**Chi-Square test to identify significant variables which have impact on employee attrition:-**

**Department is significant:**

> tbl <- table(CTDF$Department, CTDF$Attrition)

> tbl

No Yes

Human Resources 102 24

Research & Development 1656 266

Sales 708 184

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 21.592, df = 2, p-value = 2.048e-05

**Gender not significant**:

> tbl <- table(CTDF$Gender, CTDF$Attrition)

> tbl

No Yes

Female 1002 174

Male 1464 300

> chisq.test(tbl)

Pearson's Chi-squared test with Yates' continuity correction

data: tbl

X-squared = 2.3896, df = 1, p-value = 0.1221

**Age significant:**

> tbl <- table(CTDF$Age, CTDF$Attrition)

> tbl

No Yes

18 8 8

19 6 12

20 10 12

21 14 12

22 22 10

23 20 8

24 38 14

25 40 12

26 54 24

27 90 6

28 68 28

29 100 36

30 102 18

31 102 36

32 100 22

33 92 24

34 136 18

35 136 20

36 126 12

37 88 12

38 112 4

39 72 12

40 104 10

41 68 12

42 88 4

43 60 4

44 54 12

45 78 4

46 58 8

47 42 6

48 34 4

49 44 4

50 50 10

51 34 4

52 30 6

53 34 4

54 36 0

55 38 6

56 22 6

57 8 0

58 18 10

59 20 0

60 10 0

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 238.35, df = 42, p-value < 2.2e-16

Warning message:

In chisq.test(tbl) : Chi-squared approximation may be incorrect

**Age-Bracket significant:**

CTDF$Age\_Brk = cut(CTDF$Age, breaks=c(18, 25, 35, 45, 61), labels=c("18-24","25-34","35-44","45-60"),right=F)

data.frame(CTDF,CTDF$Age\_Brk)

> tbl <- table(CTDF$Age\_Brk, CTDF$Attrition)

> tbl

No Yes

18-24 118 76

25-34 884 224

35-44 908 102

45-60 556 72

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 127.14, df = 3, p-value < 2.2e-16

**Business Travel Significant:**

> tbl <- table(CTDF$BusinessTravel, CTDF$Attrition)

> tbl

No Yes

Non-Travel 276 24

Travel\_Frequently 416 138

Travel\_Rarely 1774 312

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 48.365, df = 2, p-value = 3.146e-11

**Daily Rate-Bracket significant:**

CTDF$DailyRate\_Brk = cut(CTDF$DailyRate, breaks=c(102, 500, 1000, 1499), labels=c("102-499","500-999","1000-1499"))

data.frame(CTDF,CTDF$DailyRate\_Brk)

> tbl <- table(CTDF$DailyRate\_Brk, CTDF$Attrition)

> tbl

No Yes

102-499 652 156

500-999 908 176

1000-1499 904 142

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 11.081, df = 2, p-value = 0.003925

ctrl <- chaid\_control(minbucket = 10, minsplit = 100, alpha2=.05, alpha4 = .05)

chaid.tree <-chaid(Attrition~Department+Age\_Brk+BusinessTravel+DailyRate\_Brk,data=CTDF, control = ctrl)

print(chaid.tree)

**Model formula:**

**Attrition ~ Department + Age\_Brk + BusinessTravel + DailyRate\_Brk**

Fitted party:

[1] root

| [2] Age\_Brk in 18-24

| | [3] BusinessTravel in Non-Travel, Travel\_Rarely

| | | [4] DailyRate\_Brk in 102-499, 1000-1499

| | | | [5] BusinessTravel in Non-Travel, Travel\_Frequently: No (n = 8, err = 0.0%)

| | | | [6] BusinessTravel in Travel\_Rarely: No (n = 100, err = 42.0%)

| | | [7] DailyRate\_Brk in 500-999: No (n = 92, err = 17.4%)

| | [8] BusinessTravel in Travel\_Frequently: Yes (n = 30, err = 26.7%)

| [9] Age\_Brk in 25-34

| | [10] BusinessTravel in Non-Travel, Travel\_Rarely

| | | [11] DailyRate\_Brk in 102-499: No (n = 248, err = 21.8%)

| | | [12] DailyRate\_Brk in 500-999, 1000-1499: No (n = 708, err = 14.4%)

| | [13] BusinessTravel in Travel\_Frequently

| | | [14] Department in Human Resources: Yes (n = 10, err = 20.0%)

| | | [15] Department in Research & Development, Sales: No (n = 246, err = 27.6%)

| [16] Age\_Brk in 35-44: No (n = 936, err = 9.2%)

| [17] Age\_Brk in 45-60

| | [18] Department in Human Resources, Research & Development: No (n = 374, err = 9.1%)

| | [19] Department in Sales

| | | [20] BusinessTravel in Non-Travel, Travel\_Rarely: No (n = 136, err = 16.2%)

| | | [21] BusinessTravel in Travel\_Frequently: No (n = 34, err = 35.3%)

Number of inner nodes: 9

Number of terminal nodes: 12

**Model formula:**

**Attrition ~ Department + Age\_Brk + BusinessTravel**

Fitted party:

[1] root

| [2] Age\_Brk in 18-24

| | [3] BusinessTravel in Non-Travel, Travel\_Rarely: No (n = 200, err = 29.0%)

| | [4] BusinessTravel in Travel\_Frequently: Yes (n = 30, err = 26.7%)

| [5] Age\_Brk in 25-34

| | [6] BusinessTravel in Non-Travel, Travel\_Rarely: No (n = 956, err = 16.3%)

| | [7] BusinessTravel in Travel\_Frequently

| | | [8] Department in Human Resources: Yes (n = 10, err = 20.0%)

| | | [9] Department in Research & Development, Sales: No (n = 246, err = 27.6%)

| [10] Age\_Brk in 35-44: No (n = 936, err = 9.2%)

| [11] Age\_Brk in 45-60

| | [12] Department in Human Resources, Research & Development: No (n = 376, err = 9.0%)

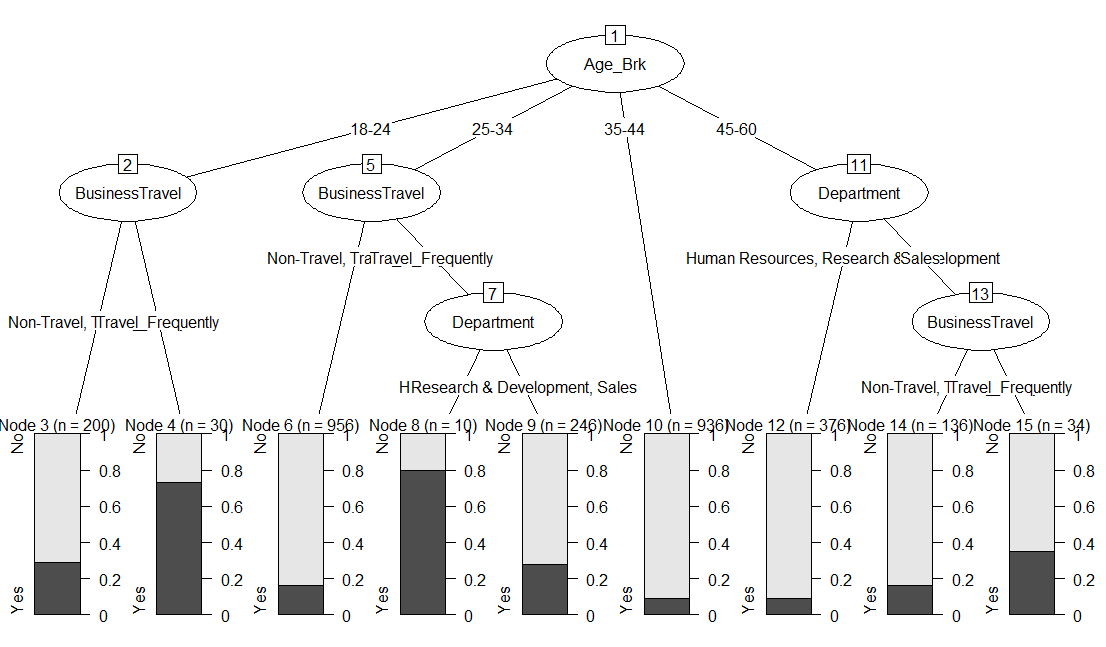
| | [13] Department in Sales

| | | [14] BusinessTravel in Non-Travel, Travel\_Rarely: No (n = 136, err = 16.2%)

| | | [15] BusinessTravel in Travel\_Frequently: No (n = 34, err = 35.3%)

Number of inner nodes: 6

Number of terminal nodes: 9



**Daily Distance from Home-Bracket significant:**

CTDF$DistHome\_Brk = cut(CTDF$DistanceFromHome, breaks=c(1, 10, 20, 29), labels=c("1-9","10-19","20-29"))

data.frame(CTDF,CTDF$DistHome\_Brk)

> tbl <- table(CTDF$DistHome\_Brk, CTDF$Attrition)

> tbl

No Yes

1-9 1400 236

10-19 384 96

20-29 318 90

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 18.247, df = 2, p-value = 0.0001091

**Education field significant:**

> tbl <- table(CTDF$EducationField, CTDF$Attrition)

> tbl

No Yes

Human Resources 40 14

Life Sciences 1034 178

Marketing 248 70

Medical 802 126

Other 142 22

Technical Degree 200 64

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 32.049, df = 5, p-value = 5.809e-06

**Environment Satisfaction field significant:**

> tbl <- table(CTDF$EnvironmentSatisfaction, CTDF$Attrition)

> tbl

No Yes

1 424 144

2 488 86

3 782 124

4 772 120

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 45.008, df = 3, p-value = 9.218e-10

**Job Involvement field significant:**

> tbl <- table(CTDF$JobInvolvement, CTDF$Attrition)

> tbl

No Yes

1 110 56

2 608 142

3 1486 250

4 262 26

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 56.984, df = 3, p-value = 2.59e-12

**Job Level field significant:**

> tbl <- table(CTDF$JobLevel, CTDF$Attrition)

> tbl

No Yes

1 800 286

2 964 104

3 372 64

4 202 10

5 128 10

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 145.06, df = 4, p-value < 2.2e-16

**Job Role field significant:**

> tbl <- table(CTDF$JobRole, CTDF$Attrition)

> tbl

No Yes

Healthcare Representative 244 18

Human Resources 80 24

Laboratory Technician 394 124

Manager 194 10

Manufacturing Director 270 20

Research Director 156 4

Research Scientist 490 94

Sales Executive 538 114

Sales Representative 100 66

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 172.38, df = 8, p-value < 2.2e-16

**Job Satisfaction field significant:**

> tbl <- table(CTDF$JobSatisfaction, CTDF$Attrition)

> tbl

No Yes

1 446 132

2 468 92

3 738 146

4 814 104

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 35.01, df = 3, p-value = 1.212e-07

**Marital Status field significant:**

> tbl <- table(CTDF$MaritalStatus, CTDF$Attrition)

> tbl

No Yes

Divorced 588 66

Married 1178 168

Single 700 240

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 92.327, df = 2, p-value < 2.2e-16

**No. of companies worked field significant:**

> tbl <- table(CTDF$NumCompaniesWorked, CTDF$Attrition)

> tbl

No Yes

0 348 46

1 846 196

2 260 32

3 286 32

4 244 34

5 94 32

6 108 32

7 114 34

8 86 12

9 80 24

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 51.489, df = 9, p-value = 5.646e-08

**Overtime field significant:**

> tbl <- table(CTDF$OverTime, CTDF$Attrition)

> tbl

No Yes

No 1888 220

Yes 578 254

> chisq.test(tbl)

Pearson's Chi-squared test with Yates' continuity correction

data: tbl

X-squared = 176.61, df = 1, p-value < 2.2e-16

**Stock option level field significant:**

> tbl <- table(CTDF$StockOptionLevel, CTDF$Attrition)

> tbl

No Yes

0 954 308

1 1080 112

2 292 24

3 140 30

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 121.2, df = 3, p-value < 2.2e-16

**Trainings last year level field significant:**

> tbl <- table(CTDF$TrainingTimesLastYear, CTDF$Attrition)

> tbl

No Yes

0 78 30

1 124 18

2 898 196

3 844 138

4 194 52

5 210 28

6 118 12

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 30.293, df = 6, p-value = 3.458e-05

**Work life balance field significant:**

> tbl <- table(CTDF$WorkLifeBalance, CTDF$Attrition)

> tbl

No Yes

1 110 50

2 572 116

3 1532 254

4 252 54

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 32.65, df = 3, p-value = 3.817e-07

**Monthly Income Break field significant:**

CTDF$MonthlyIncome\_Brk = cut(CTDF$MonthlyIncome, breaks=c(1000, 5000, 10000, 15000, 20000), labels=c("1000-4999","5000-9999","10000-14999","15000-19999"))

data.frame(CTDF,CTDF$MonthlyIncome\_Brk)

> tbl <- table(CTDF$MonthlyIncome\_Brk, CTDF$Attrition)

> tbl

No Yes

1000-4999 1172 326

5000-9999 782 98

10000-14999 256 40

15000-19999 256 10

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 82.968, df = 3, p-value < 2.2e-16

**Total Working Years field significant:**

CTDF$TotalWorkingYears\_Brk = cut(CTDF$TotalWorkingYears, breaks=c(-1, 11, 21, 31, 41), labels=c("0-10","11-20","21-30","31-40"))

data.frame(CTDF,CTDF$TotalWorkingYears\_Brk)

> tbl <- table(CTDF$TotalWorkingYears\_Brk, CTDF$Attrition)

> tbl

No Yes

0-10 1540 378

11-20 610 66

21-30 250 22

31-40 66 8

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 52.976, df = 3, p-value = 1.855e-11

**Years at company field significant:**

CTDF$YearsAtCompany\_Brk = cut(CTDF$YearsAtCompany, breaks=c(-1, 11, 21, 31, 41), labels=c("0-10","11-20","21-30","31-40"))

data.frame(CTDF,CTDF$YearsAtCompany\_Brk)

> tbl <- table(CTDF$YearsAtCompany\_Brk, CTDF$Attrition)

> tbl

No Yes

0-10 2074 438

11-20 302 22

21-30 70 8

31-40 20 6

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 26.988, df = 3, p-value = 5.923e-06

Warning message:

In chisq.test(tbl) : Chi-squared approximation may be incorrect

**Years in current role field significant:**

CTDF$YearsInCurrentRole\_Brk = cut(CTDF$YearsInCurrentRole, breaks=c(0, 7, 13, 19), labels=c("0-6","7-12","13-18"),right=F)

data.frame(CTDF,CTDF$YearsInCurrentRole\_Brk)

> tbl <- table(CTDF$YearsInCurrentRole\_Brk, CTDF$Attrition)

> tbl

No Yes

0-6 1598 372

7-12 784 94

13-18 84 8

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 33.903, df = 2, p-value = 4.346e-08

**Years with current manager field significant:**

> tbl <- table(CTDF$YearsWithCurrManager\_Brk, CTDF$Attrition)

> tbl

No Yes

0-5 1548 360

6-11 820 110

12-17 98 4

> chisq.test(tbl)

Pearson's Chi-squared test

data: tbl

X-squared = 34.546, df = 2, p-value = 3.15e-08

> chaid.tree <-chaid(Attrition~Age\_Brk+OverTime+Department+MaritalStatus+JobLevel,data=CTDF, control = ctrl)

> print(chaid.tree)

**Model formula:**

Attrition ~ Age\_Brk + OverTime + Department + MaritalStatus +

JobLevel

Fitted party:

[1] root

| [2] OverTime in No

| | [3] JobLevel in 1, 3

| | | [4] Department in Human Resources, Sales: No (n = 282, err = 23.4%)

| | | [5] Department in Research & Development

| | | | [6] MaritalStatus in Divorced, Married: No (n = 506, err = 6.7%)

| | | | [7] MaritalStatus in Single: No (n = 284, err = 19.7%)

| | [8] JobLevel in 2

| | | [9] MaritalStatus in Divorced, Married: No (n = 536, err = 4.9%)

| | | [10] MaritalStatus in Single: No (n = 240, err = 10.8%)

| | [11] JobLevel in 4, 5: No (n = 248, err = 3.2%)

| [12] OverTime in Yes

| | [13] JobLevel in 1

| | | [14] MaritalStatus in Divorced, Married: No (n = 194, err = 43.3%)

| | | [15] MaritalStatus in Single: Yes (n = 114, err = 33.3%)

| | [16] JobLevel in 2, 3, 4, 5

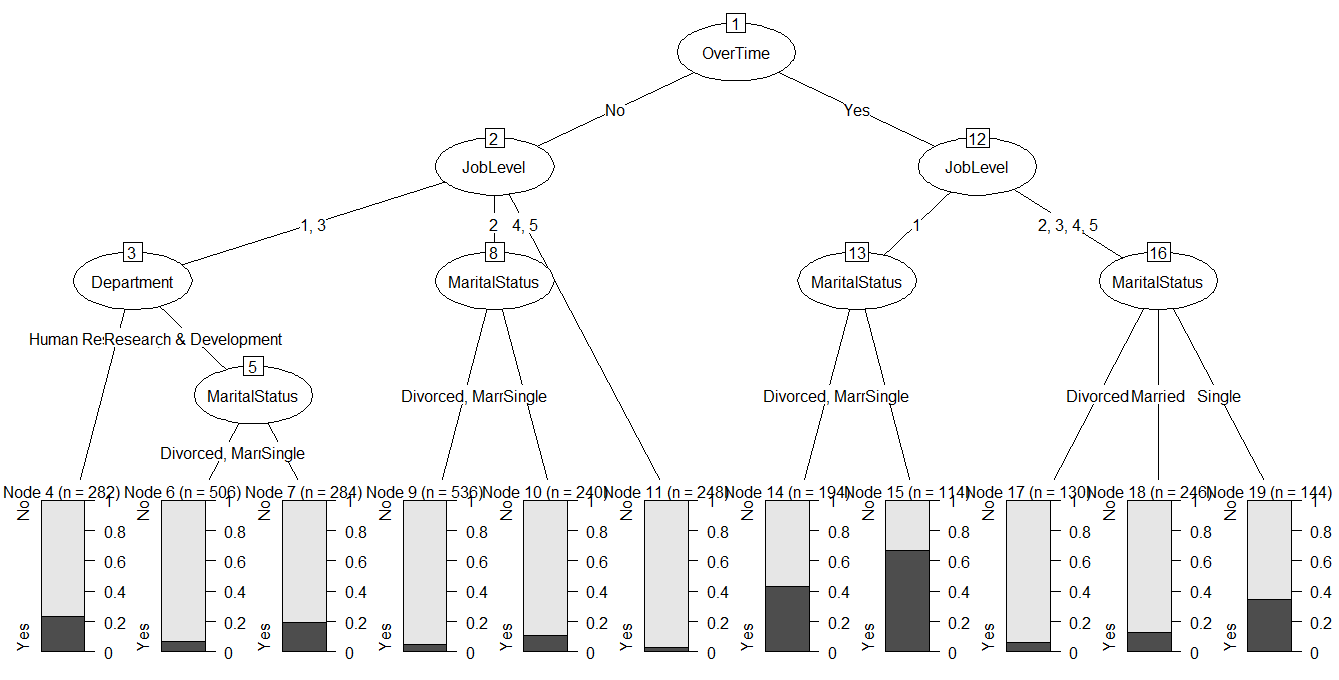
| | | [17] MaritalStatus in Divorced: No (n = 130, err = 6.2%)

| | | [18] MaritalStatus in Married: No (n = 246, err = 13.0%)

| | | [19] MaritalStatus in Single: No (n = 144, err = 34.7%)

Number of inner nodes: 8

Number of terminal nodes: 11



**Model formula:**

**Attrition ~ Age\_Brk + OverTime + Department + JobLevel**

ctrl <- chaid\_control(minbucket = 100, minsplit = 300, alpha2=.05, alpha4 = .05)

chaid.tree <-chaid(Attrition~Age\_Brk+OverTime+Department+JobLevel,data=CTDF, control = ctrl)

Fitted party:

[1] root

| [2] OverTime in No

| | [3] Age\_Brk in 18-24: No (n = 132, err = 25.8%)

| | [4] Age\_Brk in 25-34

| | | [5] JobLevel in 1, 4, 5

| | | | [6] Department in Human Resources, Sales: No (n = 86, err = 32.6%)

| | | | [7] Department in Research & Development: No (n = 310, err = 14.2%)

| | | [8] JobLevel in 2: No (n = 344, err = 7.0%)

| | | [9] JobLevel in 3: No (n = 86, err = 16.3%)

| | [10] Age\_Brk in 35-44, 45-60: No (n = 1150, err = 6.6%)

| [11] OverTime in Yes

| | [12] JobLevel in 1

| | | [13] Age\_Brk in 18-24: Yes (n = 56, err = 28.6%)

| | | [14] Age\_Brk in 25-34: Yes (n = 146, err = 43.8%)

| | | [15] Age\_Brk in 35-44, 45-60: No (n = 110, err = 38.2%)

| | [16] JobLevel in 2, 3, 4, 5

| | | [17] Department in Human Resources, Research & Development

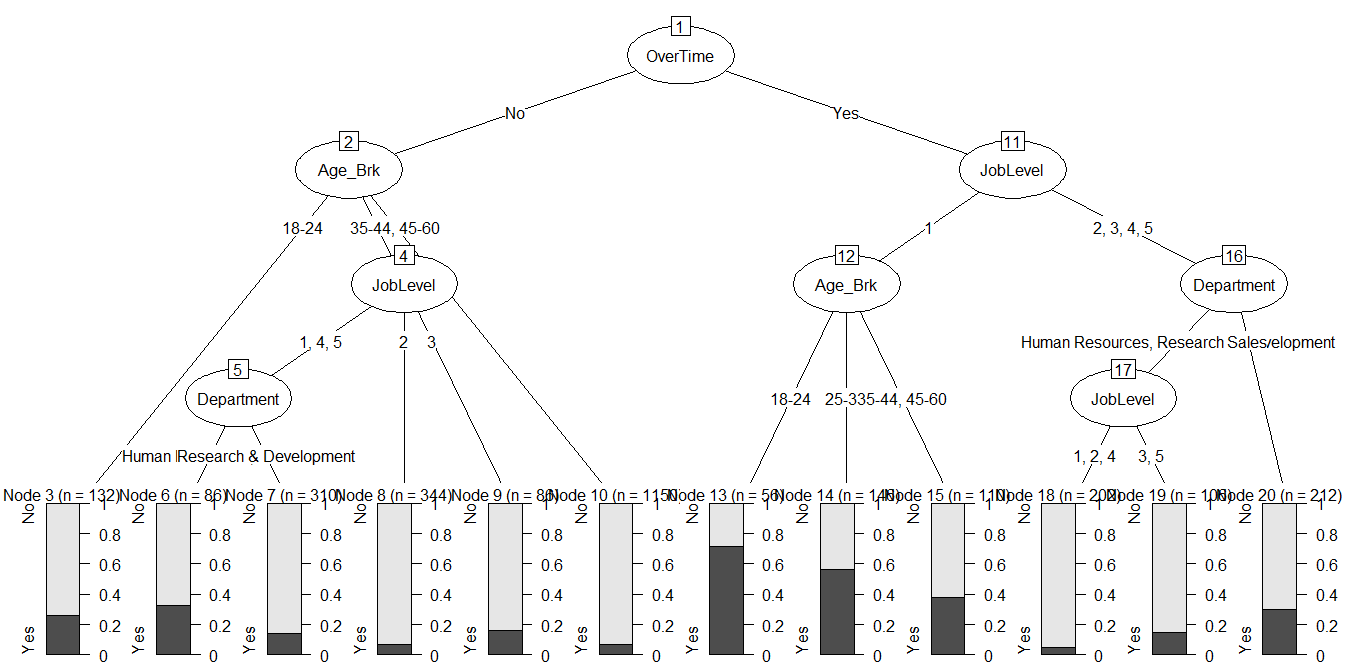
| | | | [18] JobLevel in 1, 2, 4: No (n = 202, err = 5.0%)

| | | | [19] JobLevel in 3, 5: No (n = 106, err = 15.1%)

| | | [20] Department in Sales: No (n = 212, err = 30.2%)

Number of inner nodes: 8

Number of terminal nodes: 12



**CART Model:**

**Attrition~All Columns (Dev)**

> CTDF.dev1 <- read.table("C:/Users/Owner/Documents/Dev\_HR\_Employee\_Attrition\_Data.csv", sep = ",", header = T)

> m2 <- rpart(formula = Attrition ~ ., data = CTDF.dev1, method = "class", control = r.ctrl)

> m2

n= 2040

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 2040 325 No (0.8406863 0.1593137)

2) OverTime=No 1462 147 No (0.8994528 0.1005472) \*

3) OverTime=Yes 578 178 No (0.6920415 0.3079585)

6) MonthlyIncome>=3751.5 377 69 No (0.8169761 0.1830239) \*

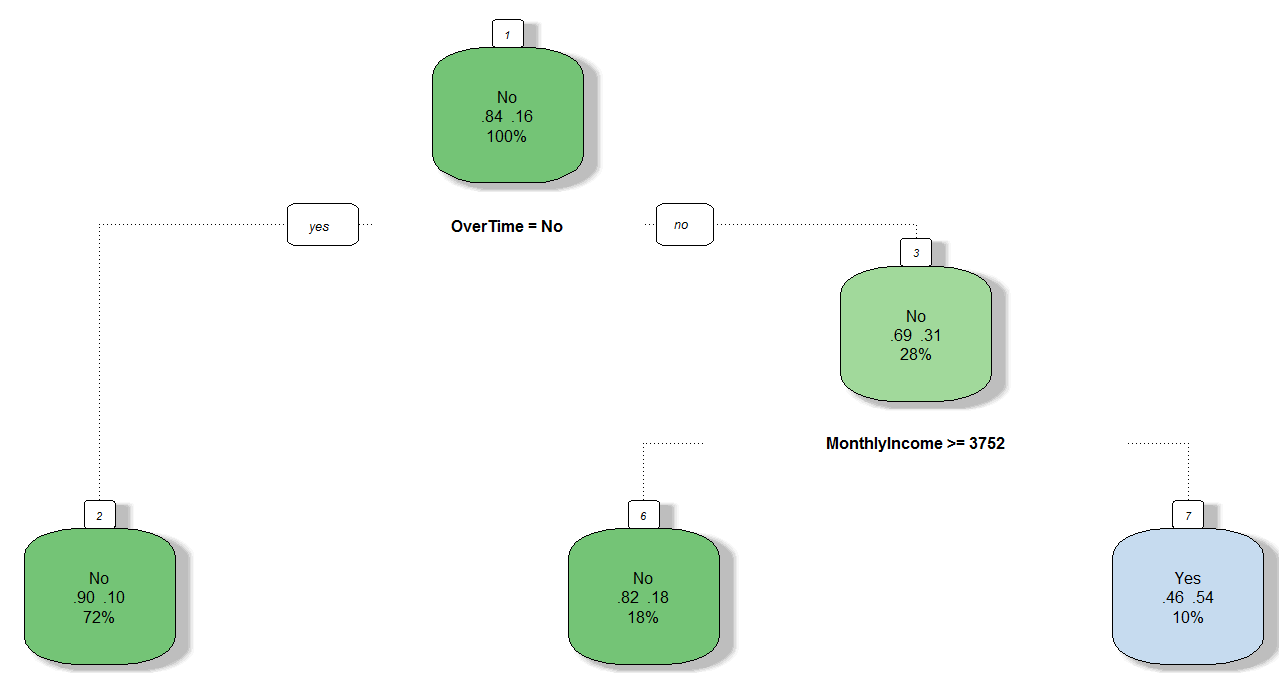
7) MonthlyIncome< 3751.5 201 92 Yes (0.4577114 0.5422886) \*

Cost Component Pruning:

325/2040 + 1Alpha = 308/2040 + 3Alpha

1 Alpha = 0.0041

>fancyRpartPlot(m2)



> printcp(m2)

Classification tree:

rpart(formula = Attrition ~ ., data = CTDF.dev1[, -1], method = "class",

control = r.ctrl)

Variables actually used in tree construction:

[1] MonthlyIncome OverTime

Root node error: 325/2040 = 0.15931

n= 2040

CP nsplit rel error xerror xstd

1 0.026154 0 1.00000 1.00000 0.050860

2 0.000000 2 0.94769 0.96615 0.050152

**Attrition~ Age\_Brk+OverTime+Department+JobLevel**

> m2 <- rpart(formula = Attrition ~ **Age\_Brk+OverTime+Department+JobLevel**, data = CTDF.dev1, method = "class", control = r.ctrl)

> m2

n= 2040

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 2040 325 No (0.8406863 0.1593137)

2) OverTime=No 1462 147 No (0.8994528 0.1005472) \*

3) OverTime=Yes 578 178 No (0.6920415 0.3079585)

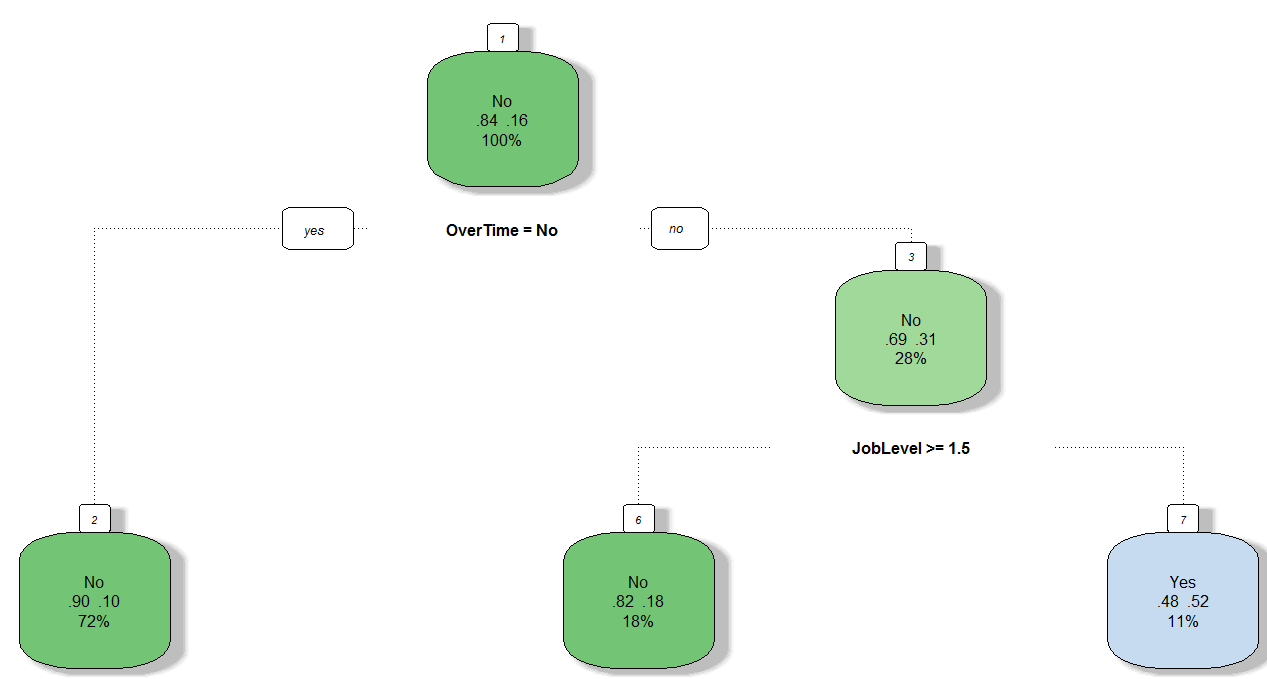
6) JobLevel>=1.5 358 63 No (0.8240223 0.1759777) \*

7) JobLevel< 1.5 220 105 Yes (0.4772727 0.5227273) \*

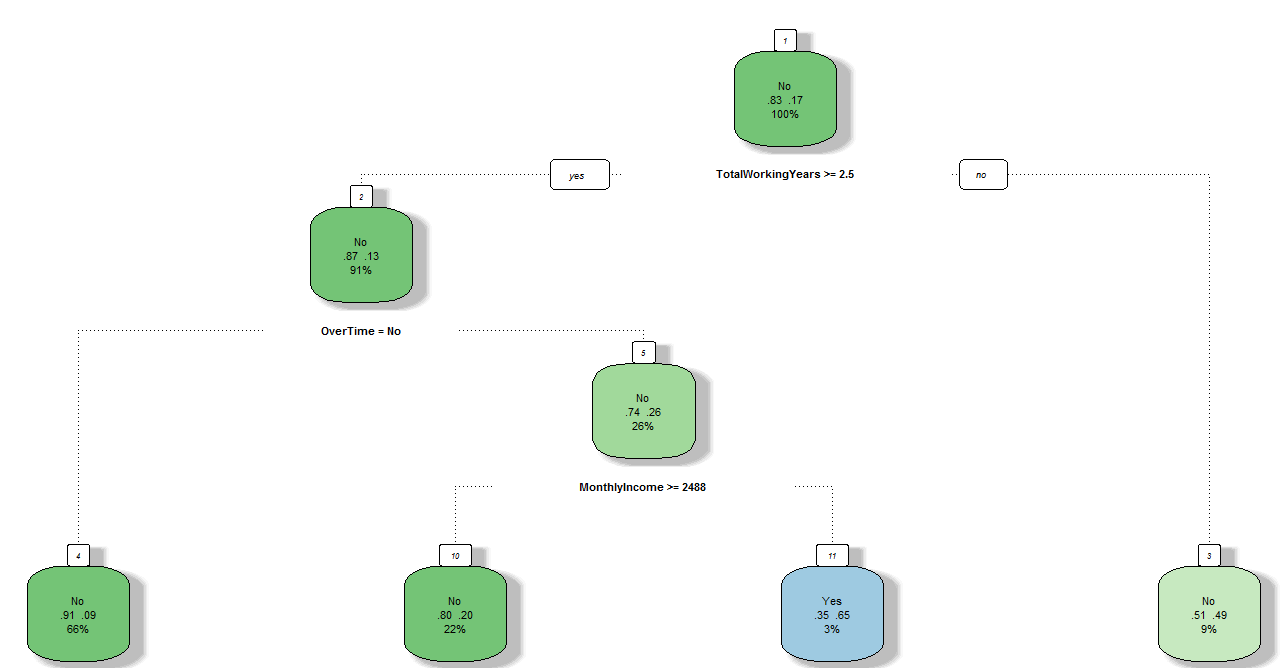
Cost Component Pruning:

325/2040 + 1Alpha = 315/2040 + 3Alpha

1 Alpha = 0.0024



* **Attrition~All Columns (Holdout Validation)**



> m3

n= 900

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 900 149 No (0.83444444 0.16555556)

2) TotalWorkingYears>=2.5 821 110 No (0.86601705 0.13398295)

4) OverTime=No 591 51 No (0.91370558 0.08629442) \*

5) OverTime=Yes 230 59 No (0.74347826 0.25652174)

10) MonthlyIncome>=2488 199 39 No (0.80402010 0.19597990) \*

11) MonthlyIncome< 2488 31 11 Yes (0.35483871 0.64516129) \*

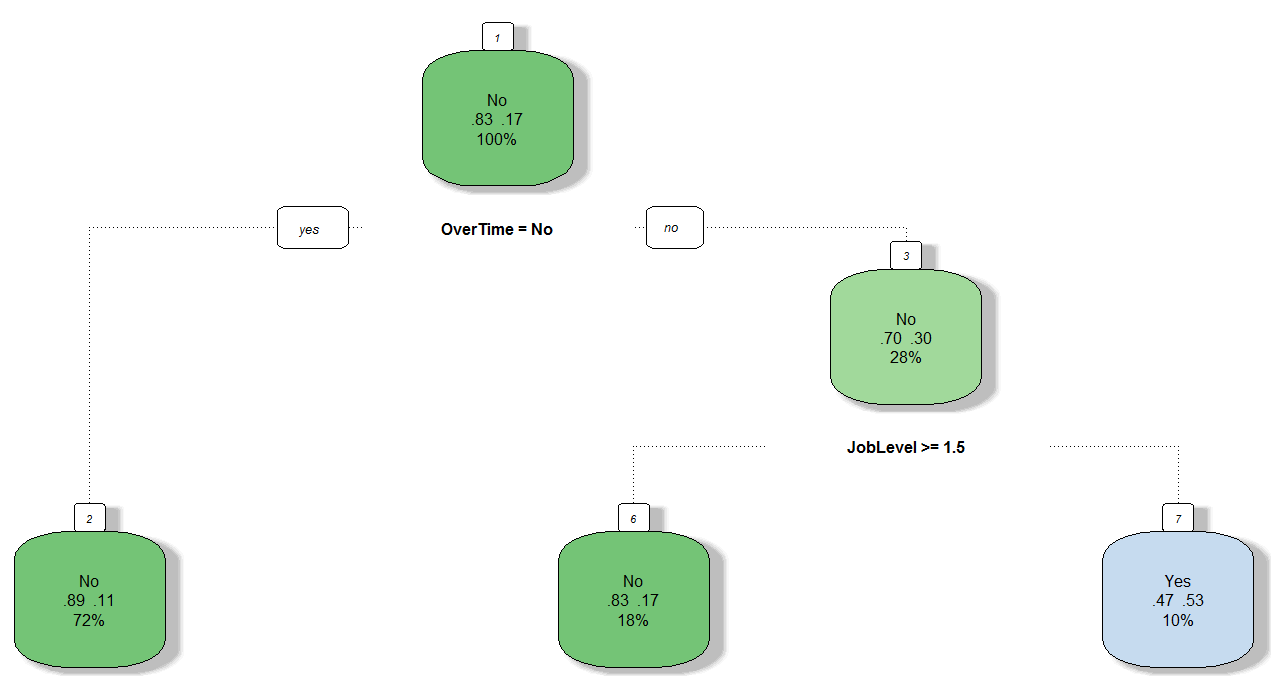
3) TotalWorkingYears< 2.5 79 39 No (0.50632911 0.49367089) \*

Cost Component Pruning:

149/900 + 1Alpha = 140/900 + 4Alpha

1 Alpha = .01/3 = 0.0033

* **Attrition~Age\_Brk+OverTime+Department+JobLevel(Holdout Validation)**



> r.ctrl = rpart.control(minsplit=100, minbucket = 30, cp = 0, xval = 10)

> m3 <- rpart(formula = Attrition ~ Age\_Brk+OverTime+Department+JobLevel, data = CTDF.holdout, method = "class", control = r.ctrl)

> m3

n= 900

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 900 149 No (0.8344444 0.1655556)

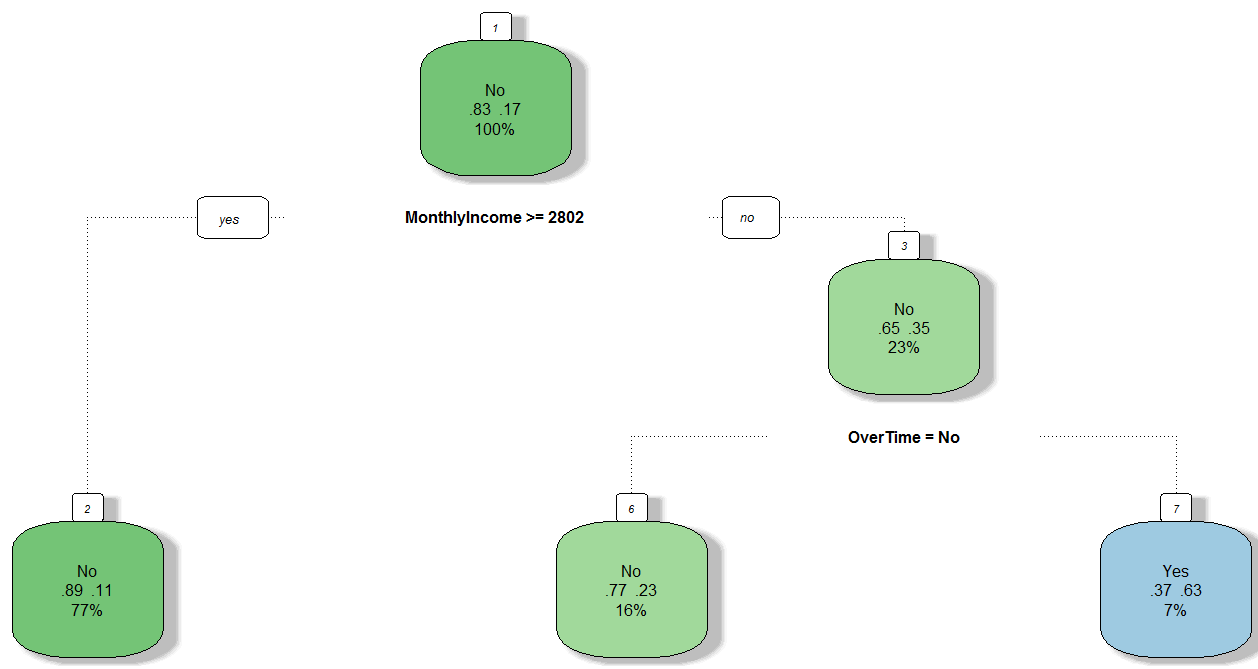
2) OverTime=No 646 73 No (0.8869969 0.1130031) \*

3) OverTime=Yes 254 76 No (0.7007874 0.2992126)

6) JobLevel>=1.5 162 27 No (0.8333333 0.1666667) \*

7) JobLevel< 1.5 92 43 Yes (0.4673913 0.5326087) \*

* **Attrition~Age\_Brk+OverTime+MonthlyIncome+JobLevel+JobRole+MaritalStatus+StockOptionLevel (Holdout Validation)**



> CTDF.holdout <- read.table("C:/Users/Owner/Documents/Holdout\_HR\_Employee\_Attrition\_Data.csv", sep = ",", header = T)

> r.ctrl = rpart.control(minsplit=100, minbucket = 30, cp = 0, xval = 10)

> m3 <- rpart(formula = Attrition ~ Age\_Brk+OverTime+MonthlyIncome+JobLevel+JobRole+MaritalStatus+StockOptionLevel, data = CTDF.holdout, method = "class", control = r.ctrl)

> m3

n= 900

node), split, n, loss, yval, (yprob)

1) root 900 149 No (0.8344444 0.1655556)

2) MonthlyIncome>=2802 697 78 No (0.8880918 0.1119082) \*

3) MonthlyIncome< 2802 203 71 No (0.6502463 0.3497537)

6) OverTime=No 141 32 No (0.7730496 0.2269504) \*

7) OverTime=Yes 62 23 Yes (0.3709677 0.6290323) \*

>fancyRpartPlot(m3)

> printcp(m3)

Classification tree:

rpart(formula = Attrition ~ Age\_Brk + OverTime + MonthlyIncome +

JobLevel + JobRole + MaritalStatus + StockOptionLevel, data = CTDF.holdout,

method = "class", control = r.ctrl)

Variables actually used in tree construction:

[1] MonthlyIncome OverTime

Root node error: 149/900 = 0.16556

n= 900

CP nsplit rel error xerror xstd

1 0.053691 0 1.00000 1.00000 0.074835

2 0.000000 2 0.89262 0.88591 0.071230

**Cost Component Pruning:**

149/900 + 1Alpha = 133/900 + 3Alpha; 1 Alpha =0.0088