R Notebook

```r
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✔ ggplot2 3.3.5     ✔ purrr   0.3.4
## ✔ tibble  3.1.6     ✔ dplyr   1.0.8
## ✔ tidyr   1.2.0     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::combine()  masks randomForest::combine()
## ✖ dplyr::filter()   masks stats::filter()
## ✖ dplyr::lag()      masks stats::lag()
## ✖ ggplot2::margin() masks randomForest::margin()
```

```r
library(ipred)
library(caTools)
```

```r
#read in csv file
churn_df <- read.csv("/Users/wyattromero/Downloads/RandomForest/churn_clean.csv")
head(churn_df)
```

	CaseOrder <int>	Customer_id <chr>	Interaction <chr>	UID <chr>
1	1	K409198	aa90260b-4141-4a24-8e36-b04ce1f4f77b	e885b299883d4f9fb18e39c75155d990
2	2	S120509	fb76459f-c047-4a9d-8af9-e0f7d4ac2524	f2de8bef964785f41a2959829830fb8a
3	3	K191035	344d114c-3736-4be5-98f7-c72c281e2d35	f1784cfa9f6d92ae816197eb175d3c71
4	4	D90850	abfa2b40-2d43-4994-b15a-989b8c79e311	dc8a365077241bb5cd5ccd305136b05e
5	5	K662701	68a861fd-0d20-4e51-a587-8a90407ee574	aabb64a116e83fdc4befc1fbab1663f9
6	6	W303516	2b451d12-6c2b-4cea-a295-ba1d6bced078	97598fd95658c80500546bc1dd312994

6 rows | 1-5 of 51 columns

```r
summary(churn_df)
```

```
##    CaseOrder     Customer_id        Interaction            UID           
##  Min.   :    1   Length:10000       Length:10000       Length:10000      
##  1st Qu.: 2501   Class :character   Class :character   Class :character  
##  Median : 5000   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 5000                                                           
##  3rd Qu.: 7500                                                           
##  Max.   :10000                                                           
##      City              State              County               Zip       
##  Length:10000       Length:10000       Length:10000       Min.   :  601  
##  Class :character   Class :character   Class :character   1st Qu.:26292  
##  Mode  :character   Mode  :character   Mode  :character   Median :48870  
##                                                           Mean   :49153  
##                                                           3rd Qu.:71866  
##                                                           Max.   :99929  
##       Lat             Lng            Population         Area          
##  Min.   :17.97   Min.   :-171.69   Min.   :     0   Length:10000      
##  1st Qu.:35.34   1st Qu.: -97.08   1st Qu.:   738   Class :character  
##  Median :39.40   Median : -87.92   Median :  2910   Mode  :character  
##  Mean   :38.76   Mean   : -90.78   Mean   :  9757                     
##  3rd Qu.:42.11   3rd Qu.: -80.09   3rd Qu.: 13168                     
##  Max.   :70.64   Max.   : -65.67   Max.   :111850                     
##    TimeZone             Job               Children           Age       
##  Length:10000       Length:10000       Min.   : 0.000   Min.   :18.00  
##  Class :character   Class :character   1st Qu.: 0.000   1st Qu.:35.00  
##  Mode  :character   Mode  :character   Median : 1.000   Median :53.00  
##                                        Mean   : 2.088   Mean   :53.08  
##                                        3rd Qu.: 3.000   3rd Qu.:71.00  
##                                        Max.   :10.000   Max.   :89.00  
##      Income           Marital             Gender             Churn          
##  Min.   :   348.7   Length:10000       Length:10000       Length:10000      
##  1st Qu.: 19224.7   Class :character   Class :character   Class :character  
##  Median : 33170.6   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 39806.9                                                           
##  3rd Qu.: 53246.2                                                           
##  Max.   :258900.7                                                           
##  Outage_sec_perweek     Email          Contacts      Yearly_equip_failure
##  Min.   : 0.09975   Min.   : 1.00   Min.   :0.0000   Min.   :0.000       
##  1st Qu.: 8.01821   1st Qu.:10.00   1st Qu.:0.0000   1st Qu.:0.000       
##  Median :10.01856   Median :12.00   Median :1.0000   Median :0.000       
##  Mean   :10.00185   Mean   :12.02   Mean   :0.9942   Mean   :0.398       
##  3rd Qu.:11.96949   3rd Qu.:14.00   3rd Qu.:2.0000   3rd Qu.:1.000       
##  Max.   :21.20723   Max.   :23.00   Max.   :7.0000   Max.   :6.000       
##     Techie            Contract          Port_modem           Tablet          
##  Length:10000       Length:10000       Length:10000       Length:10000      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##  InternetService       Phone             Multiple         OnlineSecurity    
##  Length:10000       Length:10000       Length:10000       Length:10000      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##  OnlineBackup       DeviceProtection   TechSupport        StreamingTV       
##  Length:10000       Length:10000       Length:10000       Length:10000      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##  StreamingMovies    PaperlessBilling   PaymentMethod          Tenure      
##  Length:10000       Length:10000       Length:10000       Min.   : 1.000  
##  Class :character   Class :character   Class :character   1st Qu.: 7.918  
##  Mode  :character   Mode  :character   Mode  :character   Median :35.431  
##                                                           Mean   :34.526  
##                                                           3rd Qu.:61.480  
##                                                           Max.   :71.999  
##  MonthlyCharge    Bandwidth_GB_Year     Item1           Item2      
##  Min.   : 79.98   Min.   : 155.5    Min.   :1.000   Min.   :1.000  
##  1st Qu.:139.98   1st Qu.:1236.5    1st Qu.:3.000   1st Qu.:3.000  
##  Median :167.48   Median :3279.5    Median :3.000   Median :4.000  
##  Mean   :172.62   Mean   :3392.3    Mean   :3.491   Mean   :3.505  
##  3rd Qu.:200.73   3rd Qu.:5586.1    3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :290.16   Max.   :7159.0    Max.   :7.000   Max.   :7.000  
##      Item3           Item4           Item5           Item6           Item7     
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00  
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.00  
##  Median :3.000   Median :3.000   Median :3.000   Median :3.000   Median :4.00  
##  Mean   :3.487   Mean   :3.498   Mean   :3.493   Mean   :3.497   Mean   :3.51  
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00  
##  Max.   :8.000   Max.   :7.000   Max.   :7.000   Max.   :8.000   Max.   :7.00  
##      Item8      
##  Min.   :1.000  
##  1st Qu.:3.000  
##  Median :3.000  
##  Mean   :3.496  
##  3rd Qu.:4.000  
##  Max.   :8.000
```

```r
str(churn_df)
```

```
## 'data.frame':    10000 obs. of  50 variables:
##  $ CaseOrder           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Customer_id         : chr  "K409198" "S120509" "K191035" "D90850" ...
##  $ Interaction         : chr  "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0f7d4ac2524" "3
44d114c-3736-4be5-98f7-c72c281e2d35" "abfa2b40-2d43-4994-b15a-989b8c79e311" ...
##  $ UID                 : chr  "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a2959829830fb8a" "f1784cfa9
f6d92ae816197eb175d3c71" "dc8a365077241bb5cd5ccd305136b05e" ...
##  $ City                : chr  "Point Baker" "West Branch" "Yamhill" "Del Mar" ...
##  $ State               : chr  "AK" "MI" "OR" "CA" ...
##  $ County              : chr  "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...
##  $ Zip                 : int  99927 48661 97148 92014 77461 31030 37847 73109 34771 45237 ...
##  $ Lat                 : num  56.3 44.3 45.4 33 29.4 ...
##  $ Lng                 : num  -133.4 -84.2 -123.2 -117.2 -95.8 ...
##  $ Population          : int  38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...
##  $ Area                : chr  "Urban" "Urban" "Urban" "Suburban" ...
##  $ TimeZone            : chr  "America/Sitka" "America/Detroit" "America/Los_Angeles" "America/Los_Angeles"
...
##  $ Job                 : chr  "Environmental health practitioner" "Programmer, multimedia" "Chief Financial Of
ficer" "Solicitor" ...
##  $ Children            : int  0 1 4 1 0 3 0 2 2 1 ...
##  $ Age                 : int  68 27 50 48 83 83 79 30 49 86 ...
##  $ Income              : num  28562 21705 9610 18925 40074 ...
##  $ Marital             : chr  "Widowed" "Married" "Widowed" "Married" ...
##  $ Gender              : chr  "Male" "Female" "Female" "Male" ...
##  $ Churn               : chr  "No" "Yes" "No" "No" ...
##  $ Outage_sec_perweek  : num  7.98 11.7 10.75 14.91 8.15 ...
##  $ Email               : int  10 12 9 15 16 15 10 16 20 18 ...
##  $ Contacts            : int  0 0 0 2 2 3 0 0 2 1 ...
##  $ Yearly_equip_failure: int  1 1 1 0 1 1 1 0 3 0 ...
##  $ Techie              : chr  "No" "Yes" "Yes" "Yes" ...
##  $ Contract            : chr  "One year" "Month-to-month" "Two Year" "Two Year" ...
##  $ Port_modem          : chr  "Yes" "No" "Yes" "No" ...
##  $ Tablet              : chr  "Yes" "Yes" "No" "No" ...
##  $ InternetService     : chr  "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...
##  $ Phone               : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Multiple            : chr  "No" "Yes" "Yes" "No" ...
##  $ OnlineSecurity      : chr  "Yes" "Yes" "No" "Yes" ...
##  $ OnlineBackup        : chr  "Yes" "No" "No" "No" ...
##  $ DeviceProtection    : chr  "No" "No" "No" "No" ...
##  $ TechSupport         : chr  "No" "No" "No" "No" ...
##  $ StreamingTV         : chr  "No" "Yes" "No" "Yes" ...
##  $ StreamingMovies     : chr  "Yes" "Yes" "Yes" "No" ...
##  $ PaperlessBilling    : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ PaymentMethod       : chr  "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (automatic)"
"Mailed Check" ...
##  $ Tenure              : num  6.8 1.16 15.75 17.09 1.67 ...
##  $ MonthlyCharge       : num  172 243 160 120 150 ...
##  $ Bandwidth_GB_Year   : num  905 801 2055 2165 271 ...
##  $ Item1               : int  5 3 4 4 4 3 6 2 5 2 ...
##  $ Item2               : int  5 4 4 4 4 3 5 2 4 2 ...
##  $ Item3               : int  5 3 2 4 4 3 6 2 4 2 ...
##  $ Item4               : int  3 3 4 2 3 2 4 5 3 2 ...
##  $ Item5               : int  4 4 4 5 4 4 1 2 4 5 ...
##  $ Item6               : int  4 3 3 4 4 3 5 3 3 2 ...
##  $ Item7               : int  3 4 3 3 4 3 5 4 4 3 ...
##  $ Item8               : int  4 4 3 3 5 3 5 5 4 3 ...
```

```r
#Rename survey variables for clarity
churn_df <- rename(churn_df, 'TimelyResponse' = 'Item1',
       'Fixes' = 'Item2',
       'Replacements' = 'Item3',
       'Reliability' = 'Item4',
       'Options' = 'Item5',
       'Respectfulness' = 'Item6',
       'Courteous' = 'Item7',
       'Listening' = 'Item8')
```

```r
churn_df$Gender <- ifelse(churn_df$Gender == "Female", 1, 0)
churn_df$Churn <- ifelse(churn_df$Churn == "Yes", 1, 0)
churn_df$Techie <- ifelse(churn_df$Techie == "Yes", 1, 0)
churn_df$Contract <- ifelse(churn_df$Contract == "Two Year", 1, 0)
churn_df$Port_modem <- ifelse(churn_df$Port_modem == "Yes", 1, 0)
churn_df$Tablet <- ifelse(churn_df$Tablet == "Yes", 1, 0)
churn_df$InternetService <- ifelse(churn_df$InternetService == "Fiber Optic", 1, 0)
churn_df$Phone <- ifelse(churn_df$Phone == "Yes", 1, 0)
churn_df$Multiple <- ifelse(churn_df$Multiple == "Yes", 1, 0)
churn_df$OnlineSecurity <- ifelse(churn_df$OnlineSecurity == "Yes", 1, 0)
churn_df$OnlineBackup <- ifelse(churn_df$OnlineBackup == "Yes", 1, 0)
churn_df$DeviceProtection <- ifelse(churn_df$DeviceProtection == "Yes", 1, 0)
churn_df$TechSupport <- ifelse(churn_df$TechSupport == "Yes", 1, 0)
churn_df$StreamingTV <- ifelse(churn_df$StreamingTV == "Yes", 1, 0)
churn_df$StreamingMovies <- ifelse(churn_df$StreamingMovies == "Yes", 1, 0)
churn_df$PaperlessBilling <- ifelse(churn_df$PaperlessBilling == "Yes", 1, 0)
head(churn_df)
```

	CaseOrder <int>	Customer_id <chr>	Interaction <chr>	UID <chr>
1	1	K409198	aa90260b-4141-4a24-8e36-b04ce1f4f77b	e885b299883d4f9fb18e39c75155d990
2	2	S120509	fb76459f-c047-4a9d-8af9-e0f7d4ac2524	f2de8bef964785f41a2959829830fb8a
3	3	K191035	344d114c-3736-4be5-98f7-c72c281e2d35	f1784cfa9f6d92ae816197eb175d3c71
4	4	D90850	abfa2b40-2d43-4994-b15a-989b8c79e311	dc8a365077241bb5cd5ccd305136b05e
5	5	K662701	68a861fd-0d20-4e51-a587-8a90407ee574	aabb64a116e83fdc4befc1fbab1663f9
6	6	W303516	2b451d12-6c2b-4cea-a295-ba1d6bced078	97598fd95658c80500546bc1dd312994

6 rows | 1-5 of 51 columns

```r
churn_df <- select(churn_df, -c('Population', 'CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State',
'County', 'Zip', 'Lat', 'Lng', 'Area', 'TimeZone', 'Job', 'Marital', 'PaymentMethod', 'TimelyResponse', 'Fixes',
'Replacements', 'Reliability', 'Options', 'Respectfulness', 'Courteous', 'Listening'))
head(churn_df)
```

	Children <int>	... <int>	Income <dbl>	Gender <dbl>	Churn <dbl>	Outage_sec_perweek <dbl>	Email <int>	Contacts <int>	Yearly_equip_failure <int>
1	0	68	28561.99	0	0	7.978323	10	0	1
2	1	27	21704.77	1	1	11.699080	12	0	1
3	4	50	9609.57	1	0	10.752800	9	0	1
4	1	48	18925.23	0	0	14.913540	15	2	0
5	0	83	40074.19	0	1	8.147417	16	2	1
6	3	83	22660.20	1	0	8.420993	15	3	1

6 rows | 1-10 of 27 columns

```r
churn_df <- select(churn_df, c('Children', 'Age', 'Income', 'MonthlyCharge', 'Outage_sec_perweek', 'Yearly_equip_
failure', 'Tenure','Bandwidth_GB_Year'))
head(churn_df)
```

	Children <int>	... <int>	Income <dbl>	MonthlyCharge <dbl>	Outage_sec_perweek <dbl>	Yearly_equip_failure <int>	Tenure <dbl>
1	0	68	28561.99	172.4555	7.978323	1	6.795513
2	1	27	21704.77	242.6326	11.699080	1	1.156681
3	4	50	9609.57	159.9476	10.752800	1	15.754144
4	1	48	18925.23	119.9568	14.913540	0	17.087227
5	0	83	40074.19	149.9483	8.147417	1	1.670972
6	3	83	22660.20	185.0077	8.420993	1	7.000994

6 rows | 1-8 of 9 columns

```r
dim(churn_df)
```

```
## [1] 10000     8
```

```r
churn_df[ churn_df == "?"] <- NA
colSums(is.na(churn_df))
```

```
##             Children                  Age               Income 
##                    0                    0                    0 
##        MonthlyCharge   Outage_sec_perweek Yearly_equip_failure 
##                    0                    0                    0 
##               Tenure    Bandwidth_GB_Year 
##                    0                    0
```

```r
write.csv(churn_df, "churn_prepared_D209_2.csv", row.names = TRUE)
```

```r
#Data splitting
set.seed(1)
split <- sample.split(churn_df, SplitRatio = 0.7)
train <- subset(churn_df, split == "TRUE")
test <- subset(churn_df, split == "FALSE")

dim(train)
```

```
## [1] 6250    8
```

```r
dim(test)
```

```
## [1] 3750    8
```

```r
write.csv(train, "train_dataset_D209_2", row.names = TRUE)
write.csv(test, "test_dataset_D209_2", row.names = TRUE)
```

```r
set.seed(2)
rf <- randomForest(
  formula = Bandwidth_GB_Year ~ .,
  data = train,
  mtry = 3,
  ntree = 500,
  importance = TRUE,
  type = "regression",
  na.action = na.omit
)
rf
```

```
## 
## Call:
##  randomForest(formula = Bandwidth_GB_Year ~ ., data = train, mtry = 3,      ntree = 500, importance = TRUE, ty
pe = "regression", na.action = na.omit) 
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
## 
##           Mean of squared residuals: 53432.09
##                     % Var explained: 98.88
```

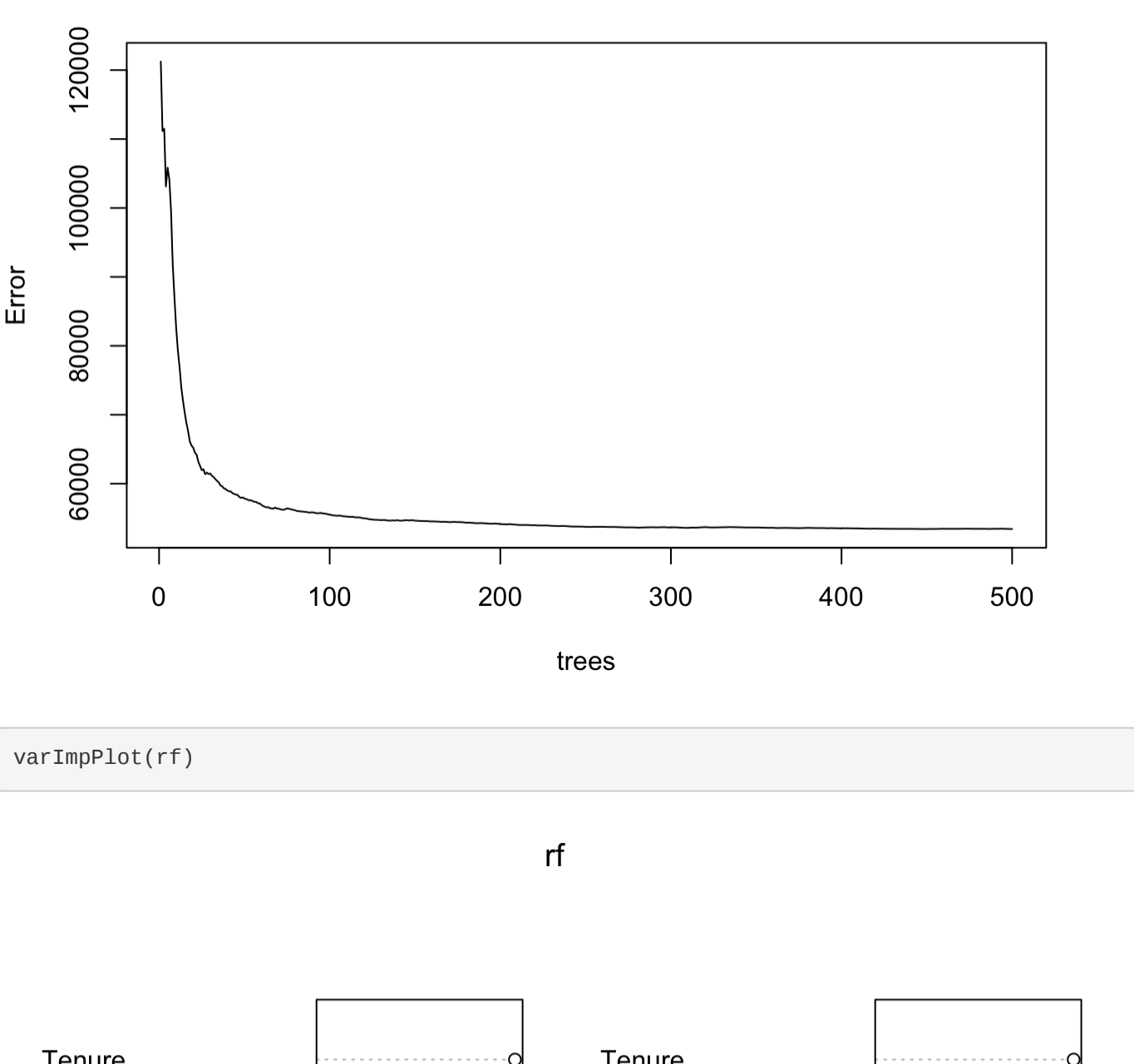
```r
mse <- which.min(rf$mse)
mse
```

```
## [1] 447
```

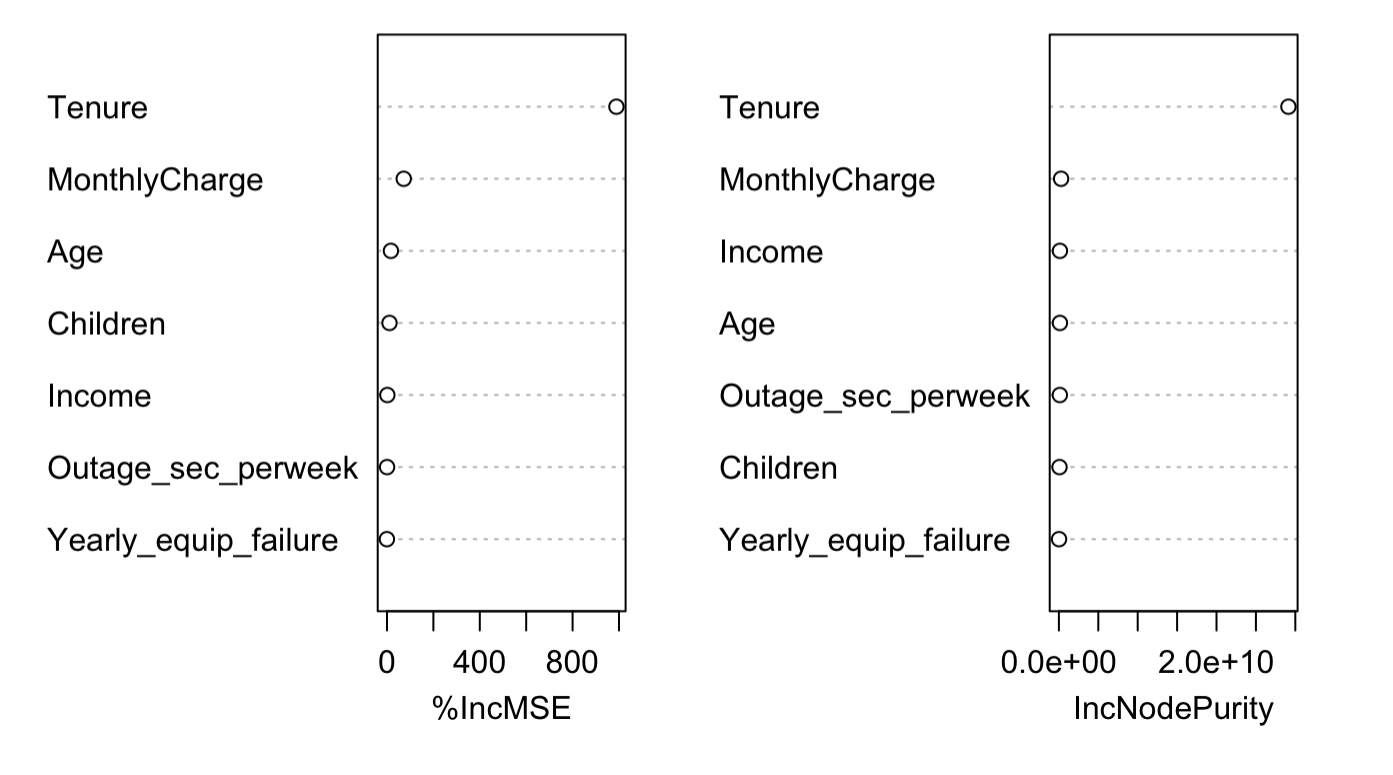
```r
accuracy <- 1 - (mse / 500)
accuracy
```

```
## [1] 0.106
```

```r
plot(rf)
```




```r
varImpPlot(rf)
```




```r
new <- data.frame(MonthlyCharge=242.6326, Tenure=1.156681, Children=1, Income=21704.77, Outage_sec_perweek=11.699
, Yearly_equip_failure=1, Age=27)
bandwidth_gb_year_pred <- predict(rf, newdata=new)
bandwidth_gb_year_pred
```

```
##        1 
## 876.4388
```