Telco Churn

Story: 14% customers left (483 of 3,333), Can we use machine learning (ML) to help inform on why they may have churned/left and reduce future churn?

Approach: Use exploratory data analysis (EDA), visualizations ML to understanding data relating to Telco churn. Demonstrate understanding of key churn data variables and provide prediction model for churn.

Data set from: https://www.kaggle.com/pangkw/telco-churn/version/3

We would like to understand using ML which data features in the Telco churn are significant to predicting churn. The data we have is 33 features with 3,333 rows/observations, of Telco data with churn results. The data provided has **14% (483)** of the total records that churned so this will be somewhat of a constraint in our ML research. First to understand the data we will use tools such as R and excel to explore.

The dependent variable in our exercise is the Churn value of 'Yes' or 'No', the independent variables are the remaining data elements that may have an impact on the dependent variable. One independent variable we can rule out is the Phone service data element which is set to 'Yes' in every record, indicating it would have no impact on the Churn since it is the same value each time. What will be of key importance to us is will one or more independent variables show more important than the rest.

The Data: Results from EDA:

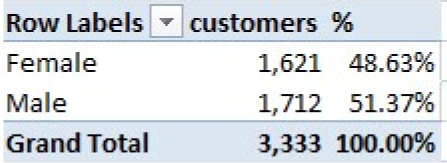
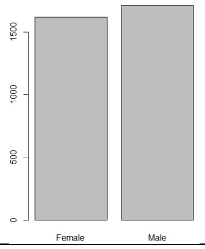
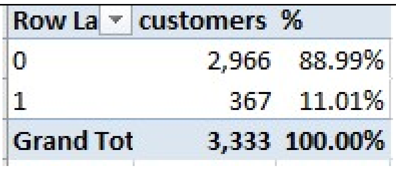
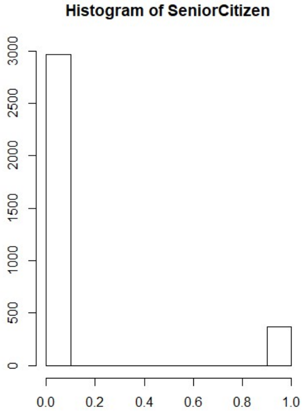
Data fields	**Definition**	**Example data**
customerID	Unique ID for customer	0002-ORFBO
gender	has: Male or Female	Female
SeniorCitizen	has: 0 or 1	0
MaritalStatus	has Yes or No	Yes
Dependents	has Yes or No	Yes
tenure	ranges from 0 to 72	9
PhoneService	all records set to yes	Yes
MultipleLines	has Yes or No	No
InternetService	has DSL, Fiber optic, No	DSL
OnlineSecurity	has No, No Internet service, Yes	No
OnlineBackup	has No, No Internet service, Yes	Yes
DeviceProtection	has No, No Internet service, Yes	No
TechSupport	has No, No Internet service, Yes	Yes

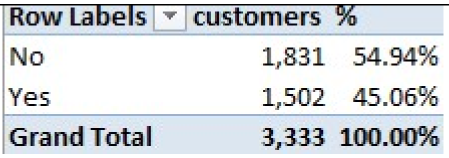
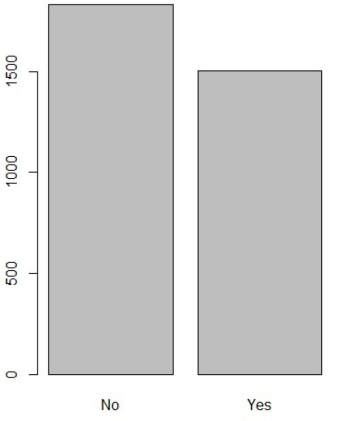
StreamingTV	has No, No Internet service, Yes	Yes
StreamingMovies	has No, No Internet service, Yes	No
Contract	has Month-to-month, One year, Two year	One year
PaperlessBilling	has Yes or No	Yes
PaymentMethod	has Bank transfer (automatic), Credit card (automatic), Electronic check, Mailed check	Mailed check
InternationalPlan	has Yes or No	No
VoiceMailPlan	has Yes or No	No
NumbervMailMessages	ranges 0 to 51	0
TotalDayMinutes	ranges 0 to 350.8	168.8
TotalDayCalls	ranges 0 to 165	137
TotalEveMinutes	ranges 0 to 363.7	241.4
TotalEveCalls	ranges 0-170	107
TotalNightMinutes	ranges 23.2 to 395	204.8
TotalNightCalls	ranges 33 to 175	106
TotalIntlMinutes	ranges 0 to 20	15.5
TotalIntlCalls	range 0 to 20	4
CustomerServiceCalls	ranges 0 to 20	0
TotalCall	ranges 194 to 418	354
TotalRevenue	ranges 18.8 to 8476.5, 5 NA's	593.3
Churn	has Yes or No	No

R code and graphics from the EDA activities

Exploratory Data Analysis (EDA)		

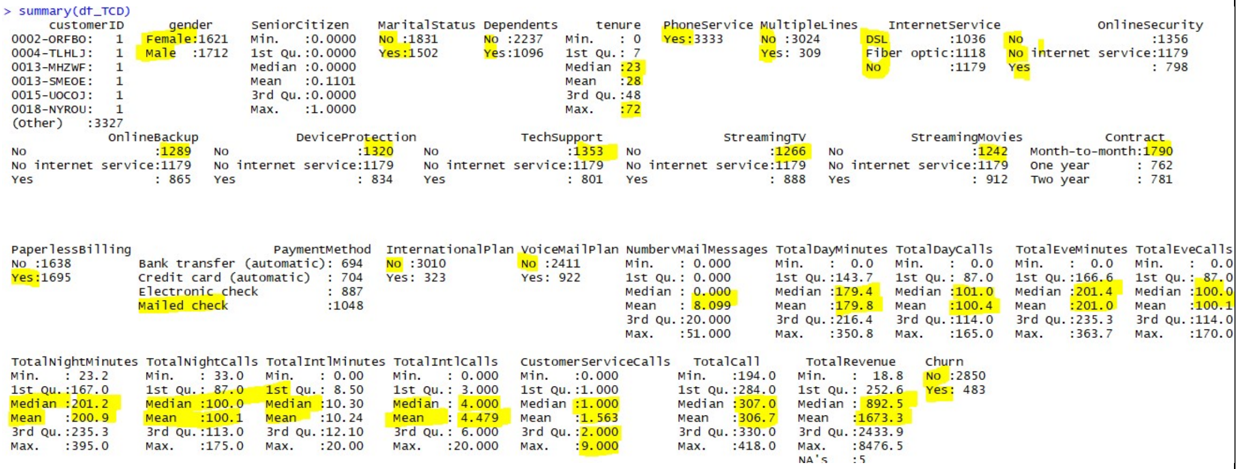
	Excel pivot	**#R plot**
		TCD <- read.csv("C:/Users/mdegra200/Documents/P2/TCD.csv")
		df_TCD = data.frame(TCD)
		attach(df_TCD)
		plot(gender)

<table>
<tr>
<td>Gender</td>
<td>
<table>
<tr><th>Row Labels</th><th>customers</th><th>%</th></tr>
<tr><td>Female</td><td>1,621</td><td>48.63%</td></tr>
<tr><td>Male</td><td>1,712</td><td>51.37%</td></tr>
<tr><td><b>Grand Total</b></td><td><b>3,333</b></td><td><b>100.00%</b></td></tr>
</table></td>
<td>
</td>
</tr>
<tr>
<td>SeniorCitizen</td>
<td>
<table>
<tr><th>Row La</th><th>customers</th><th>%</th></tr>
<tr><td>0</td><td>2,966</td><td>88.99%</td></tr>
<tr><td>1</td><td>367</td><td>11.01%</td></tr>
<tr><td><b>Grand Tot</b></td><td><b>3,333</b></td><td><b>100.00%</b></td></tr>
</table></td>
<td>

Hist(SeniorCitizen)</td>
</tr>
<tr>
<td>MaritalStatus</td>
<td>
<table>
<tr><th>Row Labels</th><th>customers</th><th>%</th></tr>
<tr><td>No</td><td>1,831</td><td>54.94%</td></tr>
<tr><td>Yes</td><td>1,502</td><td>45.06%</td></tr>
<tr><td><b>Grand Total</b></td><td><b>3,333</b></td><td><b>100.00%</b></td></tr>
</table></td>
<td>
</td>
</tr>
<tr><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td></tr>
</table>

This could take all day. So instead I run the following command:

```
summary(df_TCD) #Summarizes the data in the data frame
```



```
> summary(df_TCD)
      customerID      gender     SeniorCitizen      MaritalStatus Dependents     tenure     PhoneService MultipleLines    InternetService              OnlineSecurity
 0002-ORFBO:   1   Female:1621   Min.   :0.0000   No :1831      No :2237   Min.   : 0   Yes:3333     No :3024      DSL        :1036   No                 :1356
 0004-TLHLJ:   1   Male  :1712   1st Qu.:0.0000   Yes:1502      Yes:1096   1st Qu.: 7                Yes: 309      Fiber optic:1118   No internet service:1179
 0013-MHZWF:   1                 Median :0.0000                            Median :23                              No         :1179   Yes                : 798
 0013-SMEOE:   1                 Mean   :0.1101                            Mean   :28
 0015-UOCOJ:   1                 3rd Qu.:0.0000                            3rd Qu.:48
 0018-NYROU:   1                 Max.   :1.0000                            Max.   :72
 (Other)   :3327
             OnlineBackup             DeviceProtection               TechSupport                  StreamingTV               StreamingMovies           Contract
 No                 :1289   No                 :1320   No                 :1353   No                 :1266   No                 :1242   Month-to-month:1790
 No internet service:1179   No internet service:1179   No internet service:1179   No internet service:1179   No internet service:1179   One year      : 762
 Yes                : 865   Yes                : 834   Yes                : 801   Yes                : 888   Yes                : 912   Two year      : 781



 PaperlessBilling                   PaymentMethod   InternationalPlan VoiceMailPlan NumbervMailMessages TotalDayMinutes TotalDayCalls   TotalEveMinutes TotalEveCalls
 No :1638         Bank transfer (automatic): 694   No :3010          No :2411      Min.   : 0.000      Min.   :  0.0   Min.   :  0.0   Min.   :  0.0   Min.   :  0.0
 Yes:1695         Credit card (automatic)  : 704   Yes: 323          Yes: 922      1st Qu.: 0.000      1st Qu.:143.7   1st Qu.: 87.0   1st Qu.:166.6   1st Qu.: 87.0
                  Electronic check         : 887                                 Median : 0.000      Median :179.4   Median :101.0   Median :201.4   Median :100.0
                  Mailed check             :1048                                 Mean   : 8.099      Mean   :179.8   Mean   :100.4   Mean   :201.0   Mean   :100.1
                                                                                 3rd Qu.:20.000      3rd Qu.:216.4   3rd Qu.:114.0   3rd Qu.:235.3   3rd Qu.:114.0
                                                                                 Max.   :51.000      Max.   :350.8   Max.   :165.0   Max.   :363.7   Max.   :170.0
 TotalNightMinutes TotalNightCalls TotalIntlMinutes TotalIntlCalls   CustomerServiceCalls   TotalCall       TotalRevenue     Churn
 Min.   : 23.2     Min.   : 33.0   Min.   : 0.00    Min.   : 0.000   Min.   :0.000        Min.   :194.0   Min.   :  18.8   No :2850
 1st Qu.:167.0     1st Qu.: 87.0   1st Qu.: 8.50    1st Qu.: 3.000   1st Qu.:1.000        1st Qu.:284.0   1st Qu.: 252.6   Yes: 483
 Median :201.2     Median :100.0   Median :10.30    Median : 4.000   Median :1.000        Median :307.0   Median : 892.5
 Mean   :200.9     Mean   :100.1   Mean   :10.24    Mean   : 4.479   Mean   :1.563        Mean   :306.7   Mean   :1673.3
 3rd Qu.:235.3     3rd Qu.:113.0   3rd Qu.:12.10    3rd Qu.: 6.000   3rd Qu.:2.000        3rd Qu.:330.0   3rd Qu.:2433.9
 Max.   :395.0     Max.   :175.0   Max.   :20.00    Max.   :20.000   Max.   :9.000        Max.   :418.0   Max.   :8476.5
                                                                                                          NA's   :5
```

Gives me the contents and breakdown for the categorical fields

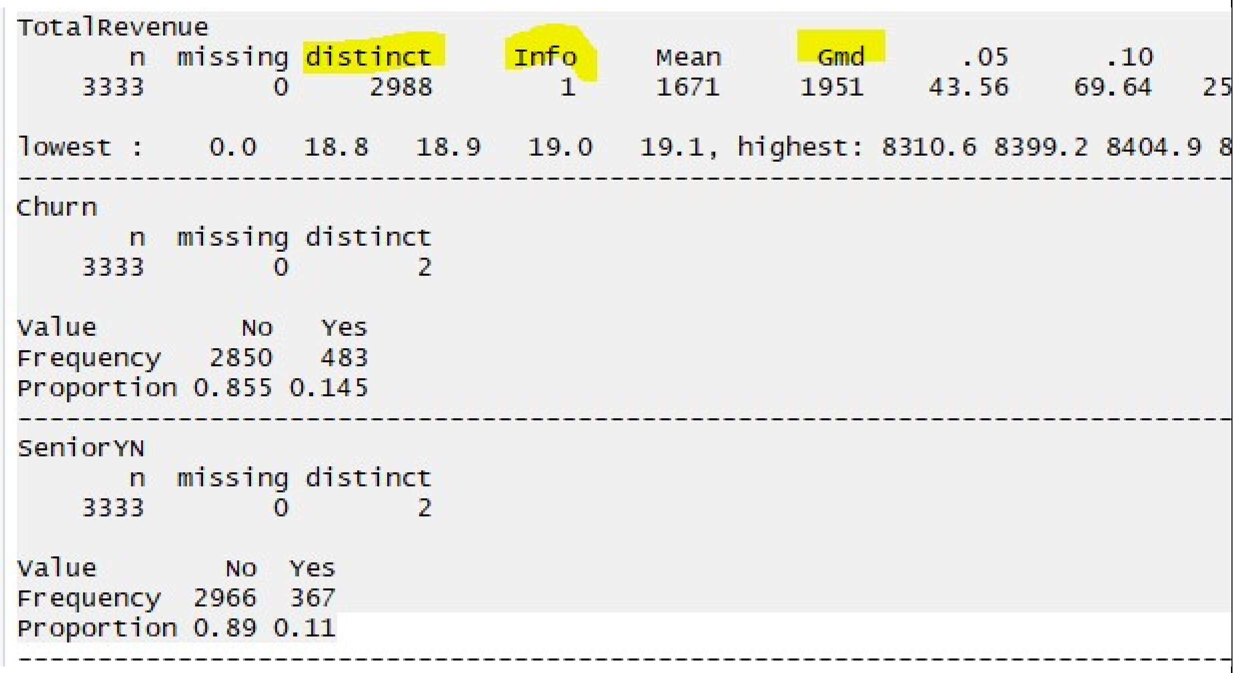
Gives me some interesting stats like Min,Max and quartiles on the continuous fields

Removed Nulls from Total revenue manually since there were only 5 of them. I set them to 0.

```
str(df_TCD)
'data.frame':   3333 obs. of  33 variables:
 $ customerID          : Factor w/ 3333 levels "0002-ORFBO","0004-TLHLJ",..
: 1 2 3 4 5 6 7 8 9 10 ...
 $ gender              : Factor w/ 2 levels "Female","Male": 1 2 1 1 1 1 1
2 1 1 ...
 $ SeniorCitizen       : int  0 0 0 1 1 0 1 1 0 0 ...
 $ MaritalStatus       : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 2 1 1 1 2
...
 $ Dependents          : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 2
...
 $ tenure              : int  9 4 9 71 7 5 1 45 3 4 ...
 $ PhoneService        : Factor w/ 1 level "Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ MultipleLines       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 1
...
 $ InternetService     : Factor w/ 3 levels "DSL","Fiber optic",..: 1 2 1 2
1 2 2 1 3 3 ...
 $ OnlineSecurity      : Factor w/ 3 levels "No","No internet service",..:
1 1 1 3 3 1 1 3 2 2 ...
 $ OnlineBackup        : Factor w/ 3 levels "No","No internet service",..:
3 1 1 3 1 1 1 1 2 2 ...
 $ DeviceProtection    : Factor w/ 3 levels "No","No internet service",..:
1 3 1 3 1 1 1 3 2 2 ...
 $ TechSupport         : Factor w/ 3 levels "No","No internet service",..:
3 1 3 3 1 1 1 1 2 2 ...
 $ StreamingTV         : Factor w/ 3 levels "No","No internet service",..:
3 1 3 3 1 1 1 1 2 2 ...
 $ StreamingMovies     : Factor w/ 3 levels "No","No internet service",..:
1 1 3 3 1 1 1 3 2 2 ...
 $ Contract            : Factor w/ 3 levels "Month-to-month",..: 2 1 1 3 1
1 1 2 1 1 ...
```

```
 $ PaperlessBilling    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 1 1 1
...
 $ PaymentMethod       : Factor w/ 4 levels "Bank transfer (automatic)",..:
4 3 2 1 3 3 3 2 4 4 ...
 $ InternationalPlan   : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 1 1
...
 $ VoiceMailPlan       : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 1
...
 $ NumbervMailMessages : int  0 0 36 0 0 0 0 0 0 0 ...
 $ TotalDayMinutes     : num  168.8 122.2 178.7 190.2 67.7 ...
 $ TotalDayCalls       : int  137 112 134 68 68 95 55 133 158 99 ...
 $ TotalEveMinutes     : num  241 132 179 262 196 ...
 $ TotalEveCalls       : int  107 94 102 64 86 128 124 86 120 93 ...
 $ TotalNightMinutes   : num  205 170 127 130 236 ...
 $ TotalNightCalls     : int  106 106 82 92 137 105 81 80 46 106 ...
 $ TotalIntlMinutes    : num  15.5 10.3 8 8.8 12 12.9 10 11.5 12.4 8 ...
 $ TotalIntlCalls      : int  4 9 4 4 2 5 7 3 3 4 ...
 $ CustomerServiceCalls: int  0 5 2 0 1 3 3 0 1 1 ...
 $ TotalCall           : int  354 326 324 228 294 336 270 302 328 303 ...
 $ TotalRevenue        : num  593 281 572 7904 340 ...
 $ Churn               : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 2 1 2
...
```

```
describe(df_TCD) # advantage over summary()? Shows distinct,missing,descrip
tive

Sample:
```



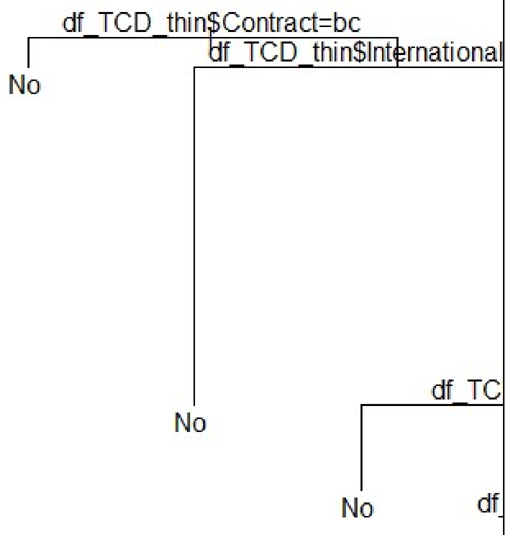
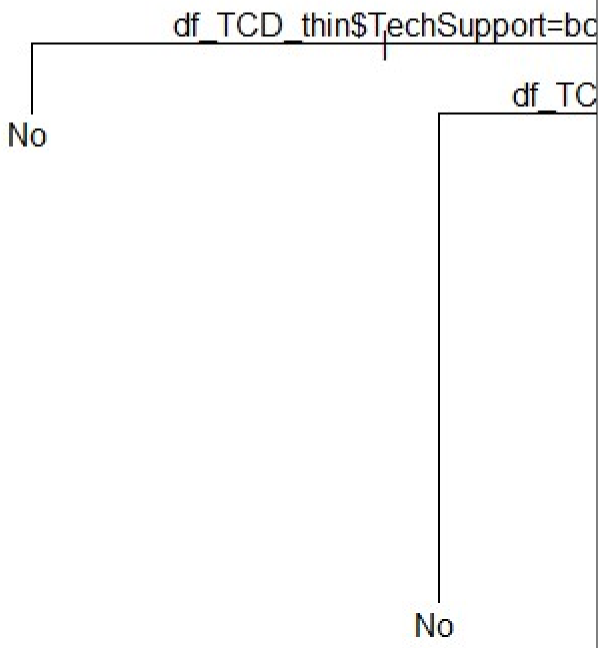
```
TotalRevenue
       n  missing distinct     Info     Mean      Gmd      .05      .10      25
    3333        0     2988        1     1671     1951    43.56    69.64    25

lowest :    0.0   18.8   18.9   19.0   19.1, highest: 8310.6 8399.2 8404.9 8
---------------------------------------------------------------------------------
Churn
       n  missing distinct
    3333        0        2

Value         No   Yes
Frequency   2850   483
Proportion 0.855 0.145
---------------------------------------------------------------------------------
SeniorYN
       n  missing distinct
    3333        0        2

Value        No  Yes
Frequency  2966  367
Proportion 0.89 0.11
---------------------------------------------------------------------------------
```

```
var(df_TCD_thin) #shows variance within the data
```

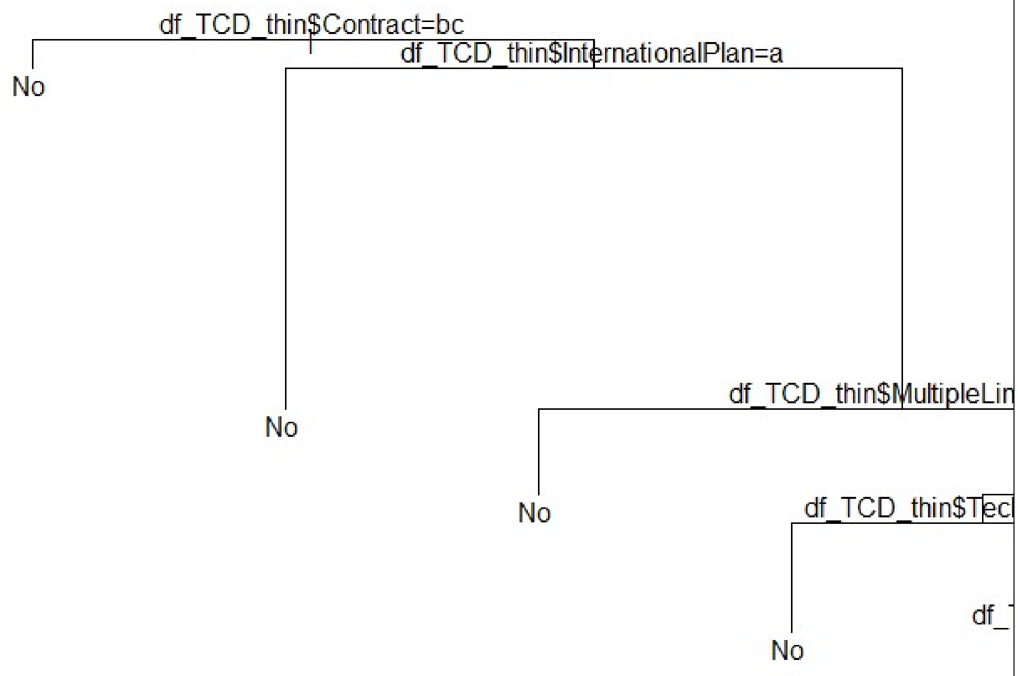
Data cleansing needed

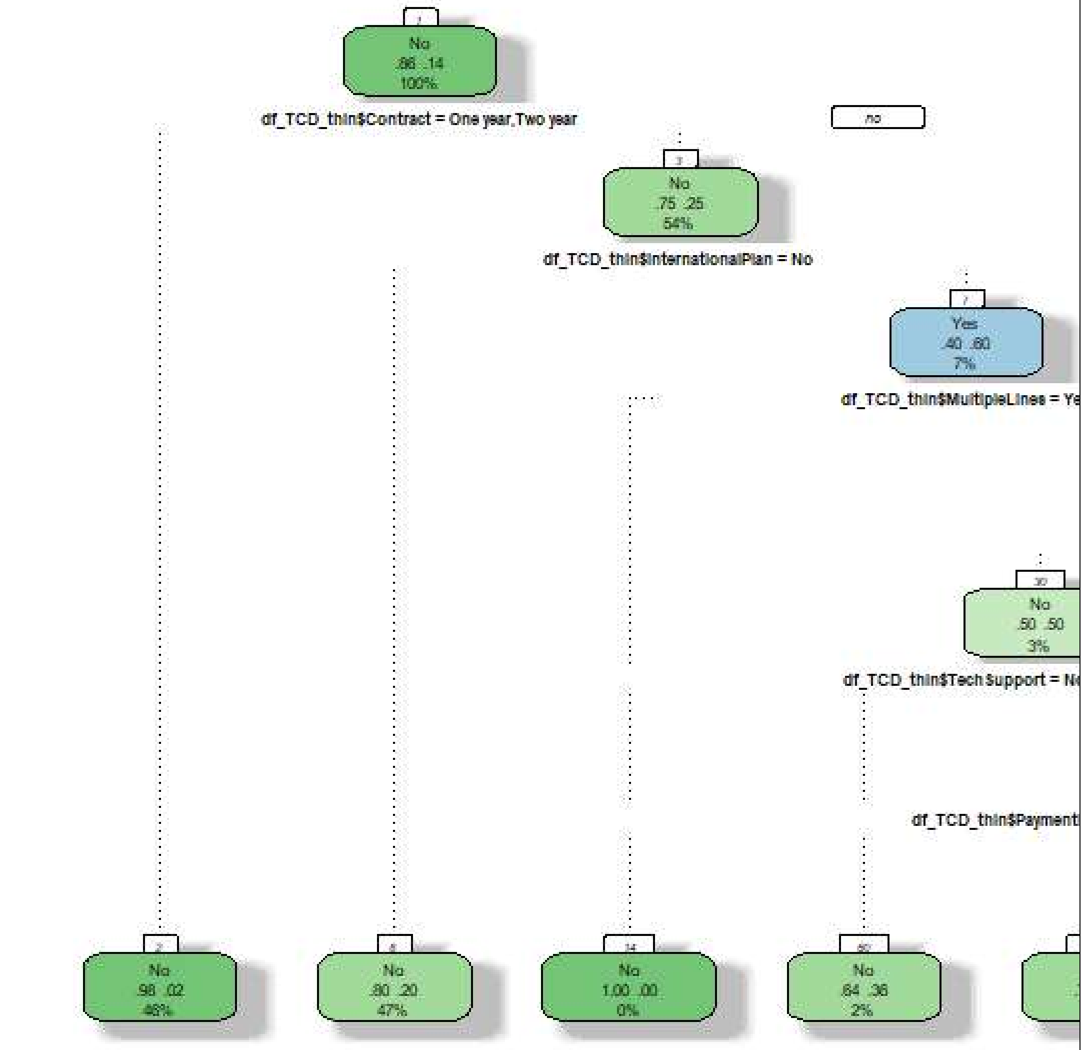
1. **SeniorCitizen** is stored as 1 or 0, I would like that to change to Yes or No to fit with rest of Y/N attributes like Marital status, Children etc.
2. **PhoneService** is always set to know which is possibly responsible for my decision tree error of needing more factor levels. I am going to remove it and try.
3. **TotalRevenue** has some Null values which I need to replace with zeros or averages.

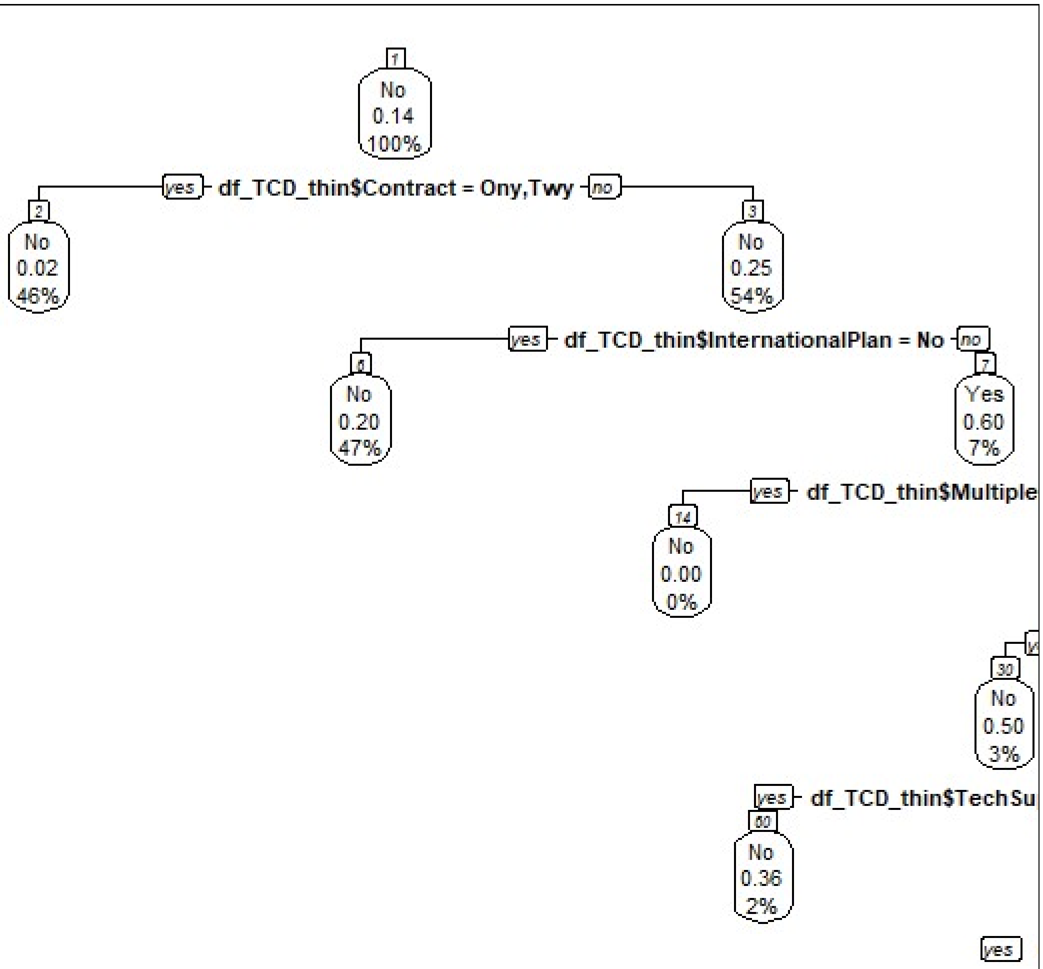
Simple tree comparison, playing around

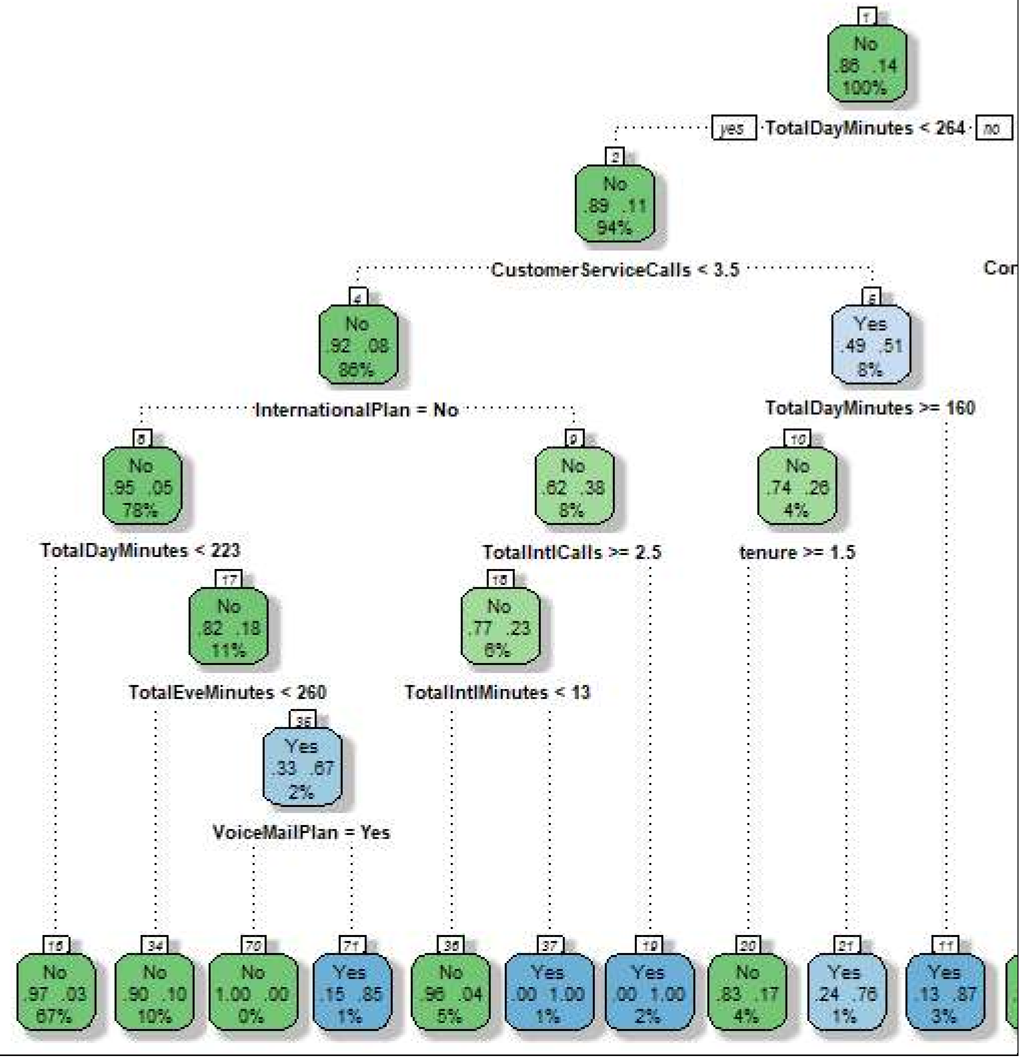
Splits on Contract	Removed Contract
$SeniorYN+$gender+$MaritalStatus+$Dependents+$MultipleLines+$InternetService+$OnlineSecurity +$OnlineBackup+$DeviceProtection+$TechSupport+$StreamingTV+$StreamingMovies+$Contract +$PaperlessBilling+$PaymentMethod+$InternationalPlan+$VoiceMailPlan	Splits on Tech support
 df_TCD_thin$Contract=bc df_TCD_thin$International No No df_TC No df_	

Simple tree results

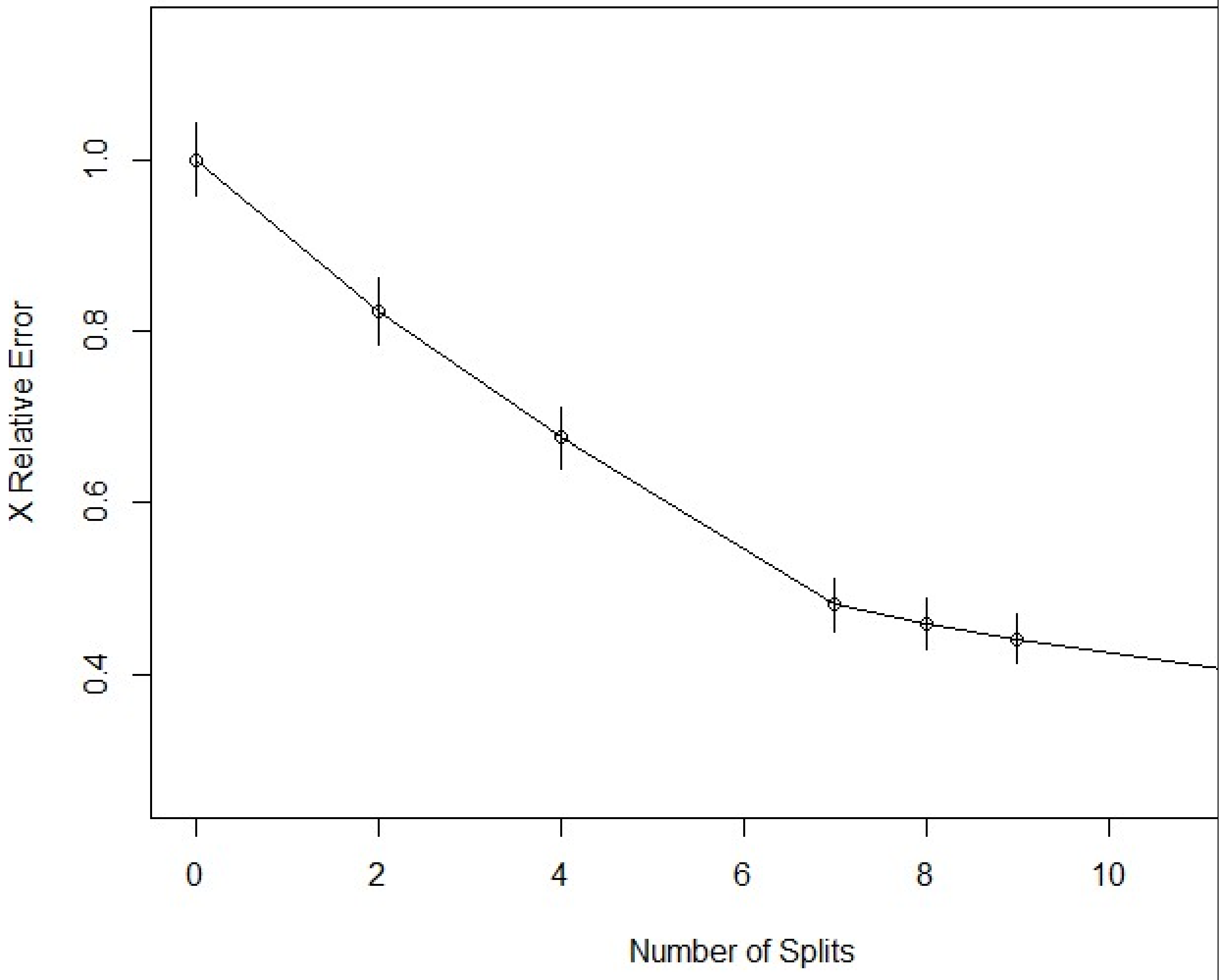
From SimpleTree_TCD.R	Tree plot output plot(fit) and text(fit)
Using rpart	 df_TCD_thin$Contract=bc df_TCD_thin$InternationalPlan=a No No df_TCD_thin$MultipleLin No df_TCD_thin$Tecl df_ No

Using the fancyRpart Plot	 1 No .86 .14 100% df_TCD_thin$Contract = One year,Two year no 3 No .75 .25 54% df_TCD_thin$InternationalPlan = No 7 Yes .40 .60 7% df_TCD_thin$MultipleLines = Ye 30 No .50 .50 3% df_TCD_thin$TechSupport = N df_TCD_thin$Payment 2 No .98 .02 46% 6 No .80 .20 47% 14 No 1.00 .00 0% 60 No .64 .36 2%

Using prp(fit) part of rpart	 1 No 0.14 100% yes df_TCD_thin$Contract = Ony,Twy no 2 No 0.02 46% 3 No 0.25 54% yes df_TCD_thin$InternationalPlan = No no 6 No 0.20 47% 7 Yes 0.60 7% yes df_TCD_thin$Multiple 14 No 0.00 0% 30 No 0.50 3% yes df_TCD_thin$TechSu 60 No 0.36 2% yes

<table>
<tr>
<td>Using all the data for decision tree</td>
<td>


</td>
</tr>
</table>



rsq.rpart(fit_all) #major improvements end after 7 splits

```
> summary(fit_all)
Call:
rpart(formula = Churn ~ SeniorYN + gender + MaritalStatus + Dependents +
    tenure + PhoneService + MultipleLines + InternetService +
    OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +
    StreamingTV + StreamingMovies + Contract + PaperlessBilling +
    PaymentMethod + InternationalPlan + VoiceMailPlan + NumbervMailMessages
+
    TotalDayMinutes + TotalDayCalls + TotalEveMinutes + TotalEveCalls +
    TotalNightMinutes + TotalNightCalls + TotalIntlMinutes +
    TotalIntlCalls + CustomerServiceCalls + TotalCall + TotalRevenue,
    data = df_TCD_thin, method = "class")
  n= 3333

          CP nsplit rel error    xerror       xstd
1 0.09316770      0 1.0000000 1.0000000 0.04207569
```

```
2 0.07867495      2 0.8136646 0.8219462 0.03871761
3 0.05279503      4 0.6563147 0.6749482 0.03550673
4 0.02277433      7 0.4616977 0.4803313 0.03041796
5 0.01863354      8 0.4389234 0.4575569 0.02974070
6 0.01759834      9 0.4202899 0.4409938 0.02923495
7 0.01000000     12 0.3623188 0.3954451 0.02778145

Variable importance
     TotalDayMinutes CustomerServiceCalls     TotalIntlMinutes    Internati
onalPlan       TotalIntlCalls               tenure
                  24                   13                   10
8                   8                    7
            Contract         TotalRevenue      TotalEveMinutes  NumbervMail
Messages         VoiceMailPlan       OnlineSecurity
                   6                    5                    4
4                   4                    2
     InternetService         OnlineBackup      TotalNightCalls    TotalNigh
tMinutes
                   1                    1                    1
1

Node number 1: 3333 observations,    complexity param=0.0931677
  predicted class=No   expected loss=0.1449145  P(node) =1
    class counts:  2850   483
   probabilities: 0.855 0.145
  left son=2 (3122 obs) right son=3 (211 obs)
  Primary splits:
      TotalDayMinutes      < 264.45  to the left,  improve=94.08310, (0 mis
sing)
      Contract             splits as  RLL,         improve=86.76350, (0 mis
sing)
      CustomerServiceCalls < 3.5     to the left,  improve=80.30617, (0 mis
sing)
      tenure               < 5.5     to the right, improve=80.25095, (0 mis
sing)
      TechSupport          splits as  RLL,         improve=71.85318, (0 mis
sing)
```