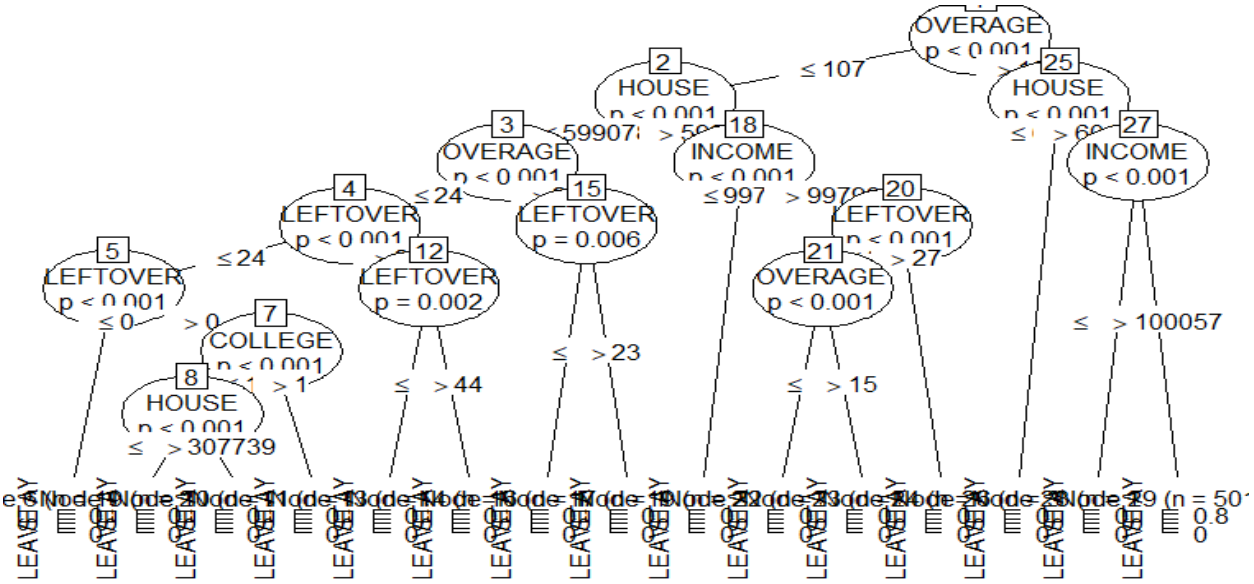
Decision Tree

The BondTelco Company need to determine the churn rate of their clients, to achieve this a data set have been provided by their IT staffs to help make the decision. The data has 20,000 rows and the 12 columns including the outcome variable (LEAVE). The outcome variable was converted to a factor variable and the remaining variables converted to numeric variables. Since it is very difficult to know the variables that would best explain the model, all the variables were included in the model. A control criterion was set to a minimum of 99 %. Below are the codes and output.

```
> cat('Importing the data\n')
Importing the data
>
> data = read.csv(file.choose(), header = T)
> cat('Checking the data set')
Checking the data set> head(data)
  COLLEGE INCOME OVERAGE LEFTOVER  HOUSE HANDSET_PRICE OVER_15MINS_CALLS_PER_MONTH AVERAGE
RATION
1    zero  40385      65       23 453600           216                            3
5
2    zero  43915     158       15 151890           197                           24
5
3    zero  70863     186        9 705316           546                           19
5
4     one  27886      63       63 461456           241                            1
2
5    zero  31556      71       76 324804           195                           15
1
6    zero  84992     197        8 736073           396                            1
4
  REPORTED_SATISFACTION REPORTED_USAGE_LEVEL CONSIDERING_CHANGE_OF_PLAN LEAVE
1              very_sat               little                        no  STAY
2            very_unsat            very_high                   perhaps LEAVE
3                   sat            very_high                        no  STAY
4            very_unsat               little                considering  STAY
5                 unsat               little             never_thought LEAVE
6                 unsat          very_little                        no  STAY
> cat('Checking the structure if the data\n')
Checking the structure if the data
>
> str(data)
'data.frame':   20000 obs. of  12 variables:
 $ COLLEGE                   : Factor w/ 2 levels "one","zero": 2 2 2 1 2 2 2 2 2 1 ...
 $ INCOME                    : int  40385 43915 70863 27886 31556 84992 63466 55347 24146
..
 $ OVERAGE                   : int  65 158 186 63 71 197 67 59 203 0 ...
 $ LEFTOVER                  : int  23 15 9 63 76 8 0 0 8 59 ...
 $ HOUSE                     : int  453600 151890 705316 461456 324804 736073 728459 9224
1 693081 ...
 $ HANDSET_PRICE             : int  216 197 546 241 195 396 254 331 332 225 ...
 $ OVER_15MINS_CALLS_PER_MONTH: int  3 24 19 1 15 1 4 13 25 0 ...
 $ AVERAGE_CALL_DURATION     : int  5 5 5 2 1 4 14 15 6 2 ...
```

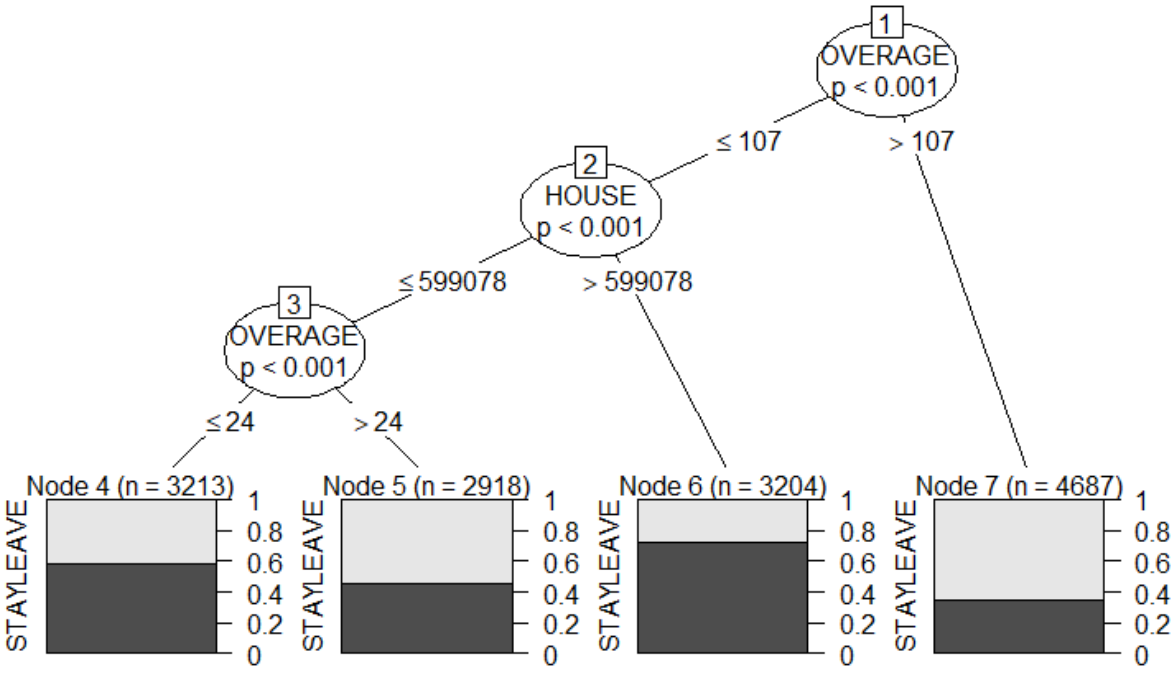
```
 $ REPORTED_SATISFACTION      : Factor w/ 5 levels "avg","sat","unsat",..: 4 5 2 5 3 3 4 4
 $ REPORTED_USAGE_LEVEL       : Factor w/ 5 levels "avg","high","little",..: 3 4 4 3 3 5 3
..
 $ CONSIDERING_CHANGE_OF_PLAN : Factor w/ 5 levels "actively_looking_into_it",..: 4 5 4 2
1 2 ...
 $ LEAVE                      : Factor w/ 2 levels "LEAVE","STAY": 2 1 2 2 1 2 2 2 1 1 ...
> library(dplyr)
>
> head(data)
  COLLEGE INCOME OVERAGE LEFTOVER  HOUSE HANDSET_PRICE OVER_15MINS_CALLS_PER_MONTH AVERAGE
RATION
1    zero  40385      65       23 453600           216                            3
5
2    zero  43915     158       15 151890           197                           24
5
3    zero  70863     186        9 705316           546                           19
5
4     one  27886      63       63 461456           241                            1
2
5    zero  31556      71       76 324804           195                           15
1
6    zero  84992     197        8 736073           396                            1
4
  REPORTED_SATISFACTION REPORTED_USAGE_LEVEL CONSIDERING_CHANGE_OF_PLAN LEAVE
1              very_sat               little                        no  STAY
2            very_unsat            very_high                   perhaps LEAVE
3                   sat            very_high                        no  STAY
4            very_unsat               little                considering  STAY
5                 unsat               little             never_thought LEAVE
6                 unsat          very_little                        no  STAY
> cat('Converting the variables to numeric')
Converting the variables to numeric> data$COLLEGE = as.numeric(data$COLLEGE)
> data$INCOME = as.numeric(data$INCOME)
> data$OVERAGE = as.numeric(data$OVERAGE)
> data$LEFTOVER = as.numeric(data$LEFTOVER)
> data$HOUSE= as.numeric(data$HOUSE)
> data$HANDSET_PRICE = as.numeric(data$HANDSET_PRICE)
> data$OVER_15MINS_CALLS_PER_MONTH = as.numeric(data$OVER_15MINS_CALLS_PER_MONTH)
> data$AVERAGE_CALL_DURATION = as.numeric(data$AVERAGE_CALL_DURATION)
> data$REPORTED_SATISFACTION = as.numeric(data$REPORTED_SATISFACTION)
> data$REPORTED_USAGE_LEVEL = as.numeric(data$REPORTED_USAGE_LEVEL)
> data$CONSIDERING_CHANGE_OF_PLAN = as.numeric(data$CONSIDERING_CHANGE_OF_PLAN)
> data$LEAVE = data$LEAVE
> cat('PArtitioning data into training and testing')
PArtitioning data into training and testing> set.seed(1234)
> ind= sample(2, nrow(data), replace = T, prob = c(0.7,0.3))
> trainData = data[ind==1,]
> testData = data[ind==2,]
> cat('Building a decision tree')
Building a decision tree> library(party)
> myTree = ctree(LEAVE~., data= trainData, controls = ctree_control(mincriterion = 0.99, m
=500 ))
> myTree

         Conditional inference tree with 15 terminal nodes

Response:  LEAVE
Inputs:  COLLEGE, INCOME, OVERAGE, LEFTOVER, HOUSE, HANDSET_PRICE, OVER_15MINS_CALLS_PER_M
ERAGE_CALL_DURATION, REPORTED_SATISFACTION, REPORTED_USAGE_LEVEL, CONSIDERING_CHANGE_OF_PL
Number of observations:  14022

1) OVERAGE <= 107; criterion = 1, statistic = 783.133
  2) HOUSE <= 599078; criterion = 1, statistic = 219.846
    3) OVERAGE <= 24; criterion = 1, statistic = 83.231
```

```
      4) LEFTOVER <= 24; criterion = 1, statistic = 43.033
        5) LEFTOVER <= 0; criterion = 1, statistic = 270.5
          6)*  weights = 1067
        5) LEFTOVER > 0
          7) COLLEGE <= 1; criterion = 1, statistic = 36.446
            8) HOUSE <= 307739; criterion = 1, statistic = 46.315
              9)*  weights = 262
            8) HOUSE > 307739
              10)*  weights = 295
          7) COLLEGE > 1
            11)*  weights = 536
      4) LEFTOVER > 24
        12) LEFTOVER <= 44; criterion = 0.998, statistic = 14.33
          13)*  weights = 314
        12) LEFTOVER > 44
          14)*  weights = 739
    3) OVERAGE > 24
      15) LEFTOVER <= 23; criterion = 0.994, statistic = 11.954
        16)*  weights = 1884
      15) LEFTOVER > 23
        17)*  weights = 1034
  2) HOUSE > 599078
    18) INCOME <= 99793; criterion = 1, statistic = 212.799
      19)*  weights = 2126
    18) INCOME > 99793
      20) LEFTOVER <= 27; criterion = 1, statistic = 27.211
        21) OVERAGE <= 15; criterion = 1, statistic = 21.667
          22)*  weights = 356
        21) OVERAGE > 15
          23)*  weights = 359
      20) LEFTOVER > 27
        24)*  weights = 363
1) OVERAGE > 107
  25) HOUSE <= 601673; criterion = 1, statistic = 593.974
    26)*  weights = 3093
  25) HOUSE > 601673
    27) INCOME <= 100057; criterion = 1, statistic = 398.734
      28)*  weights = 1093
    27) INCOME > 100057
      29)*  weights = 501
> cat('Plotting the decision Tree')
Plotting the decision Tree> plot(myTree)

>
```



The figure above shows that OVERAGE AND HOUSE were the best variables to be used in the model. To confirm this, restriction were added on the model. Below is the results obtained

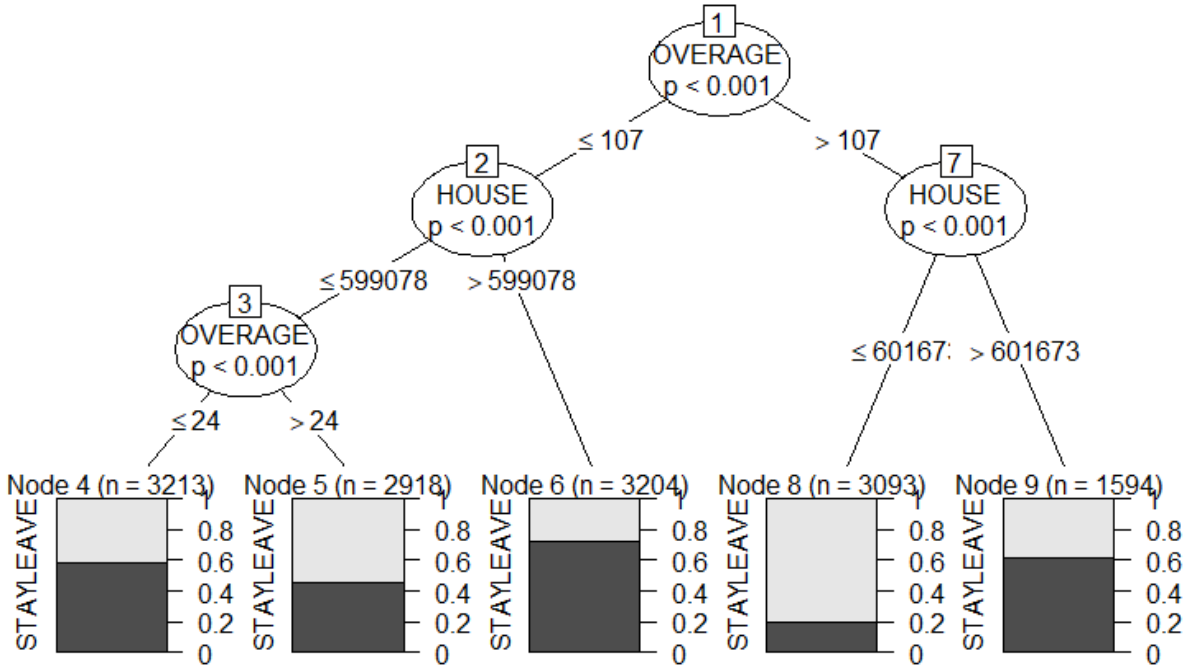
```
> cat('changing the restrictions')
changing the restrictions> myTree = ctree(LEAVE~., data= trainData, controls
= ctree_control(mincriterion = 0.99, minsplit =5000 ))
> #myTree
> plot(myTree)
```



The figure below confirms our statement. OVERAGE and HOUSE were the best variables for the prediction. Therefore, we can subset the data and select the two variables for the model.

```
> # importing the library to be used to subset the data
> library(dplyr)
>
> # Converting the data into the required data type
> ddt = select(data, LEAVE, HOUSE, OVERAGE)
> ddt$LEAVE = as.factor(ddt$LEAVE)
> ddt$HOUSE= as.numeric(ddt$HOUSE)
> ddt$OVERAGE = as.numeric(ddt$OVERAGE)
> # partitioning into training and testing
> set.seed(1234)
> ind= sample(2, nrow(ddt), replace = T, prob = c(0.7,0.3))
> trainData = ddt[ind==1,]
> testData = ddt[ind==2,]
> # Building a decision tree
> library(party)
> myTree = ctree(LEAVE~., data= trainData, controls = ctree_control(mincriter
ion = 0.99, minsplit =4500 ))
> myTree

         Conditional inference tree with 5 terminal nodes

Response:  LEAVE
Inputs:  HOUSE, OVERAGE
Number of observations:  14022
```

```
1) OVERAGE <= 107; criterion = 1, statistic = 783.133
  2) HOUSE <= 599078; criterion = 1, statistic = 219.846
    3) OVERAGE <= 24; criterion = 1, statistic = 83.231
      4)*  weights = 3213
    3) OVERAGE > 24
      5)*  weights = 2918
  2) HOUSE > 599078
    6)*  weights = 3204
1) OVERAGE > 107
  7) HOUSE <= 601673; criterion = 1, statistic = 593.974
    8)*  weights = 3093
  7) HOUSE > 601673
    9)*  weights = 1594
> plot(myTree)
```



This is the final prediction for the model.