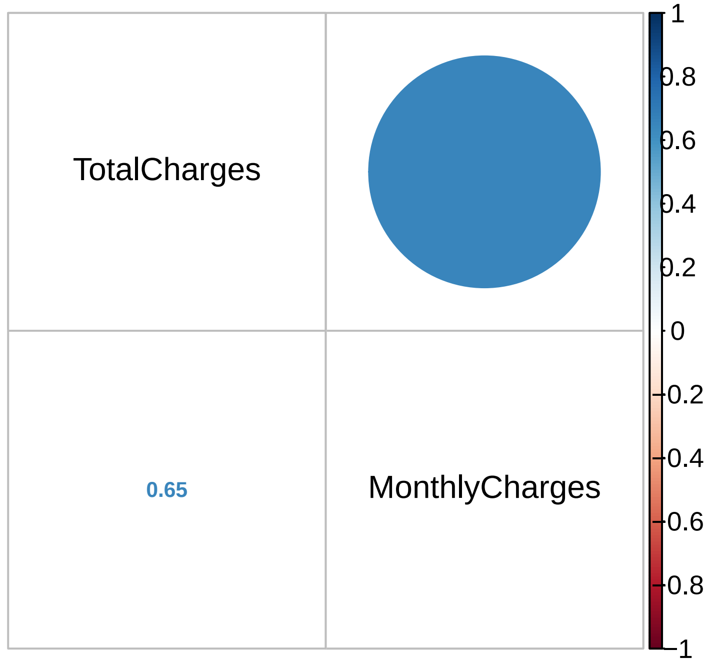
5. Remove the columns we do not need for the analysis.

Exploratory data analysis and feature selection

Correlation between numeric variables




```
churn %>%
  dplyr::select (TotalCharges, MonthlyCharges) %>%
  cor() %>%
  corrplot.mixed(upper = "circle", tl.col = "black", number.cex = 0.7)
```



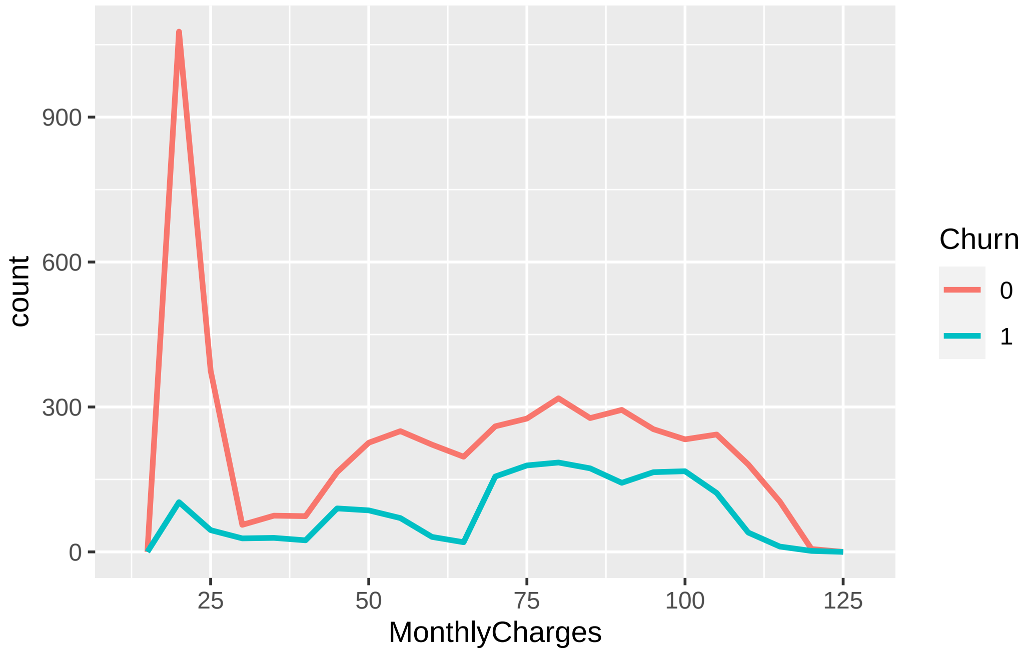
The plot shows high correlations between Totalcharges and tenure and between TotalCharges and MonthlyCharges. Pay attention to these variables while training models later. Multi-collinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set. 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 remove Total Charges.

Remove TotalCharges

Continuous Variables

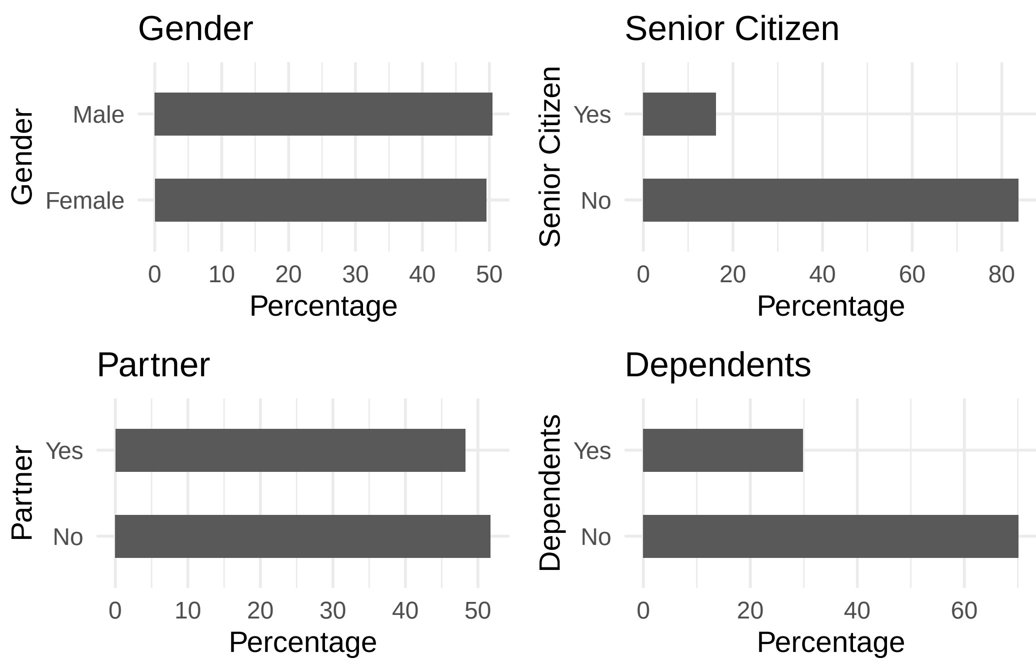
For continuous variables, let's check for distributions.



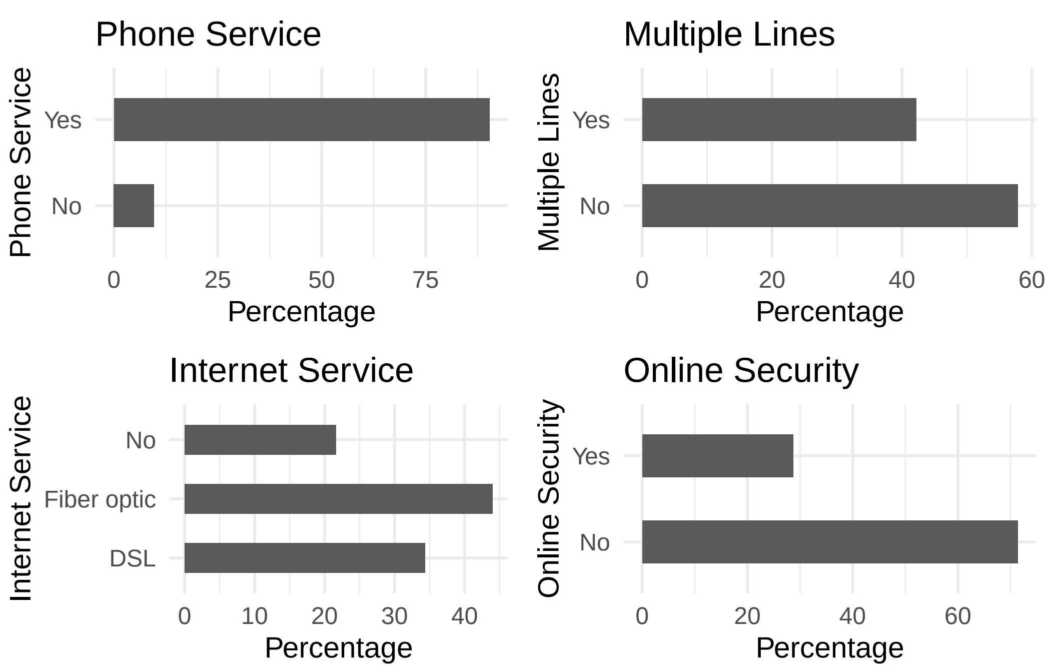
The number of current customers with MonthlyCharges below $25 is extremly high. For the customers with Monthlycharges greater than $30, the distributions are similar between who churned and who did not churn.

Bar plots of categorical variables

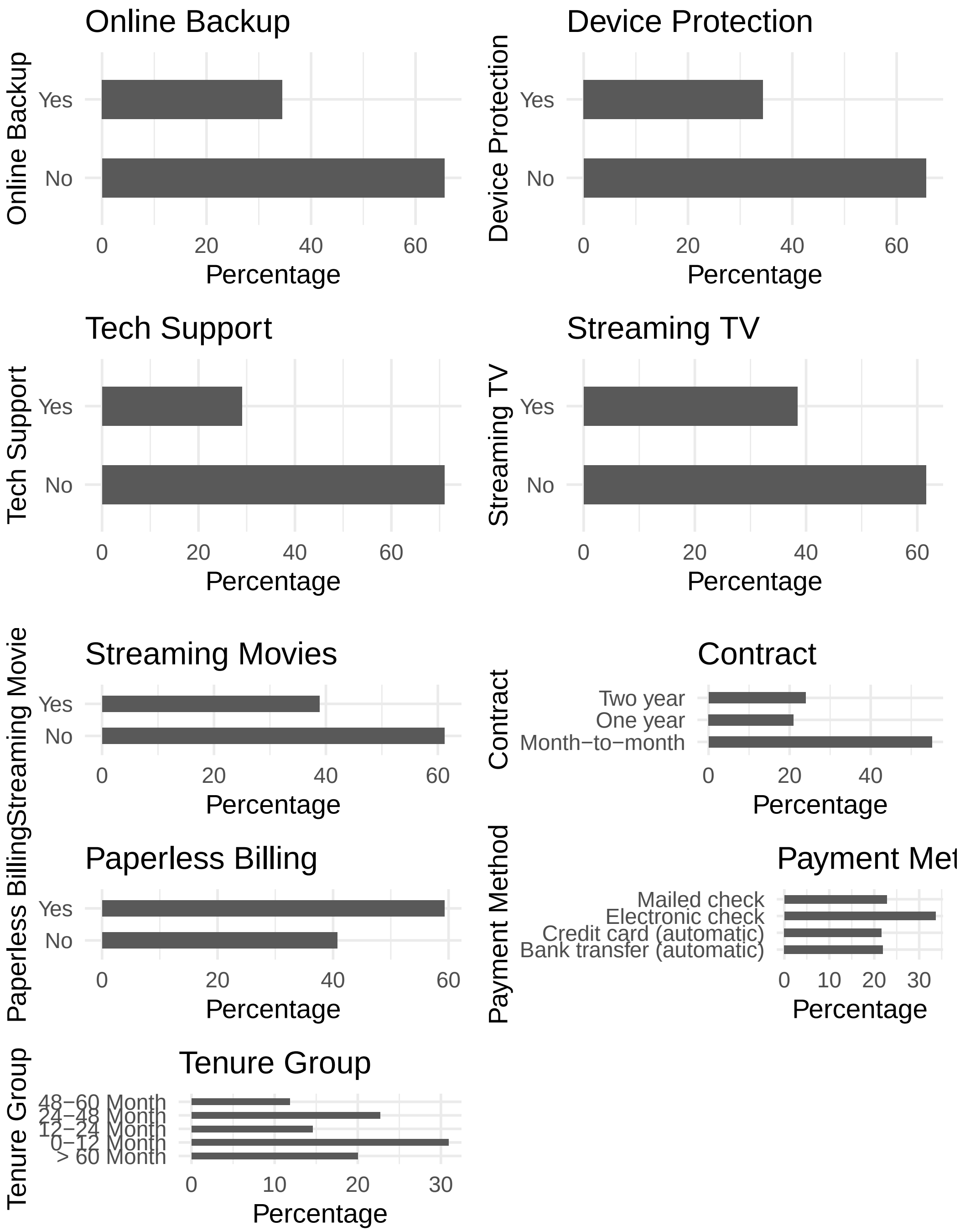
```
p1 <- ggplot(churn, aes(x=gender)) + ggtitle("Gender") + xlab("Gender") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + co
p2 <- ggplot(churn, aes(x=SeniorCitizen)) + ggtitle("Senior Citizen") + xlab("Senior Citiz
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + co
p3 <- ggplot(churn, aes(x=Partner)) + ggtitle("Partner") + xlab("Partner") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + co
p4 <- ggplot(churn, aes(x=Dependents)) + ggtitle("Dependents") + xlab("Dependents") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + co
grid.arrange(p1, p2, p3, p4, ncol=2)
```



Gender
Gender
Male
Female
0
10
20
30
40
50
Percentage
Senior Citizen
Senior Citizen
Yes
No
0
20
40
60
80
Percentage
Partner
Partner
Yes
No
0
10
20
30
40
50
Percentage
Dependents
Dependents
Yes
No
0
20
40
60
Percentage



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



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

Logistic Regression

First, we split the data into training and testing sets

Check out the results if correct

```
[1] 4924   19
```

```
[1] 2108   19
```

Fitting the Logistic Regression Model

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0223  -0.6827  -0.2943   0.6591   3.0797

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -1.718434   0.991254  -1.734  0.08299 .
genderMale                            0.025449   0.077522   0.328  0.74270
SeniorCitizenYes                      0.135299   0.101989   1.327  0.18464
PartnerYes                           -0.010482   0.092400  -0.113  0.90968
DependentsYes                        -0.115191   0.106844  -1.078  0.28098
PhoneServiceYes                      -0.314323   0.779740  -0.403  0.68687
MultipleLinesYes                      0.302531   0.211921   1.428  0.15342
InternetServiceFiber optic            1.053578   0.957783   1.100  0.27132
InternetServiceNo                    -0.921366   0.966874  -0.953  0.34062
OnlineSecurityYes                    -0.374999   0.216227  -1.734  0.08287 .
OnlineBackupYes                      -0.188508   0.210327  -0.896  0.37011
DeviceProtectionYes                   0.043049   0.211563   0.203  0.83876
TechSupportYes                       -0.357279   0.215674  -1.657  0.09761 .
StreamingTVYes                        0.362818   0.392445   0.925  0.35522
StreamingMoviesYes                    0.467447   0.392981   1.189  0.23425
ContractOne year                     -0.679920   0.125521  -5.417 6.07e-08 ***
ContractTwo year                     -1.703434   0.221138  -7.703 1.33e-14 ***
PaperlessBillingYes                   0.361303   0.088747   4.071 4.68e-05 ***
PaymentMethodCredit card (automatic) -0.166205   0.135479  -1.227  0.21990
PaymentMethodElectronic check         0.294830   0.110773   2.662  0.00778 **
```

```
PaymentMethodMailed check            -0.040806   0.133750  -0.305  0.76030
MonthlyCharges                       -0.007911   0.038097  -0.208  0.83550
tenure_group0-12 Month                1.686048   0.201812   8.355  < 2e-16 ***
tenure_group12-24 Month               0.817898   0.197138   4.149 3.34e-05 ***
tenure_group24-48 Month               0.326604   0.181574   1.799  0.07206 .
tenure_group48-60 Month               0.169953   0.196791   0.864  0.38779
---
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: 4108.4  on 4898  degrees of freedom
AIC: 4160.4

Number of Fisher Scoring iterations: 6
```

Feature Analysis

The top three most-relevant features include Contract, tenure_group and PaperlessBilling.

```
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     0.01      4922     5702.8  0.924233
SeniorCitizen     1    82.48      4921     5620.3 < 2.2e-16 ***
Partner           1   119.70      4920     5500.6 < 2.2e-16 ***
Dependents        1    34.86      4919     5465.7 3.546e-09 ***
PhoneService      1     1.65      4918     5464.1  0.198719
MultipleLines     1     6.72      4917     5457.3  0.009534 **
InternetService   2   465.51      4915     4991.8 < 2.2e-16 ***
OnlineSecurity    1   177.39      4914     4814.5 < 2.2e-16 ***
OnlineBackup      1    81.50      4913     4733.0 < 2.2e-16 ***
DeviceProtection  1    33.49      4912     4699.5 7.161e-09 ***
TechSupport       1    82.20      4911     4617.3 < 2.2e-16 ***
```

```
StreamingTV       1     3.75      4910     4613.5  0.052662 .
StreamingMovies   1     3.33      4909     4610.2  0.068162 .
Contract          2   280.15      4907     4330.0 < 2.2e-16 ***
PaperlessBilling  1    19.26      4906     4310.8 1.141e-05 ***
PaymentMethod     3    37.63      4903     4273.1 3.390e-08 ***
MonthlyCharges    1     0.18      4902     4273.0  0.669074
tenure_group      4   164.58      4898     4108.4 < 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 Logistic Regression model

```
[1] "Logistic Regression Accuracy 0.79696394686907"
```

Logistic Regression Confusion Matrix

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

```
   FALSE TRUE
  0  1409  139
  1   289  271
```

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.

```
Waiting for profiling to be done...
```

```
                                               OR      2.5 %     97.5 %
(Intercept)                             0.1793467 0.02564311  1.2502683
genderMale                              1.0257760 0.88120001  1.1941936
SeniorCitizenYes                        1.1448795 0.93713478  1.3978985
PartnerYes                              0.9895731 0.82570196  1.1862218
DependentsYes                           0.8911961 0.72230050  1.0981891
PhoneServiceYes                         0.7302832 0.15832153  3.3679810
MultipleLinesYes                        1.3532792 0.89341436  2.0508714
InternetServiceFiber optic              2.8678952 0.43931006 18.7842421
InternetServiceNo                       0.3979750 0.05976483  2.6482240
OnlineSecurityYes                       0.6872899 0.44953668  1.0494912
OnlineBackupYes                         0.8281941 0.54825169  1.2506807
DeviceProtectionYes                     1.0439886 0.68953415  1.5806270
TechSupportYes                          0.6995774 0.45802014  1.0669947
StreamingTVYes                          1.4373740 0.66623367  3.1040516
StreamingMoviesYes                      1.5959144 0.73912213  3.4508941
ContractOne year                        0.5066575 0.39516159  0.6465321
ContractTwo year                        0.1820572 0.11629758  0.2772623
PaperlessBillingYes                     1.4351985 1.20644574  1.7085688
PaymentMethodCredit card (automatic) 0.8468730 0.64891637  1.1039472
PaymentMethodElectronic check           1.3428977 1.08146722  1.6698217
PaymentMethodMailed check               0.9600158 0.73884042  1.2483336
MonthlyCharges                          0.9921201 0.92068723  1.0690280
tenure_group0-12 Month                  5.3981068 3.64716507  8.0490601
tenure_group12-24 Month                 2.2657314 1.54358458  3.3449863
tenure_group24-48 Month                 1.3862521 0.97377240  1.9855210
tenure_group48-60 Month                 1.1852494 0.80608468  1.7448971
```

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

```
Call:
glm(formula = Churn ~ SeniorCitizen + PhoneService + OnlineSecurity +
    OnlineBackup + DeviceProtection + TechSupport + Contract +
    PaperlessBilling + PaymentMethod + MonthlyCharges + tenure_group,
    family = binomial(link = "logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9904  -0.6867  -0.2959   0.6715   3.0832
```

```
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         -2.732498   0.262870 -10.395  < 2e-16 ***
SeniorCitizenYes                     0.157093   0.099739   1.575  0.11525    
PhoneServiceYes                     -1.078416   0.150781  -7.152 8.54e-13 ***
OnlineSecurityYes                   -0.602009   0.100127  -6.012 1.83e-09 ***
OnlineBackupYes                     -0.398881   0.093163  -4.282 1.86e-05 ***
DeviceProtectionYes                 -0.163802   0.096077  -1.705  0.08821 .  
TechSupportYes                      -0.576492   0.101647  -5.672 1.42e-08 ***
ContractOne year                    -0.694263   0.123342  -5.629 1.82e-08 ***
ContractTwo year                    -1.713239   0.218958  -7.825 5.10e-15 ***
PaperlessBillingYes                  0.367608   0.088554   4.151 3.31e-05 ***
PaymentMethodCredit card (automatic) -0.168582   0.135205  -1.247  0.21245    
PaymentMethodElectronic check        0.291114   0.110575   2.633  0.00847 ** 
PaymentMethodMailed check           -0.035926   0.133237  -0.270  0.78744    
MonthlyCharges                       0.033539   0.002231  15.032  < 2e-16 ***
tenure_group0-12 Month               1.646261   0.193295   8.517  < 2e-16 ***
tenure_group12-24 Month              0.787423   0.193501   4.069 4.71e-05 ***
tenure_group24-48 Month              0.306900   0.180063   1.704  0.08831 .  
tenure_group48-60 Month              0.166575   0.196124   0.849  0.39570    
---
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: 4113.1  on 4906  degrees of freedom
AIC: 4149.1

Number of Fisher Scoring iterations: 6
```

use AIC to exclude variables based on their significance and create a new model thenuUse VIF function to check multicollinearity

```
                    SeniorCitizenYes                      PhoneServiceYes 
                            1.094939                             1.370370 
                   OnlineSecurityYes                      OnlineBackupYes 
                            1.103391                             1.244177 
                 DeviceProtectionYes                       TechSupportYes 
                            1.333299                             1.158991 
                    ContractOne year                     ContractTwo year 
                            1.312726                             1.389866 
```

```
          PaperlessBillingYes PaymentMethodCredit card (automatic)
                     1.128080                             1.565086
PaymentMethodElectronic check           PaymentMethodMailed check
                     2.038338                             1.951592
               MonthlyCharges             tenure_group0-12 Month
                     2.427078                             6.187351
     tenure_group12-24 Month             tenure_group24-48 Month
                     3.615622                             3.580184
     tenure_group48-60 Month
                     2.122871
```

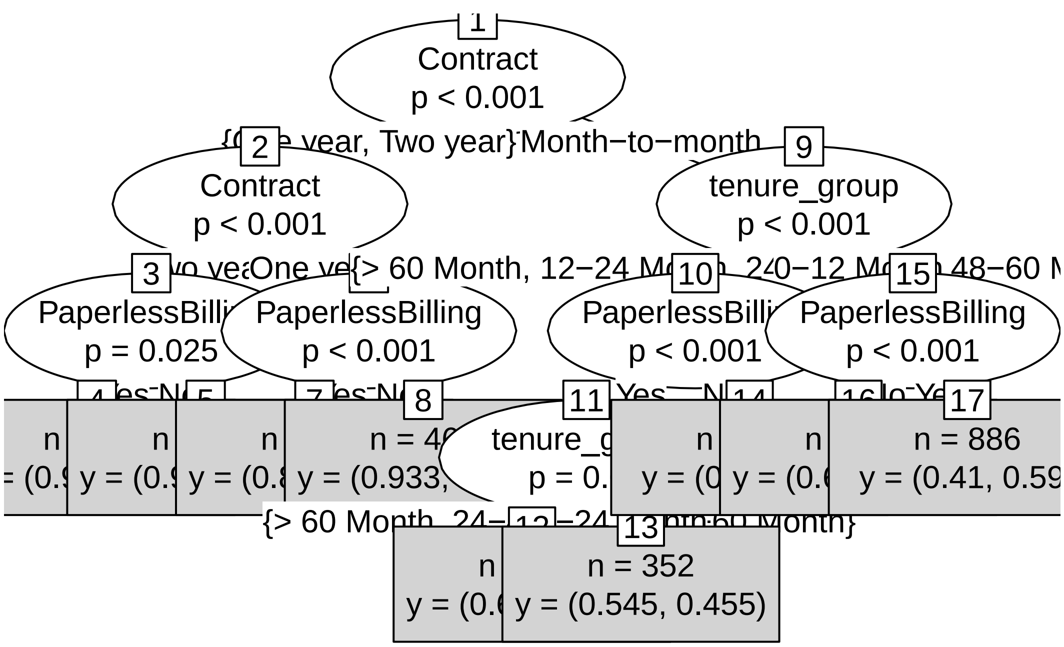
Decision Tree

Decision Tree visualization

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

Grouping

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



```
{# {r} # library(rpart) # library(rpart.plot)  # rpart.plot(tree, tweak =
1.8)     #generate the decision tree # rpart.plot(tree, type = 4, extra =
101, tweak = 1.8) #generate decision tree with more descriptions # fancyRpartPlot(tree)
```

1. Out of three variables we use, Contract is the most important variable to predict customer churn or not churn.
2. If a customer in a one-year or two-year contract, no matter 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, and in the tenure group of 0–12 month, and using PaperlessBilling, then this customer is more likely to churn.

Decision Tree Confusion Matrix

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

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

```
         Actual
Predicted   No  Yes
      No  1412  350
      Yes  136  210
```

Decision Tree Accuracy

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

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

Random Forest

Random Forest Initial Model

```
Call:
 randomForest(formula = Churn ~ ., data = training)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4
```

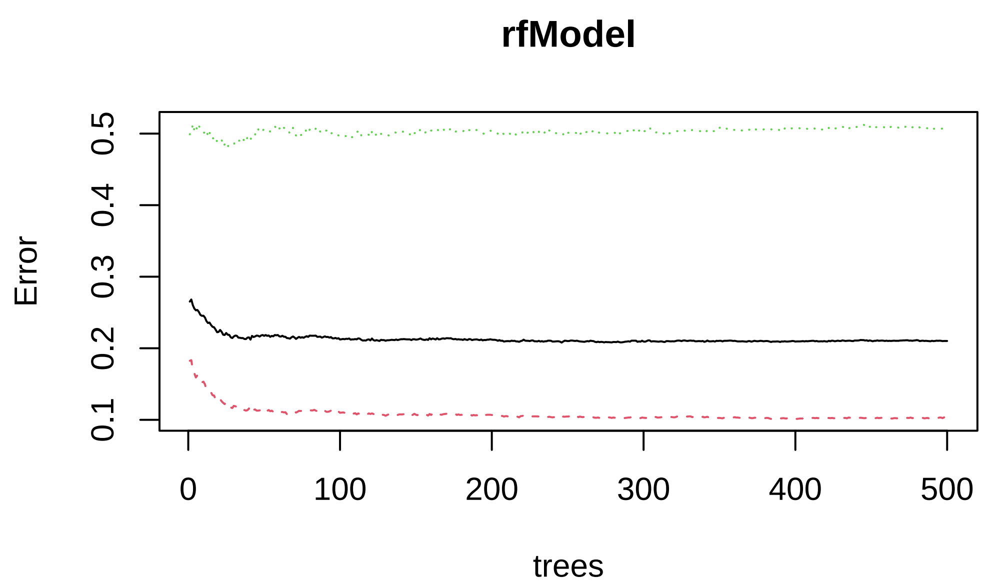
```
        OOB estimate of  error rate: 21.02%
Confusion matrix:
      No Yes class.error
No  3244 371   0.1026279
Yes  664 645   0.5072574
```

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

```
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
```

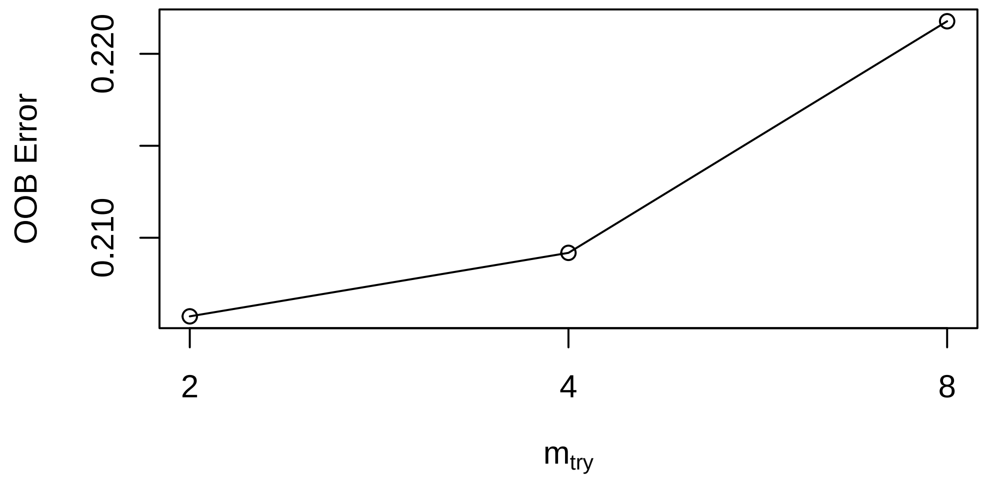
Random Forest Error Rate



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 the lowest when mtry is 2. Therefore, we choose mtry=2.

Fit the Random Forest Model After Tuning

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

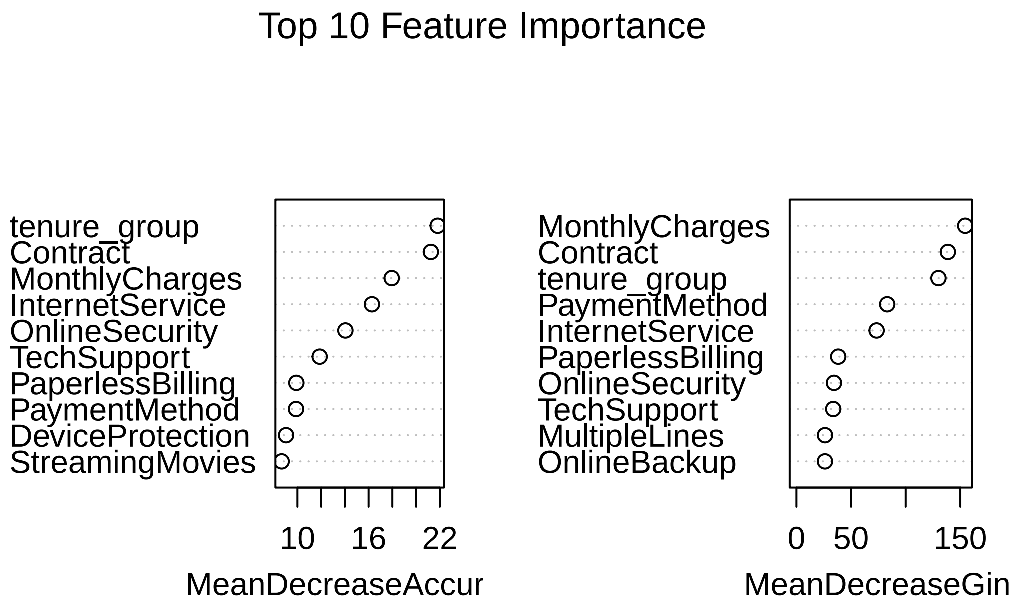
OOB error rate decreased to 20.98% from 21.81%% earlier.

Random Forest Predictions and Confusion Matrix After Tuning

```
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 the sensitivity improved, compared with the initial Random Forest model.

Random Forest 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 learnt several important things::

1. Features such as tenure_group, Contract, PaperlessBilling, MonthlyCharges and InternetService appear to play a role in customer churn.
2. There does not seem to be a relationship between gender and churn.
3. Customers in a month-to-month contract, with PaperlessBilling and are within 12 months tenure, are more likely to churn; On the other hand, customers with one or two year contract, with longer than 12 months tenure, that are not using PaperlessBilling, are less likely to churn.

References

Alboukaey, Nadia, Ammar Joukhadar, and Nada Ghneim. 2020. "Dynamic Behavior Based Churn Prediction in Mobile Telecom." *Expert Systems with Applications* 162 (December): 113779. https://doi.org/10.1016/j.eswa.2020.113779.

Al-Mashraie, Mohammed, Sung Hoon Chung, and Hyun Woo Jeon. 2020. "Customer Switching Behavior Analysis in the Telecommunication Industry via Push-Pull-Mooring Framework: A Machine Learning Approach." *Computers & Industrial Engineering* 144 (June): 106476. https://doi.org/10.1016/j.cie.2020.106476.

De Caigny, Arno, Kristof Coussement, and Koen W. De Bock. 2018. "A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees." *European Journal of Operational Research* 269 (2): 760–72. https://doi.org/10.1016/j.ejor.2018.02.009.

Idris, Adnan, Aksam Iftikhar, and Zia ur Rehman. 2017. "Intelligent Churn Prediction for Telecom Using GP-AdaBoost Learning and PSO Undersampling." *Cluster Computing* 22 (S3): 7241–55. https://doi.org/10.1007/s10586-017-1154-3.

KhakAbi, Sahand, Mohammad R. Gholamian, and Morteza Namvar. 2010. "Data Mining Applications in Customer Churn Management." *2010 International Conference on Intelligent Systems, Modelling and Simulation*, January. https://doi.org/10.1109/isms.2010.49.

Lalwani, Praveen, Manas Kumar Mishra, Jasroop Singh Chadha, and Pratyush Sethi. 2021. "Customer Churn Prediction System: A Machine Learning Approach." *Computing* 104 (2): 271–94. https://doi.org/10.1007/s00607-021-00908-y.

Malikireddy, Venkata Pullareddy, and Madhavi Kasa. 2021. "Customer Churns Prediction Model Based on Machine Learning Techniques: A Systematic Review." *Atlantis Highlights in Computer Sciences*. https://doi.org/10.2991/ahis.k.210913.021.

"Marketing Metrics: The Definitive Guide to Measuring Marketing Performance." 2010. *Choice Reviews Online* 48 (01). https://doi.org/10.5860/choice.48-0373.

Ngai, E. W. T., Li Xiu, and D. C. K. Chau. 2009. "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification." *Expert Systems with Applications* 36 (2): 2592–602. https://doi.org/10.1016/j.eswa.2008.02.021.

Vijaya, J., and E. Sivasankar. 2018. "Computing Efficient Features Using Rough Set Theory Combined with Ensemble Classification Techniques to Improve the Customer Churn Prediction in Telecommunication Sector." *Computing* 100 (8): 839–60. https://doi.org/10.1007/s00607-018-0633-6.

Vo, Nhi N. Y., Shaowu Liu, Xitong Li, and Guandong Xu. 2021. "Leveraging Unstructured Call Log Data for Customer Churn Prediction." *Knowledge-Based Systems* 212 (January): 106586. https://doi.org/10.1016/j.knosys.2020.106586.