Logistic Regression & Decision Trees: Additional Examples with Answers

Example 1: Predicting Annual Income of Individuals

In this example, we will use logistic regression to predict if an individual earns more than $50k in a year or not. First read the 'Adult' CSV file using the following command (you need to be connected to the internet and it will take few moments for data to be loaded):

inputData <- read.csv("http://rstatistics.net/wp-content/uploads/2015/09/adult.csv")

Let's have a look at the first 6 records and also a summary of the dataset

```
> inputData <- read.csv("http://rstatistics.net/wp-content/uploads/2015/09/adult.csv")
> head(inputData)
  AGE        WORKCLASS FNLWGT  EDUCATION EDUCATIONNUM      MARITALSTATUS
1  39        State-gov  77516  Bachelors           13      Never-married
2  50 Self-emp-not-inc  83311  Bachelors           13 Married-civ-spouse
3  38          Private 215646    HS-grad            9           Divorced
4  53          Private 234721       11th            7 Married-civ-spouse
5  28          Private 338409  Bachelors           13 Married-civ-spouse
6  37          Private 284582    Masters           14 Married-civ-spouse
         OCCUPATION   RELATIONSHIP  RACE     SEX CAPITALGAIN CAPITALLOSS HOURSPERWEEK
1      Adm-clerical  Not-in-family White    Male        2174           0           40
2   Exec-managerial        Husband White    Male           0           0           13
3 Handlers-cleaners  Not-in-family White    Male           0           0           40
4 Handlers-cleaners        Husband Black    Male           0           0           40
5    Prof-specialty           Wife Black  Female           0           0           40
6   Exec-managerial           Wife White  Female           0           0           40
  NATIVECOUNTRY ABOVE50K
1 United-States        0
2 United-States        0
3 United-States        0
4 United-States        0
5          Cuba        0
6 United-States        0
```

And

```
> summary(inputData)
      AGE                    WORKCLASS          FNLWGT                 EDUCATION    
 Min.   :17.00   Private         :22696   Min.   :  12285   HS-grad     :10501  
 1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827   Some-college: 7291  
 Median :37.00   Local-gov       : 2093   Median : 178356   Bachelors   : 5355  
 Mean   :38.58   ?               : 1836   Mean   : 189778   Masters     : 1723  
 3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051   Assoc-voc   : 1382  
 Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705   11th        : 1175  
                (Other)          :  981                    (Other)      : 5134  
  EDUCATIONNUM                  MARITALSTATUS              OCCUPATION   
 Min.   : 1.00   Divorced             : 4443   Prof-specialty :4140  
 1st Qu.: 9.00   Married-AF-spouse    :   23   Craft-repair   :4099  
 Median :10.00   Married-civ-spouse   :14976   Exec-managerial:4066  
 Mean   :10.08   Married-spouse-absent:  418   Adm-clerical   :3770  
 3rd Qu.:12.00   Never-married        :10683   Sales          :3650  
 Max.   :16.00   Separated            : 1025   Other-service  :3295  
                 Widowed              :  993  (Other)         :9541  
          RELATIONSHIP                    RACE              SEX          CAPITALGAIN   
  Husband       :13193   Amer-Indian-Eskimo:  311   Female:10771   Min.   :    0  
  Not-in-family : 8305   Asian-Pac-Islander: 1039   Male  :21790   1st Qu.:    0  
  Other-relative:  981   Black             : 3124                  Median :    0  
  Own-child     : 5068   Other             :  271                  Mean   : 1078  
  Unmarried     : 3446   White             :27816                  3rd Qu.:    0  
  Wife          : 1568                                             Max.   :99999  

  CAPITALLOSS      HOURSPERWEEK          NATIVECOUNTRY       ABOVE50K     
 Min.   :   0.0   Min.   : 1.00   United-States:29170   Min.   :0.0000  
 1st Qu.:   0.0   1st Qu.:40.00   Mexico       :  643   1st Qu.:0.0000  
 Median :   0.0   Median :40.00   ?            :  583   Median :0.0000  
 Mean   :  87.3   Mean   :40.44   Philippines  :  198   Mean   :0.2408  
 3rd Qu.:   0.0   3rd Qu.:45.00   Germany      :  137   3rd Qu.:0.0000  
 Max.   :4356.0   Max.   :99.00   Canada       :  121   Max.   :1.0000  
                                 (Other)       : 1709                   
```

The variable 'ABOVE50K' is the variable that we are trying to predict. Currently, the variable is coded as a numeric variable that takes 0 and 1s. To use logistic regression, that is a classification method, we need to convert this variable to a factor (i.e. categorical variable):

inputData$ABOVE50K=as.factor(inputData$ABOVE50K)

let's look at the summary again:

```
> inputData$ABOVE50K=as.factor(inputData$ABOVE50K)
> summary(inputData)
      AGE                   WORKCLASS           FNLWGT                 EDUCATION
 Min.   :17.00   Private          :22696   Min.   :  12285   HS-grad     :10501
 1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827   Some-college: 7291
 Median :37.00   Local-gov        : 2093   Median : 178356   Bachelors   : 5355
 Mean   :38.58   ?                : 1836   Mean   : 189778   Masters     : 1723
 3rd Qu.:48.00   State-gov        : 1298   3rd Qu.: 237051   Assoc-voc   : 1382
 Max.   :90.00   Self-emp-inc     : 1116   Max.   :1484705   11th        : 1175
                (Other)          :  981                     (Other)     : 5134
  EDUCATIONNUM               MARITALSTATUS              OCCUPATION              RELATIONSHIP
 Min.   : 1.00   Divorced               : 4443   Prof-specialty :4140   Husband       :13193
 1st Qu.: 9.00   Married-AF-spouse      :   23   Craft-repair   :4099   Not-in-family : 8305
 Median :10.00   Married-civ-spouse     :14976   Exec-managerial:4066   Other-relative:  981
 Mean   :10.08   Married-spouse-absent:  418   Adm-clerical   :3770   Own-child     : 5068
 3rd Qu.:12.00   Never-married          :10683   Sales          :3650   Unmarried     : 3446
 Max.   :16.00   Separated              : 1025   Other-service  :3295   Wife          : 1568
                 Widowed                :  993   (Other)        :9541
                 RACE                SEX            CAPITALGAIN      CAPITALLOSS      HOURSPERWEEK
  Amer-Indian-Eskimo:  311   Female:10771   Min.   :     0   Min.   :   0.0   Min.   : 1.00
  Asian-Pac-Islander: 1039   Male  :21790   1st Qu.:     0   1st Qu.:   0.0   1st Qu.:40.00
  Black             : 3124                  Median :     0   Median :   0.0   Median :40.00
  Other             :  271                  Mean   : 1078   Mean   :  87.3   Mean   :40.44
  White             :27816                  3rd Qu.:     0   3rd Qu.:   0.0   3rd Qu.:45.00
                                            Max.   :99999   Max.   :4356.0   Max.   :99.00

        NATIVECOUNTRY   ABOVE50K
  United-States:29170   0:24720
  Mexico       :  643   1: 7841
  ?            :  583
  Philippines  :  198
  Germany      :  137
  Canada       :  121
```

Now let us build a logistic regression model based on AGE, EDUCATIONNUM (Number of years of Education), SEX and HOURSPERWEEK (number of hours worked per week) to predict ABOVE50K.

Model<-glm(ABOVE50K~AGE+EDUCATIONNUM+SEX+HOURSPERWEEK,data=inputData,family = binomial)

```
> Model<-glm(ABOVE50K~AGE+EDUCATIONNUM+SEX+HOURSPERWEEK,data=inputData,family=binomial)
> summary(Model)

Call:
glm(formula = ABOVE50K ~ AGE + EDUCATIONNUM + SEX + HOURSPERWEEK,
    family = binomial, data = inputData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6871  -0.6670  -0.4117  -0.1096   3.2214

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.133399   0.115709  -78.93   <2e-16 ***
AGE           0.045604   0.001186   38.47   <2e-16 ***
EDUCATIONNUM  0.355114   0.006617   53.67   <2e-16 ***
SEX Male      1.161158   0.037694   30.80   <2e-16 ***
HOURSPERWEEK  0.035637   0.001293   27.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35948  on 32560  degrees of freedom
Residual deviance: 27917  on 32556  degrees of freedom
AIC: 27927

Number of Fisher Scoring iterations: 5
```

Which of these variable are statistically significant?

The z-value is very high resulting in very small p-values for coefficients of variables, implying that they are all statistically significant.

How does the probability of earning above $50k changes with these variables?

The coefficient is positive for all variables implying that the probability of earnings being above the $50k increases with all variables. As for SEX, the default value is Female (alphabetically before Male) which is used for the base model. If the SEX is male there will be an additional 1.16 added to the output (that is the logarithm of the odds of earning more than $50k)

Remember James and Hannah? Apparently, they have both accepted to the graduate program and met each other. Now they are now married! Given the information below, what is the probability that each of them is earning more than $50k a year?

Hannah: Female, 25 Year, 16 Years of Education, Working part time 25 hours a week
James: Male, 31 Year, 16 Years of Education, Working 44 hours a week

We create a new dataset, first record for Hannah and second record for James:

Data_to_Predict=data.frame(SEX=c(' Female' ,' Male'), AGE=c(25,31), EDUCATIONNUM=c(16,16), HOURSPERWEEK=c(25,44))

Now you should have a new dataframe called Data_to_Predict in your global environment area. Double click on it to see the content of it.

2 observations of 4 variables

	SEX	AGE	EDUCATIONNUM	HOURSPERWEEK
1	Female	25	16	25
2	Male	31	16	44

Import Dataset Clear

Global Environment

Data

Data_to_Pre.2 obs. of 4 variables

Let's try to predict now:

predict(Model,newdata = Data_to_Predict, type='response')

```
> predict(Model,newdata = Data_to_Predict, type='response')
        1         2 
0.1945785 0.6662708 
```

So the probability of Hannah (the first record) to earn more than $50k is 19% (perhaps her mother, Elizabeth, is now worried again!) and the probability that James earns more than $50k is 66%.

What is the accuracy of this model in terms of Area Under Curve (AUC) of ROC ?

library(pROC)
Predicted_Values<-predict(Model, newdata= inputData,type='response')
roc(inputData$ ABOVE50K, Predicted_Values)

If you are getting error by calling the pROC library, you need to install the package first that is :

install.packages('pROC')

```
> Model<-glm(ABOVE50K~AGE+EDUCATIONNUM+SEX+HOURSPERWEEK,data=inputData,family = binomial)
> library(pROC)
Type 'citation("pROC")' for a citation.

Attaching package: 'pROC'

The following objects are masked from 'package:stats':

    cov, smooth, var

> Predicted_Values<-predict(Model, newdata= inputData,type='response')
> roc(inputData$ ABOVE50K, Predicted_Values)

Call:
roc.default(response = inputData$ABOVE50K, predictor = Predicted_Values)

Data: Predicted_Values in 24720 controls (inputData$ABOVE50K 0) < 7841 cases (inputData$ABOVE50K 1).
Area under the curve: 0.8164
```

AUC is 0.81. The model is pretty accurate!

Now let's try to solve the same questions but using decision trees as a classification method (instead of logistic regression). Let's build a model first

```
> Model_2=rpart(ABOVE50K~AGE+EDUCATIONNUM+SEX+HOURSPERWEEK,data=inputData,method='class')
> summary(Model_2)
Call:
rpart(formula = ABOVE50K ~ AGE + EDUCATIONNUM + SEX + HOURSPERWEEK,
    data = inputData, method = "class")
  n= 32561

          CP nsplit rel error    xerror        xstd
1 0.05764571      0 1.0000000 1.0000000 0.009839876
2 0.01492157      3 0.8168601 0.8177528 0.009151746
3 0.01000000      4 0.8019385 0.8052544 0.009098542

Variable importance
EDUCATIONNUM          AGE          SEX HOURSPERWEEK
          62           23           12            4
```

How do we judge the statistical significance and the importance of variables in decision tree models?

In decision tree models, we do not have coefficients for variables so we cannot use z-test or t-test to check the importance of variables. The variable importance field gives a measure of importance and significance of variables at the same time. The variable importance values are usually normalized so that the sum of all variable importance to be 100. In this example, EDUCATIONNUM, that is the number of years of education, is by far the most important variable followed by AGE and SEX and finally by the HOURSPERWEEK, which represents the number of hours worked per week. The same order of variable importance was suggested by the logistic regression model above.

How does the probability of earning above \$50k changes with these variables?

We can answer this question by ploting the decision tree model. You can use the plot() function from R-base (no need for additional library), or use fancyRpartPlot() from the 'rattle' library which has a nicer presentation.

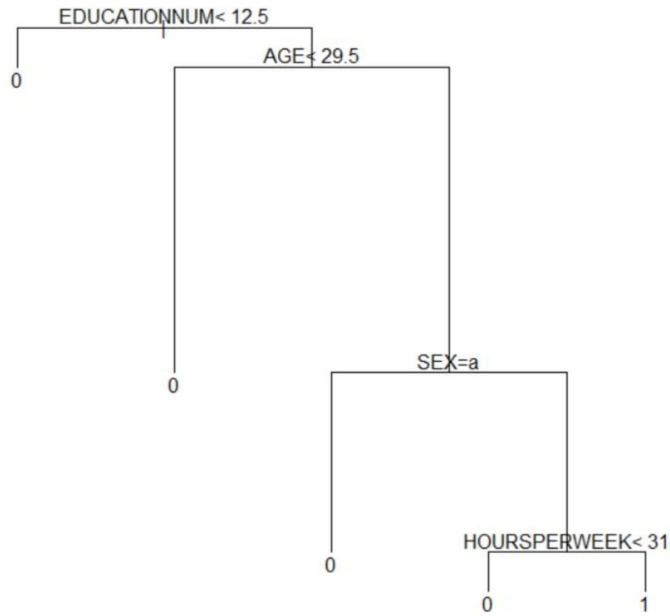
Using the plot() function:

plot(Model_2) #to plot the tree
text(Model_2) #to add labels

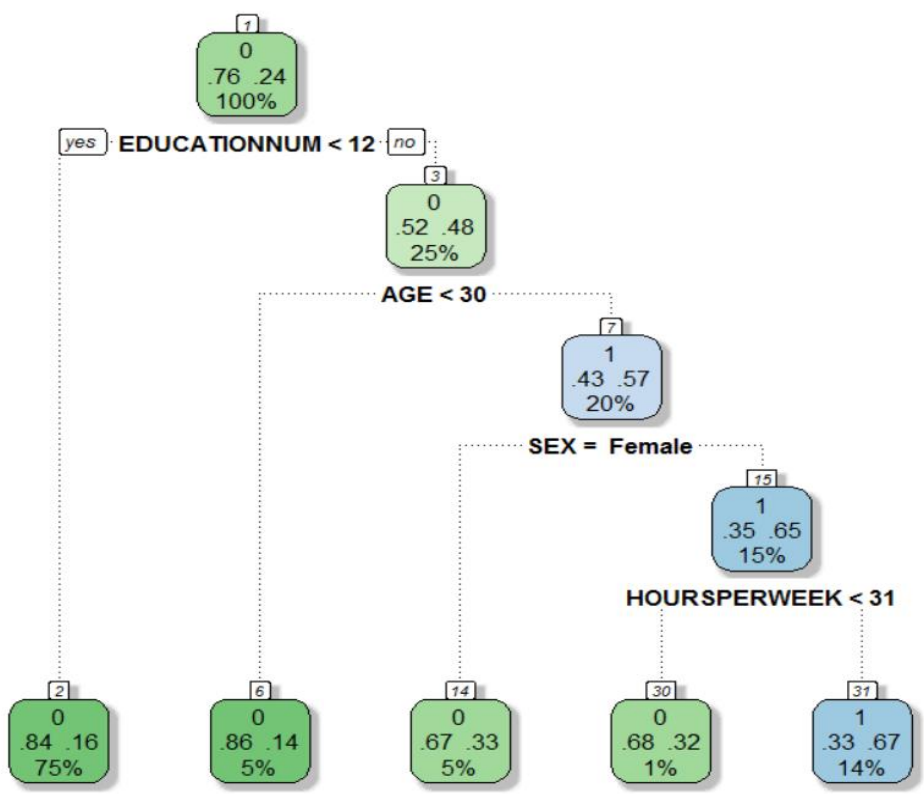
Using the fancyRpartPlot()

library(rattle)
fancyRpartPlot(Model_2)

With plot() function



With fancyRpartPlot()



Remember James and Hannah? Apparently, they have both accepted to the graduate program and met each other. Now they are now married! Given the information below, what is the probability that each of them is earning more than $50k a year?

Hannah: Female, 25 Year, 16 Years of Education, Working part time 25 hours a week
James: Male, 31 Year, 16 Years of Education, Working 44 hours a week

We create a new dataset, first record for Hannah and second record for James:

Data_to_Predict=data.frame(SEX=c(' Female' ,' Male'), AGE=c(25,31),

EDUCATIONNUM=c(16,16), HOURSPERWEEK=c(25,44))

predict(Model_2,newdata= Data_to_Predict,type='prob')

```
> Data_to_Predict=data.frame(SEX=c(' Female' ,' Male'), AGE=c(25,31),
+                             EDUCATIONNUM=c(16,16), HOURSPERWEEK=c(25,44))
>
> predict(Model_2,newdata= Data_to_Predict,type='prob')
          0         1
1 0.8565244 0.1434756
2 0.3278652 0.6721348
```

Unlike logistic regression model which provides the probability of the second level class (by default alphabetically), rpart decision tree models gives the probability of each class explicitly. In our example, the probability of 1 (i.e. earning above 50K) is 14.3% for the first observation (i.e. Hannah) and the probability of 1 (again earning abobr50K) for the second observation (i.e. James) is 67.2%. These values are somehow similar to what we got from the logistic regression: 19.4% for Hannah and 66.6% for James.

What is the accuracy of this model in terms of Area Under Curve (AUC) of ROC ?

library(pROC)

Predicted_Values<-predict(Model_2, newdata= inputData,type='prob')
roc(inputData$ ABOVE50K, Predicted_Values[,2])

```
> library(pROC)
> Predicted_Values<-predict(Model_2, newdata= inputData,type='prob')
> roc(inputData$ ABOVE50K, Predicted_Values[,2])

Call:
roc.default(response = inputData$ABOVE50K, predictor = Predicted_Values[,     2])

Data: Predicted_Values[, 2] in 24720 controls (inputData$ABOVE50K 0) < 7841 cases (inputData$ABOVE50K 1).
Area under the curve: 0.6878
>
```

We passed Predicted_Values[,2] to the predict function because the second column of the vector Predicted_Values contains probabilities for 1 (i.e. income above50K), the first column is the probability for 0.

Comparing the decision tree and the logistic regression model, it is apparent that the logistic regression model was more accurate where the AUC was 0.81.

Example 2: Predicting Restaurant Tip!

The owner of a restaurant was interested in studying the tipping patterns of his customers. He collected restaurant bills over a two week period that he believes provide a good sample of his customers. The data recorded include the amount of the bill, size of the tip, percentage tip, number of customers in the group, whether or not a credit card was used, day of the week, and a coded identity of the server.

Use the following line to read the data into a new dataframe called mydata.

mydata = read.csv("http://bit.ly/1StTazL",header=T)

There are seven variables being measured in the data set. These variables are

- bill amount,
- tip amount,
- method of payment (“credit”),
- number of guests,
- day of week,
- server, and
- percent tip.

The numerical variables are bill, tip, guest, and percent tip. The categorical data includes method of payment (whether or not a credit card was used), day of the week, and server.

```
> summary(mydata)
      Bill            Tip          Credit      Guests         Day     Server      PctTip     
 Min.   : 1.66   Min.   : 0.250   n:92   Min.   :1.000   F:25   A:55   Min.   : 6.70  
 1st Qu.:15.37   1st Qu.: 2.145   y:48   1st Qu.:2.000   M:18   B:55   1st Qu.:14.28  
 Median :19.95   Median : 3.340          Median :2.000   R:32   C:30   Median :16.35  
 Mean   :23.08   Mean   : 3.925          Mean   :2.129   T:13          Mean   :16.70  
 3rd Qu.:28.92   3rd Qu.: 5.000          3rd Qu.:2.000   W:52          3rd Qu.:18.20  
 Max.   :70.51   Max.   :15.000          Max.   :7.000                 Max.   :42.20  
```

Let’s consider linear regression model first and to see if we can predict the tip percentage from the available variables. We start by just considering a single variable, bill amount:

Model<-lm(PctTip~ Bill,data=mydata)

summary(Model)

looking at the output (next page), the tip percentage equation would be:

Tip_percentage=15.63+0.045*Bill

For example if the bill is $20 we should expect 16.53% tip that is $3.3.

```
> Model<-lm(PctTip~ Bill,data=mydata)
> summary(Model)

Call:
lm(formula = PctTip ~ Bill, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1099 -2.5014 -0.6406  1.5424 25.4749 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 15.63782    0.80641  19.392   <2e-16 ***
Bill         0.04588    0.03073   1.493    0.138    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.543 on 138 degrees of freedom
Multiple R-squared:  0.01589,	Adjusted R-squared:  0.00876 
F-statistic: 2.228 on 1 and 138 DF,  p-value: 0.1378
```

Is this model accurate?

No, not all! The R-squared (R^2) is 0.015 that suggest the model can only explain 1.5% of target variability.

Let’s add the payment type (credit card versus cash) and see if that improve the model.

```
> Model<-lm(PctTip~ Bill+Credit,data=mydata)
> summary(Model)

Call:
lm(formula = PctTip ~ Bill + Credit, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.0797 -2.4971 -0.6393  1.4481 25.6121 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 15.62773    0.80907  19.316   <2e-16 ***
Bill         0.04051    0.03306   1.226    0.222    
Credity      0.39029    0.87004   0.449    0.654    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.556 on 137 degrees of freedom
Multiple R-squared:  0.01733,	Adjusted R-squared:  0.002989 
F-statistic: 1.208 on 2 and 137 DF,  p-value: 0.3019
```

The accuracy is still very bad. Improvement is very small.

Let’s add the number of guests and see if that improve the model.

```
> Model<-lm(PctTip~ Bill+Credit+Guests,data=mydata)
> summary(Model)

Call:
lm(formula = PctTip ~ Bill + Credit + Guests, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.0996 -2.4733 -0.6583  1.4317 25.6184 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 15.59603    0.97758  15.954   <2e-16 ***
Bill         0.03919    0.04019   0.975    0.331    
Credity      0.39583    0.87839   0.451    0.653    
Guests       0.02832    0.48624   0.058    0.954    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.572 on 136 degrees of freedom
Multiple R-squared:  0.01736,	Adjusted R-squared:  -0.004317 
F-statistic: 0.8008 on 3 and 136 DF,  p-value: 0.4955
```

Almost no improvement at all!

Let's add the server.

```
> Model<-lm(PctTip~ Bill+Credit+Guests+Server,data=mydata)
> summary(Model)

Call:
lm(formula = PctTip ~ Bill + Credit + Guests + Server, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5309 -2.4302 -0.4073  1.7612 24.5375 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.35814    1.03189  15.853   <2e-16 ***
Bill         0.03096    0.04003   0.773   0.4407    
Credity      0.34493    0.88035   0.392   0.6958    
Guests       0.28537    0.49420   0.577   0.5646    
ServerB     -1.81404    0.88789  -2.043   0.0430 *  
ServerC     -1.81563    1.05482  -1.721   0.0875 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.522 on 134 degrees of freedom
Multiple R-squared:  0.0532,	Adjusted R-squared:  0.01787 
F-statistic: 1.506 on 5 and 134 DF,  p-value: 0.1921
```

Some minor improvement.R2 has jump to 0.053 which is still very small.

Let's add the day of the week.

```
> Model<-lm(PctTip~ Bill+Credit+Guests+Server+Day,data=mydata)
> summary(Model)

Call:
lm(formula = PctTip ~ Bill + Credit + Guests + Server + Day,
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1811 -2.1981 -0.4597  1.8566 24.8859

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.01493    1.31105  12.215   <2e-16 ***
Bill         0.03220    0.04098   0.786    0.434
Credity      0.26308    0.91795   0.287    0.775
Guests       0.26804    0.51625   0.519    0.604
ServerB     -1.52300    1.02334  -1.488    0.139
ServerC     -1.77563    1.07337  -1.654    0.100
DayM        -0.26769    1.56270  -0.171    0.864
DayR         0.34586    1.23785   0.279    0.780
DayT         0.97060    1.65175   0.588    0.558
DayW         0.32799    1.13633   0.289    0.773
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.582 on 130 degrees of freedom
Multiple R-squared:  0.05694, Adjusted R-squared:  -0.008354
F-statistic: 0.872 on 9 and 130 DF,  p-value: 0.552
```

Again, very small improvement.

Let's look at the variable importance once again using ANOVA:

```
> anova(Model)
Analysis of Variance Table

Response: PctTip
           Df  Sum Sq Mean Sq F value  Pr(>F)
Bill        1   45.98  45.983  2.1905 0.14128
Credit      1    4.18   4.177  0.1990 0.65630
Guests      1    0.07   0.071  0.0034 0.95374
Server      2  103.71  51.855  2.4703 0.08852 .
Day         4   10.81   2.703  0.1288 0.97173
Residuals 130 2728.92  20.992
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

We can see the Sum sq is highest for Server, followed by Bill. However, we can still see that the total variance captured by all variables is still too small compared to what is left out (i.e. Residuals) hence small R^2.

Conclusion: It is not possible to predict the tip percentage accurately with the given data.

Previously, we have been trying to predict the Tip Percentage. How about attempting to predict the Tip value itself? This should be easier since we all know the higher the bill the higher will be the tip:

```
> Model<-lm(Tip~ Bill,data=mydata)
> summary(Model)

Call:
lm(formula = Tip ~ Bill, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4037 -0.5167 -0.1043  0.2763  5.9616 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.283217   0.181713  -1.559    0.121    
Bill         0.182349   0.006925  26.332   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.024 on 138 degrees of freedom
Multiple R-squared:  0.834,	Adjusted R-squared:  0.8328 
F-statistic: 693.4 on 1 and 138 DF,  p-value: < 2.2e-16
```

Even with only a single variable, Bill, we can get R^2 of 0.834. This intuitively makes sense as well, higher the bill, the higher is the Tip. Predicting the percentage is more tricky.

Now if I give the model the Bill and the Tip Percentage, I should expect the model to have a very easy time telling me the Tip amount right? Let’s try:

```
> Model<-lm(Tip~ Bill+PctTip,data=mydata)
> summary(Model)

Call:
lm(formula = Tip ~ Bill + PctTip, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.27272 -0.17601 -0.04493  0.07691  2.92888 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.477561   0.148595  -23.40   <2e-16 ***
Bill         0.172978   0.002958   58.48   <2e-16 ***
PctTip       0.204270   0.008127   25.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4337 on 137 degrees of freedom
Multiple R-squared:  0.9704,	Adjusted R-squared:   0.97 
F-statistic:  2247 on 2 and 137 DF,  p-value: < 2.2e-16
```

Why R^2 is not 1 here?

Note: This was obviously a cheating and in modelling this is called leaking. The variable PctTip is leaking the information that we are trying to predict that is the Tip amount. You should always be careful of the leaks in your data.

Let’s now convert the problem to a classification question: Let’s see if we can predict if the tip will be above 15% or not.

First we create a new variable called Tip_above_15 as follows

mydata$Tip_above_15=as.factor(mydata$PctTip>15)

```
> mydata$Tip_above_15=as.factor(mydata$PctTip>15)
> summary(mydata)
      Bill             Tip          Credit      Guests        Day    Server      PctTip       Tip_above_15
 Min.   : 1.66   Min.   : 0.250   n:92   Min.   :1.000   F:25   A:55   Min.   : 6.70   FALSE:52
 1st Qu.:15.37   1st Qu.: 2.145   y:48   1st Qu.:2.000   M:18   B:55   1st Qu.:14.28   TRUE :88
 Median :19.95   Median : 3.340          Median :2.000   R:32   C:30   Median :16.35
 Mean   :23.08   Mean   : 3.925          Mean   :2.129   T:13          Mean   :16.70
 3rd Qu.:28.92   3rd Qu.: 5.000          3rd Qu.:2.000   W:52          3rd Qu.:18.20
 Max.   :70.51   Max.   :15.000          Max.   :7.000                 Max.   :42.20
```

Now this is a classification problem and we need to use logistic regression.

```
> Model<-glm(Tip_above_15~Bill+Credit+Guests+Server+Day,data=mydata,family =
> summary(Model)

Call:
glm(formula = Tip_above_15 ~ Bill + Credit + Guests + Server +
    Day, family = binomial, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0583  -1.0793   0.6749   0.9825   1.3952

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.927199   0.675954  -1.372   0.1702
Bill         0.005625   0.021165   0.266   0.7904
Credity      1.049190   0.470834   2.228   0.0259 *
Guests       0.399559   0.289757   1.379   0.1679
ServerB     -0.284177   0.482199  -0.589   0.5556
ServerC     -0.528130   0.525042  -1.006   0.3145
DayM         0.050723   0.714997   0.071   0.9434
DayR         0.796494   0.584012   1.364   0.1726
DayT         0.252470   0.800976   0.315   0.7526
DayW         0.564832   0.520988   1.084   0.2783
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 184.72  on 139  degrees of freedom
Residual deviance: 168.63  on 130  degrees of freedom
```

Use the model to predict the probability of having a tip above 15% for records in the mydata dataframe. Add this as a new variable to the dataframe and call it Tip_above_15_prob

mydata$Tip_above_15_prob<-predict(Model,newdata =mydata,type = 'response')

```
> mydata$Tip_above_15_prob<-predict(Model,newdata =mydata,type = 'response')
> View(mydata)
```

	Bill	Tip	Credit	Guests	Day	Server	PctTip	Tip_above_15	Tip_above_15_prob
1	10.17	1.83	n	1	W	A	18.0	TRUE	0.5235824
2	18.40	2.75	n	2	M	B	14.9	FALSE	0.4358475
3	11.72	2.28	y	1	W	A	19.5	TRUE	0.7599323
4	9.20	1.80	n	1	W	A	19.6	TRUE	0.5222211
5	18.14	4.00	n	3	W	C	22.1	TRUE	0.6011407
6	20.87	3.13	y	2	W	B	15.0	FALSE	0.7890421
7	25.09	5.00	y	2	R	C	19.9	TRUE	0.7909413
8	18.62	3.35	y	2	T	A	18.0	TRUE	0.7821636
9	39.75	7.25	y	2	W	A	18.2	TRUE	0.8467768
10	22.36	3.00	n	2	F	C	13.4	FALSE	0.3704162
11	32.31	4.69	n	2	W	A	14.5	FALSE	0.6498774

Is this model accurate? Use AUC as a metric.

```
> library(pROC)
> mydata$Tip_above_15_prob<-predict(Model,newdata =mydata,type = 'response')
> roc(mydata$Tip_above_15,mydata$Tip_above_15_prob)

Call:
roc.default(response = mydata$Tip_above_15, predictor = mydata$Tip_above_15_prob)

Data: mydata$Tip_above_15_prob in 52 controls (mydata$Tip_above_15 FALSE) < 88 cases (mydata$Tip_above_15 TRUE).
Area under the curve: 0.6983
> |
```

The ROC is nearly 0.7 which is not too bad! In other words, while the actual prediction of Tip Percentage using linear regression analysis was difficult (the R2 was around 0.05), just predicting weather the tip would be above or below 15% can be done with a much better accuracy.

Now let us build a decision tree to do the same prediction.

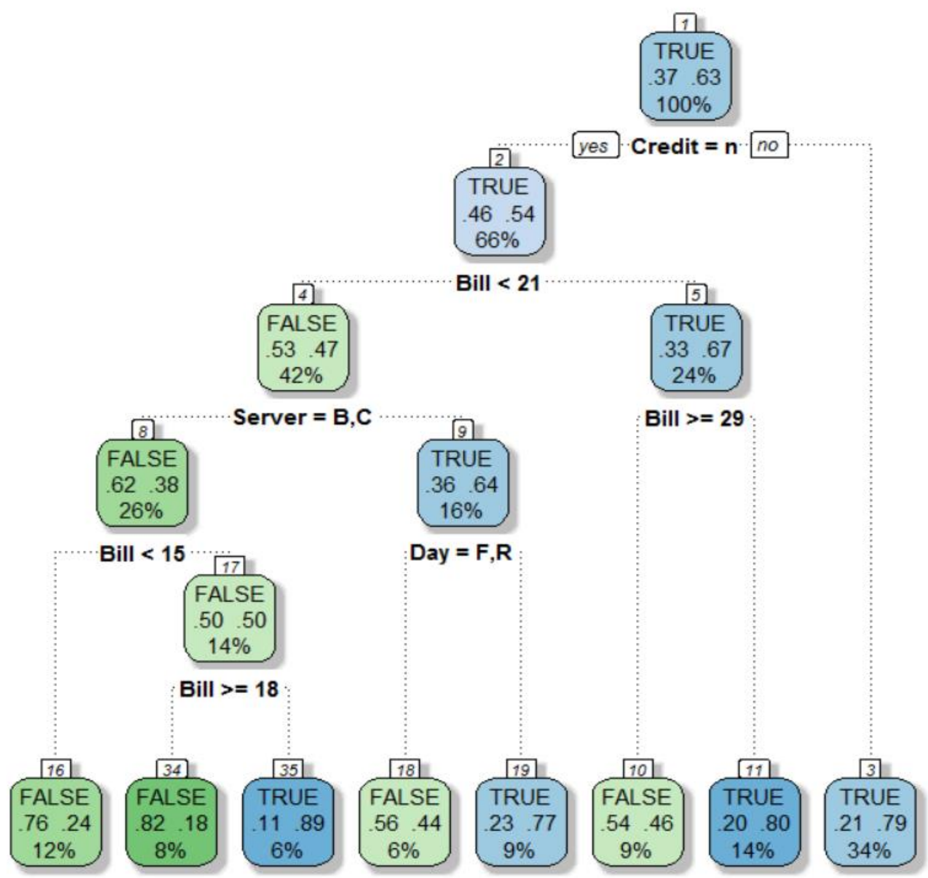
```
> library(rpart)
Warning message:
package ‘rpart’ was built under R version 3.4.2
> Model_3=rpart(Tip_above_15~Bill+Credit+Server+Day,data=mydata,method='class')
> summary(Model_3)
Call:
rpart(formula = Tip_above_15 ~ Bill + Credit + Server + Day,
    data = mydata, method = "class")
  n= 140

          CP nsplit rel error   xerror      xstd
1 0.05769231      0 1.0000000 1.000000 0.1099450
2 0.01923077      5 0.6923077 1.134615 0.1123573
3 0.01000000      7 0.6538462 1.019231 0.1103647

Variable importance
  Bill Credit    Day Server
    55     20     15     10
```

Unlike the logistic regression model above, the decision tree model considers the Bill amount as the most important variable as oppose to Credit variable which was the most important variable in the logistic regression.

Plot the decision tree.



Just look at the tree above and predict the probability of receiving a tip higher than 15% if the bill was $22, the server was A, the payment was made using a credit card on a Friday. Use the predict function to additionally verify that.

Just by looking at the tree, the payment was made using credit card so we move to node 2 (a yes at node one will move us to node 2). At node 2, the bill is above $21 so we have a No and therefore we take the right branch and move to node 5. At node 5, we have the bill less than $29, so we have a 'No' and therefore will take the right branch and end up in the terminal node 11. In terminal node 11, the probability of TRUE (i.e. tip percentage above 15%) is 80%. S final answer 80%

```
> predict(Model_3, newdata=data.frame(Bill=22,Credit='y',Server='A',Day='F'),type='prob')
      FALSE      TRUE
1 0.2083333 0.7916667
```

Same number (sometimes values are rounded on the plot by 1% to make it easier to read).

What is the AUC of the model?

```
> library(pROC)
> Predicted_Values<-predict(Model_3, newdata= mydata,type='prob')
> roc(mydata$Tip_above_15, Predicted_Values[,2])

Call:
roc.default(response = mydata$Tip_above_15, predictor = Predicted_Values[,     2])

Data: Predicted_Values[, 2] in 52 controls (mydata$Tip_above_15 FALSE) < 88 cases (mydata$Tip_above_15 TRUE).
Area under the curve: 0.7646
```