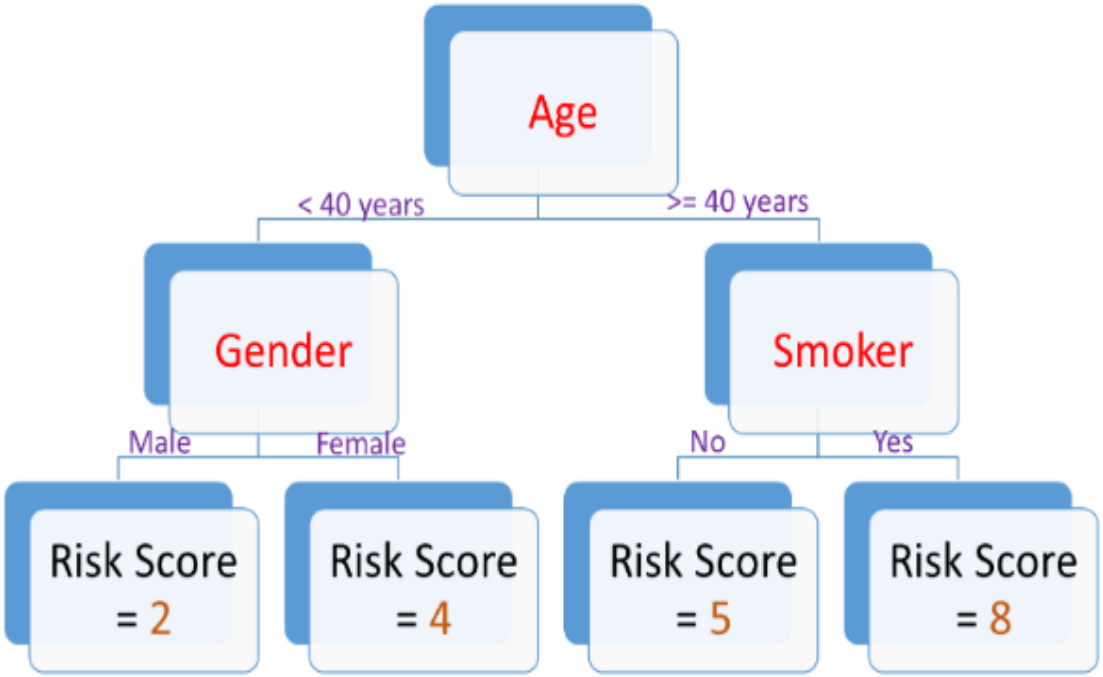
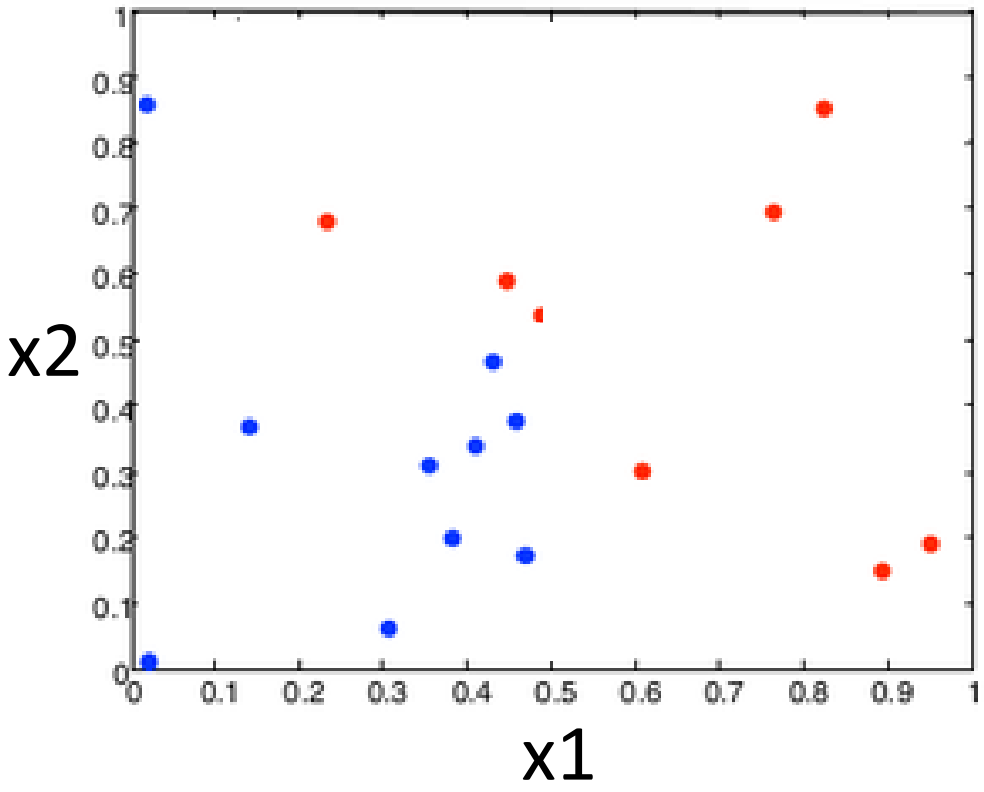
Short Course
Modern Methods in Sport Statistics

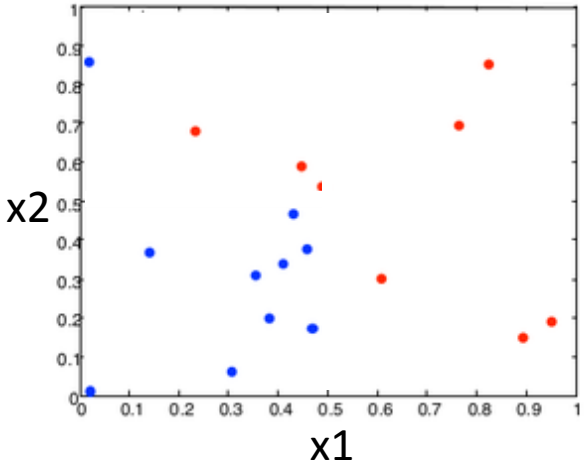
Session 2

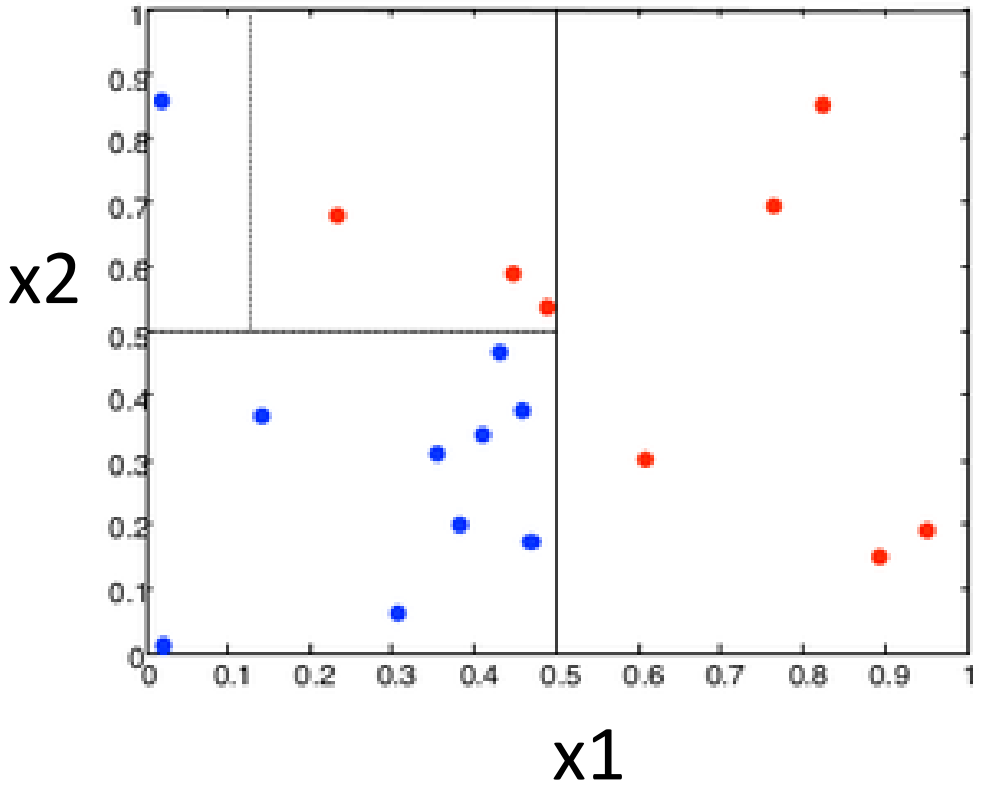
Trees in Sport

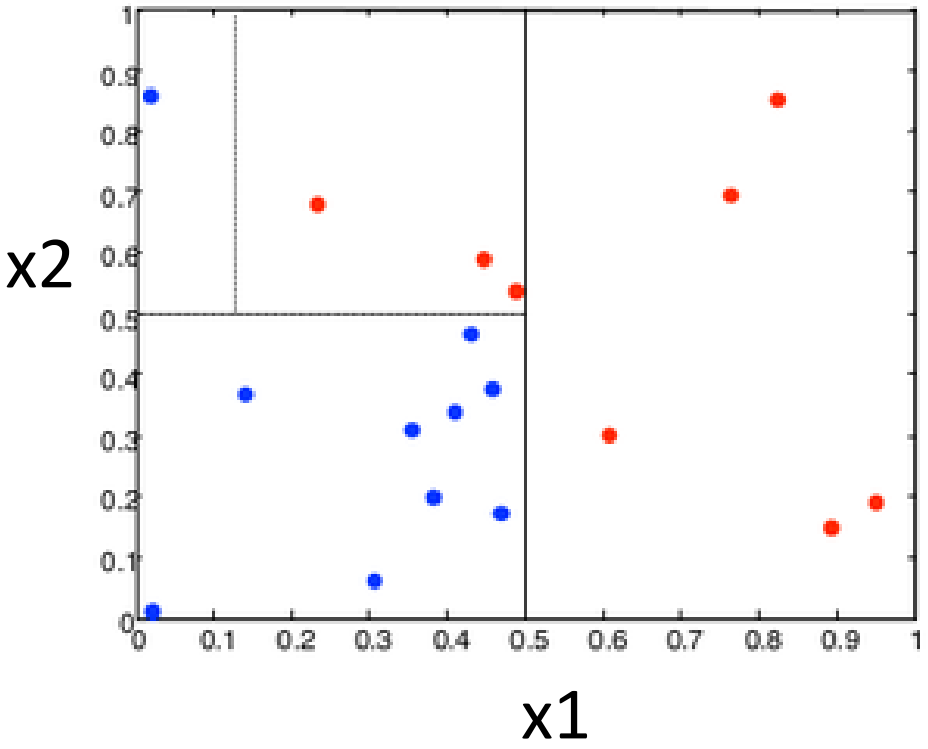
Classification and Regression Trees (CART)

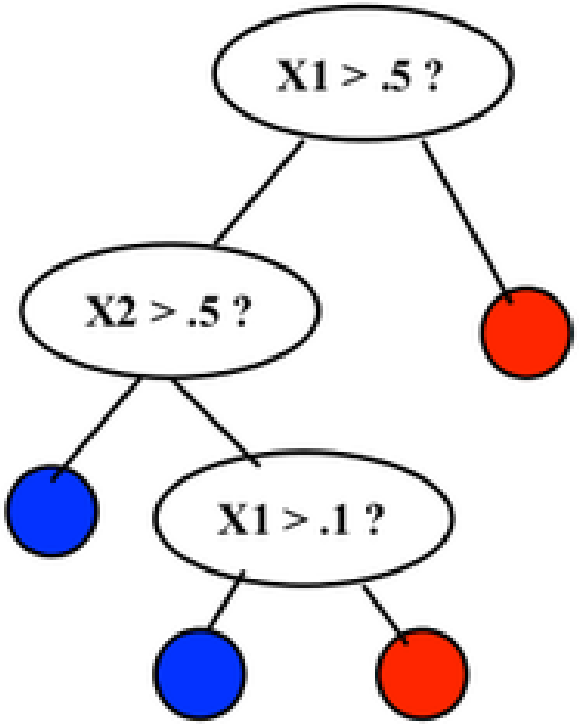



x2
x1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


x2
x1


x2
x1


x2
x1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


X1 > .5 ?
X2 > .5 ?
X1 > .1 ?

Regression Trees

1. Start with a sample of n observations $(x_1, y_1),\ldots, (x_n, y_n)$.
   In a regression tree, the response y is assumed to be continuous.
   x represents one or more covariates.

2. Use the covariates x to split the response y into two 'similar' groups, then split these groups, and so on, in a tree-like fashion.

3. The aim is to minimise the difference in responses within the groups and maximise the difference in average response between the groups.

4. The result is a set of subgroups with "acceptably" similar responses.

5. The predicted response for a new object is the average (and variance) of the observations in the terminal (leaf) node it falls into, based on following the branches of the tree.

Classification Trees

1. Start with a sample of n observations $(x_1, y_1),\ldots, (x_n, y_n)$.
   In a classification tree, the response y is assumed to be categorical
   x represents one or more covariates.

2. Use the covariates x to split the response y into two 'similar' groups, then split these groups, and so on, in a tree-like fashion.

3. The aim is to minimise the misclassification cost ('impurity').

4. The result is a set of subgroups with "acceptably" similar classes.

5. The predicted response for a new object is the most common (modal) class among the observations in the terminal (leaf) node it falls into, based on following the branches of the tree.

Breiman et al. (1984) Classification and Regression Trees. Wadsworth International Group.

Introductions to CART:

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

https://www.statmethods.net/advstats/cart.html

CART in Sport: Case Study

- 1,404 balanced games (score-differences: 1-14 points) from the Spanish EBA Basketball League.

- The games were split into faster- and slower-paced games according to ball possessions per game (using a cluster k-means).

- CART was used to predict which game-related variable/s better classified winning and losing teams during slower- and faster-paced games.

Gomez et al. The use of classification and regression tree when classifying winning and losing basketball teams. Kinesiology 49(2017)1:47-56

Game-related statistics

- 2-point and 3-point field-goals (successful & unsuccessful)
- free-throws (successful & unsuccessful)
- offensive & defensive rebounds, steals, turnovers, assists, blocks (performed & received)
- personal fouls (committed &received).

Situational variable effects: k-means cluster analyses

Game type: k=2 clusters

1. balanced games: score diffs 1-14 points (n=1,404 games)
2. unbalanced games: score diffs >15 points (n=670 games)

(Only the balanced games were included in the analysis.)

Game pace (for the balanced games): k=2 clusters

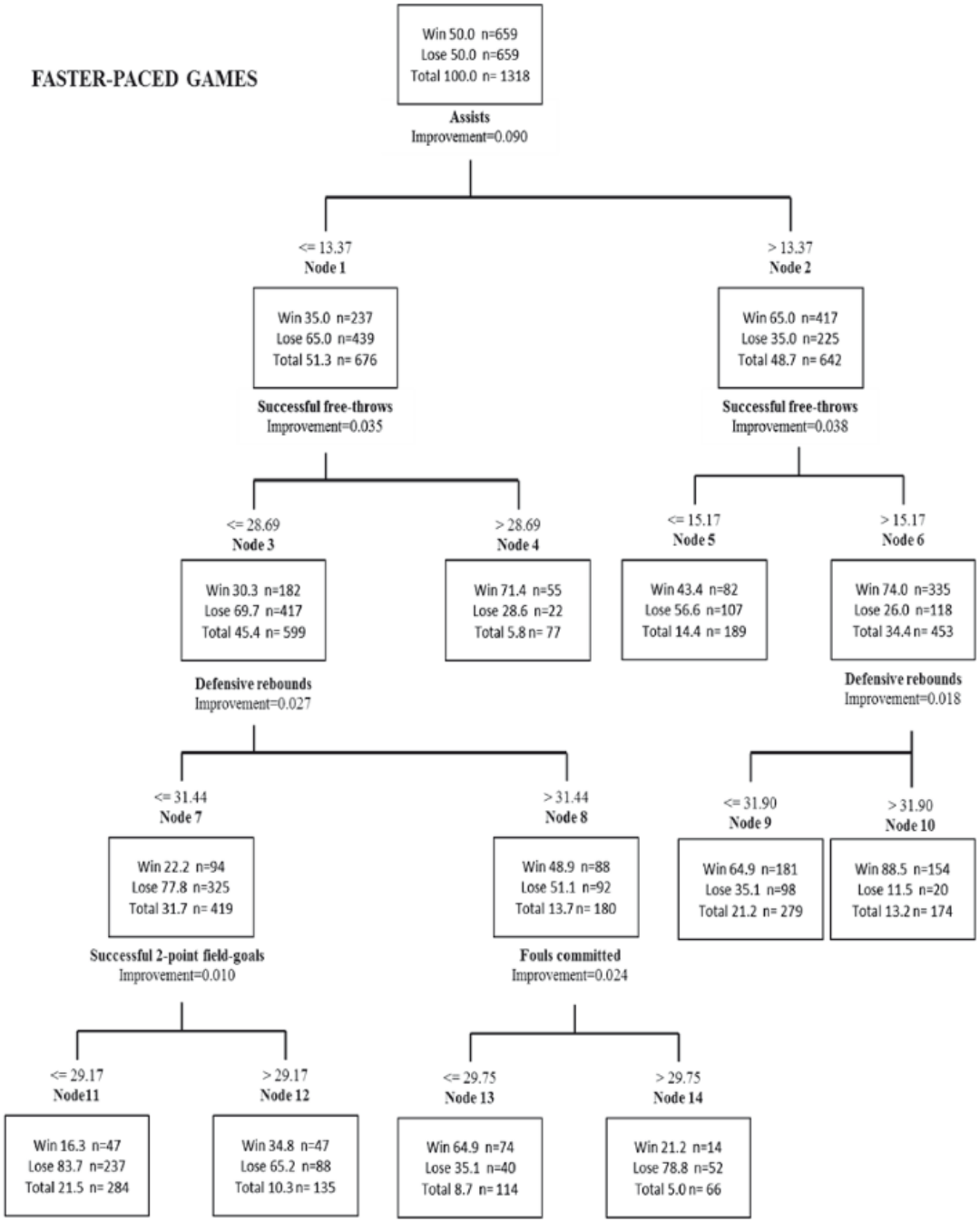
1. slower-paced games: 57.03±13.32 ball possessions (n=745 games)
2. faster-paced games: 79.31±6.45 ball possessions, n=659 games

Team quality: k=2 clusters

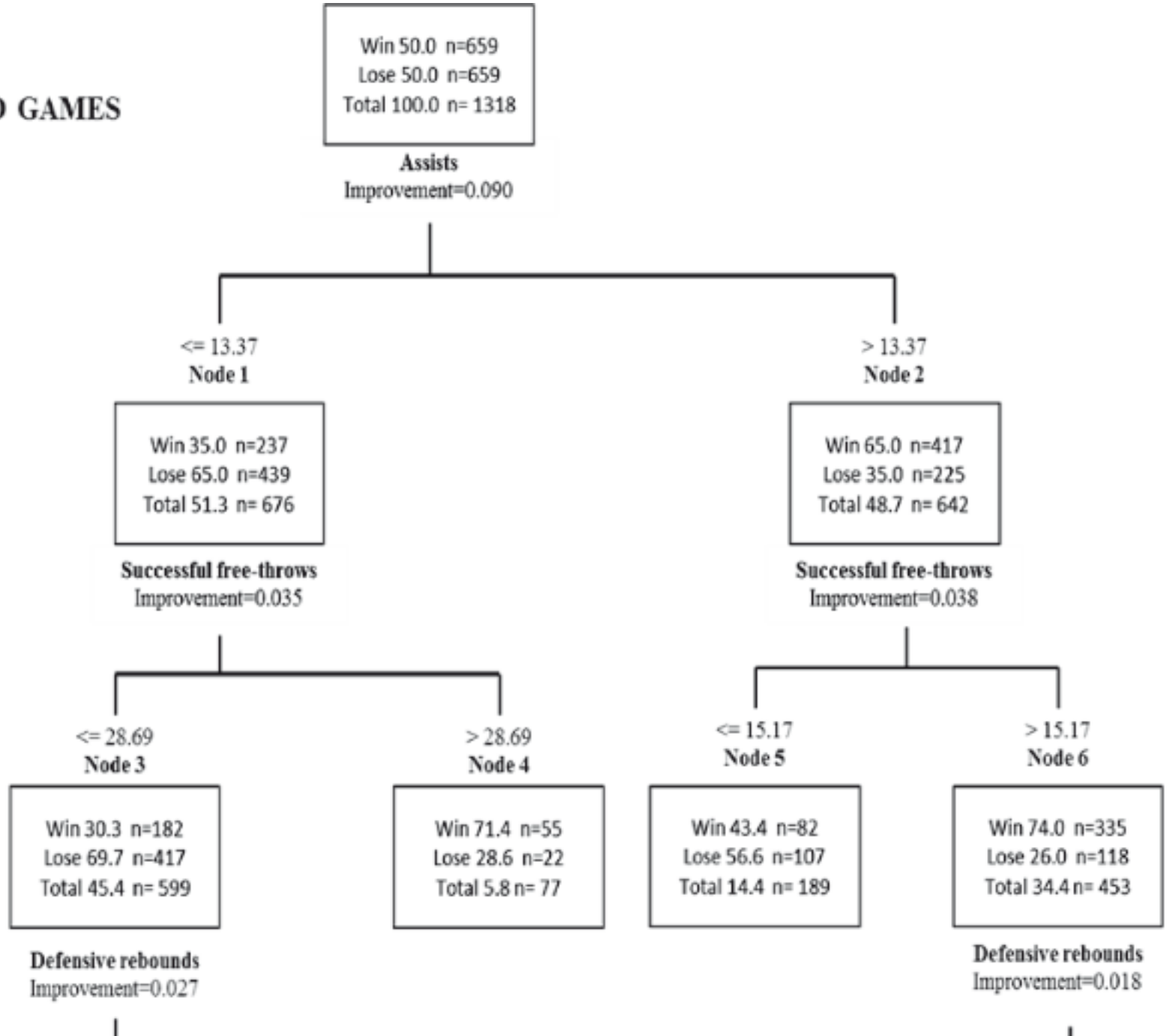
1. worst teams (winning%=42.25±9.32)
2. best teams (winning%=61.51±12.24)

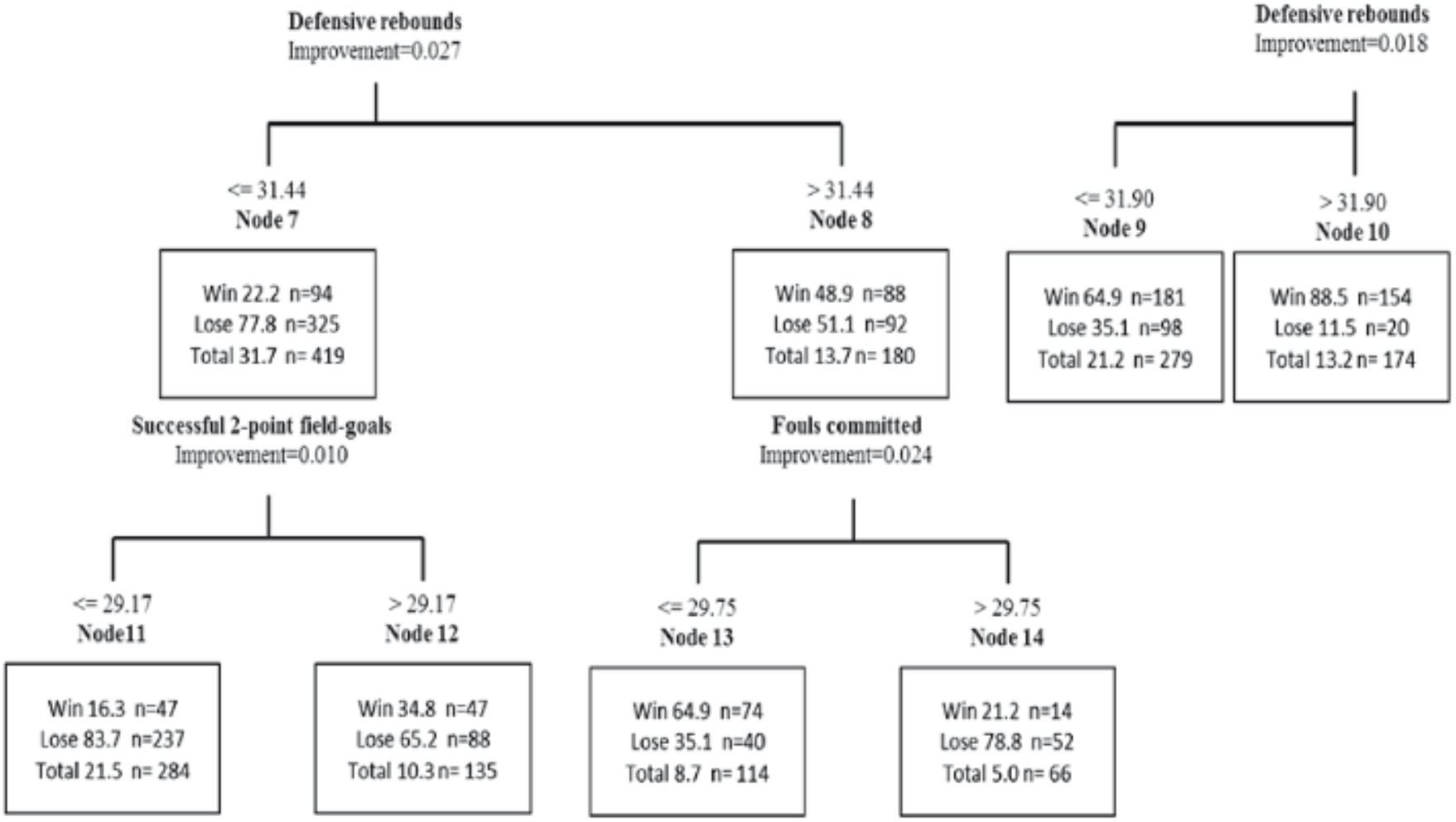
Game location: playing at home or away

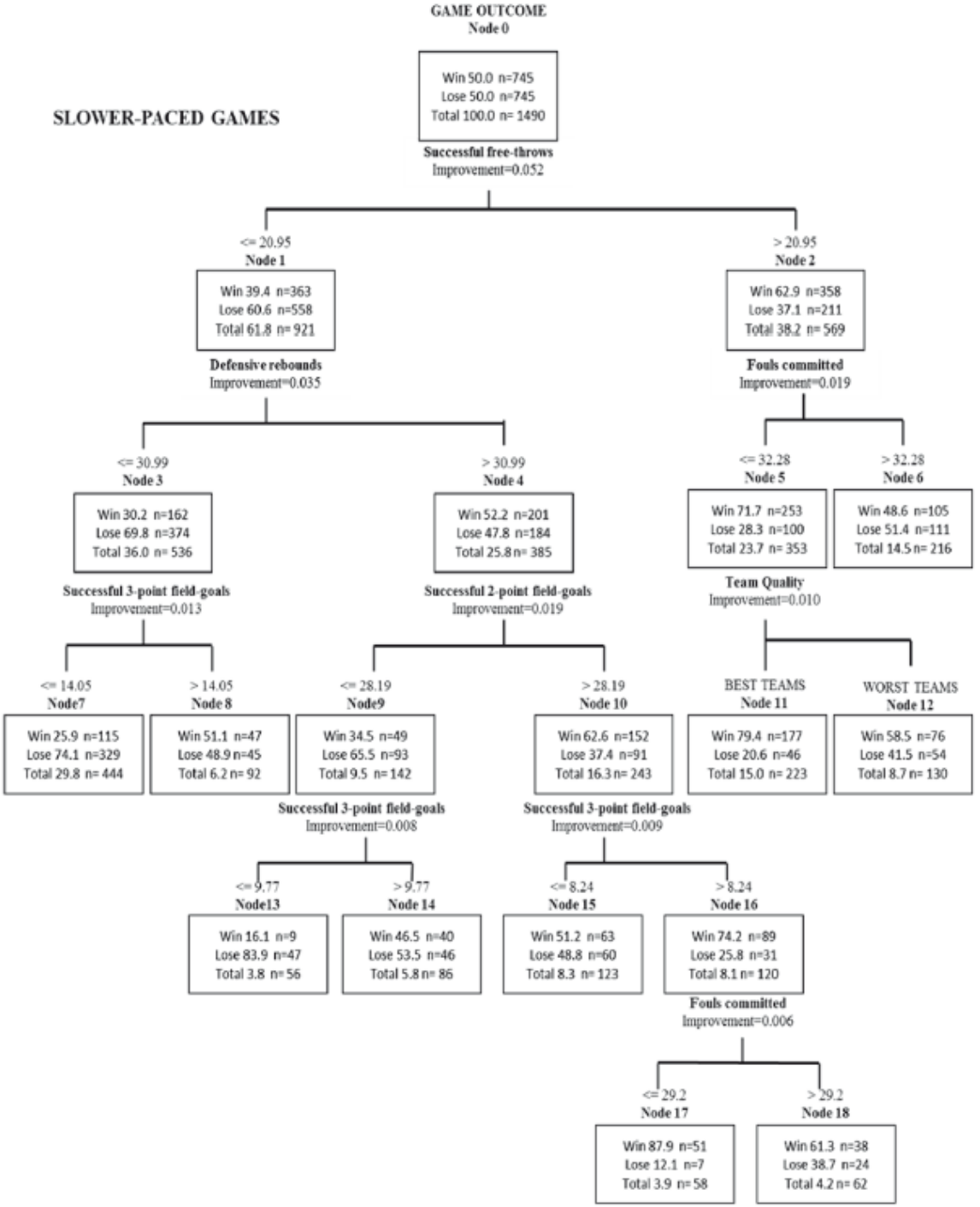
FASTER-PACED GAMES

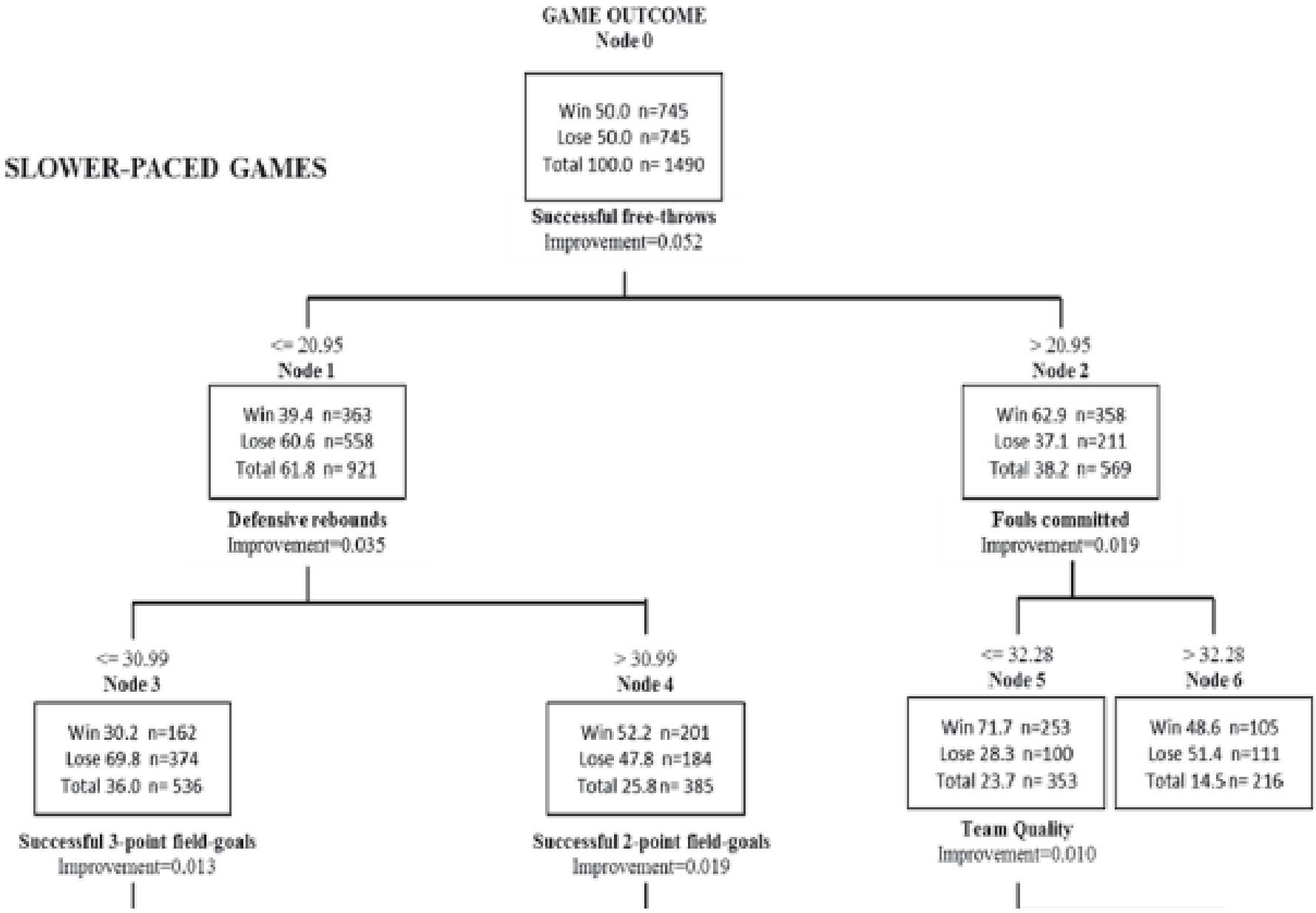


FASTER-PACED GAMES




Defensive rebounds
Improvement=0.027
Defensive rebounds
Improvement=0.018
<= 31.44
Node 7
> 31.44
Node 8
<= 31.90
Node 9
> 31.90
Node 10
Win 22.2 n=94
Lose 77.8 n=325
Total 31.7 n= 419
Win 48.9 n=88
Lose 51.1 n=92
Total 13.7 n= 180
Win 64.9 n=181
Lose 35.1 n=98
Total 21.2 n= 279
Win 88.5 n=154
Lose 11.5 n=20
Total 13.2 n= 174
Successful 2-point field-goals
Improvement=0.010
Fouls committed
Improvement=0.024
<= 29.17
Node11
> 29.17
Node 12
<= 29.75
Node 13
> 29.75
Node 14
Win 16.3 n=47
Lose 83.7 n=237
Total 21.5 n= 284
Win 34.8 n=47
Lose 65.2 n=88
Total 10.3 n= 135
Win 64.9 n=74
Lose 35.1 n=40
Total 8.7 n= 114
Win 21.2 n=14
Lose 78.8 n=52
Total 5.0 n= 66


SLOWER-PACED GAMES
GAME OUTCOME
Node 0
Win 50.0 n=745
Lose 50.0 n=745
Total 100.0 n= 1490
Successful free-throws
Improvement=0.052
<= 20.95
Node 1
Win 39.4 n=363
Lose 60.6 n=558
Total 61.8 n= 921
Defensive rebounds
Improvement=0.035
> 20.95
Node 2
Win 62.9 n=358
Lose 37.1 n=211
Total 38.2 n= 569
Fouls committed
Improvement=0.019
<= 30.99
Node 3
Win 30.2 n=162
Lose 69.8 n=374
Total 36.0 n= 536
Successful 3-point field-goals
Improvement=0.013
> 30.99
Node 4
Win 52.2 n=201
Lose 47.8 n=184
Total 25.8 n= 385
Successful 2-point field-goals
Improvement=0.019
<= 32.28
Node 5
Win 71.7 n=253
Lose 28.3 n=100
Total 23.7 n= 353
> 32.28
Node 6
Win 48.6 n=105
Lose 51.4 n=111
Total 14.5 n= 216
Team Quality
Improvement=0.010
<= 14.05
Node7
Win 25.9 n=115
Lose 74.1 n=329
Total 29.8 n= 444
> 14.05
Node 8
Win 51.1 n=47
Lose 48.9 n=45
Total 6.2 n= 92
<= 28.19
Node9
Win 34.5 n=49
Lose 65.5 n=93
Total 9.5 n= 142
> 28.19
Node 10
Win 62.6 n=152
Lose 37.4 n=91
Total 16.3 n= 243
BEST TEAMS
Node 11
Win 79.4 n=177
Lose 20.6 n=46
Total 15.0 n= 223
WORST TEAMS
Node 12
Win 58.5 n=76
Lose 41.5 n=54
Total 8.7 n= 130
Successful 3-point field-goals
Improvement=0.008
Successful 3-point field-goals
Improvement=0.009
<= 9.77
Node13
Win 16.1 n=9
Lose 83.9 n=47
Total 3.8 n= 56
> 9.77
Node 14
Win 46.5 n=40
Lose 53.5 n=46
Total 5.8 n= 86
<= 8.24
Node 15
Win 51.2 n=63
Lose 48.8 n=60
Total 8.3 n= 123
> 8.24
Node 16
Win 74.2 n=89
Lose 25.8 n=31
Total 8.1 n= 120
Fouls committed
Improvement=0.006
<= 29.2
Node 17
Win 87.9 n=51
Lose 12.1 n=7
Total 3.9 n= 58
> 29.2
Node 18
Win 61.3 n=38
Lose 38.7 n=24
Total 4.2 n= 62


SLOWER-PACED GAMES
GAME OUTCOME
Node 0
Win 50.0 n=745
Lose 50.0 n=745
Total 100.0 n= 1490
Successful free-throws
Improvement=0.052
<= 20.95
Node 1
Win 39.4 n=363
Lose 60.6 n=558
Total 61.8 n= 921
Defensive rebounds
Improvement=0.035
> 20.95
Node 2
Win 62.9 n=358
Lose 37.1 n=211
Total 38.2 n= 569
Fouls committed
Improvement=0.019
<= 30.99
Node 3
Win 30.2 n=162
Lose 69.8 n=374
Total 36.0 n= 536
Successful 3-point field-goals
Improvement=0.013
> 30.99
Node 4
Win 52.2 n=201
Lose 47.8 n=184
Total 25.8 n= 385
Successful 2-point field-goals
Improvement=0.019
<= 32.28
Node 5
Win 71.7 n=253
Lose 28.3 n=100
Total 23.7 n= 353
Team Quality
Improvement=0.010
> 32.28
Node 6
Win 48.6 n=105
Lose 51.4 n=111
Total 14.5 n= 216

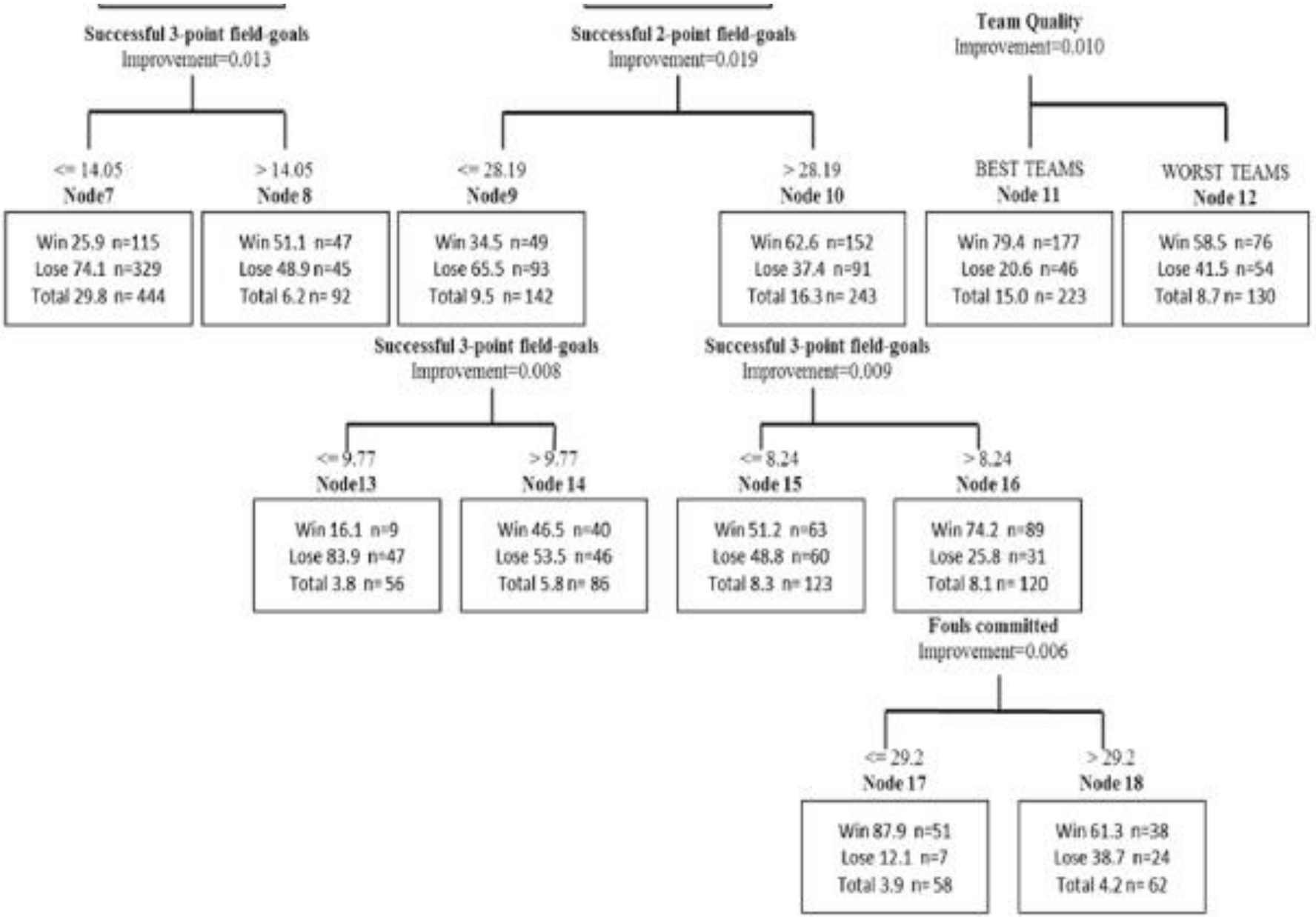

Successful 3-point field-goals
Improvement=0.013
<= 14.05
Node7
Win 25.9 n=115
Lose 74.1 n=329
Total 29.8 n= 444
> 14.05
Node 8
Win 51.1 n=47
Lose 48.9 n=45
Total 6.2 n= 92
Successful 2-point field-goals
Improvement=0.019
<= 28.19
Node9
Win 34.5 n=49
Lose 65.5 n=93
Total 9.5 n= 142
> 28.19
Node 10
Win 62.6 n=152
Lose 37.4 n=91
Total 16.3 n= 243
Team Quality
Improvement=0.010
BEST TEAMS
Node 11
Win 79.4 n=177
Lose 20.6 n=46
Total 15.0 n= 223
WORST TEAMS
Node 12
Win 58.5 n=76
Lose 41.5 n=54
Total 8.7 n= 130
Successful 3-point field-goals
Improvement=0.008
<= 9.77
Node13
Win 16.1 n=9
Lose 83.9 n=47
Total 3.8 n= 56
> 9.77
Node 14
Win 46.5 n=40
Lose 53.5 n=46
Total 5.8 n= 86
Successful 3-point field-goals
Improvement=0.009
<= 8.24
Node 15
Win 51.2 n=63
Lose 48.8 n=60
Total 8.3 n= 123
> 8.24
Node 16
Win 74.2 n=89
Lose 25.8 n=31
Total 8.1 n= 120
Fouls committed
Improvement=0.006
<= 29.2
Node 17
Win 87.9 n=51
Lose 12.1 n=7
Total 3.9 n= 58
> 29.2
Node 18
Win 61.3 n=38
Lose 38.7 n=24
Total 4.2 n= 62

Table 1. Cross-validation results of CRT models: estimations (standard errors of classification) and importance of the independent variables for faster- and slower-paced games

Faster-paced games			Slower-paced games		
	Importance			Importance	
	Absolute	%		Absolute	%
Defensive rebounds	0.10	100	Successful free-throws	0.06	100
Successful free-throws	0.10	94.73	Defensive rebounds	0.05	82.27
Assists	0.09	86.13	Committed fouls	0.04	68.44
Committed fouls	0.06	55.86	Assists	0.04	66.95
Team ability	0.05	46.70	Successful 2-pt field-goal	0.03	62.24
Game location	0.05	44.05	Successful 3-pt field-goal	0.03	61.59
Received fouls	0.04	34.16	Received fouls	0.02	37.93
Successful 2-pt field-goal	0.04	33.73	Unsuccessful 2-pt field-goals	0.02	35.69
Unsuccessful 3-pt field-goals	0.03	30.63	Blocks received	0.02	34.58
Unsuccessful 2-pt field-goals	0.03	26.12	Team ability	0.02	31.23
Unsuccessful free-throws	0.02	23.35	Unsuccessful 3-pt field-goals	0.02	30.26
Blocks made	0.02	21.33	Unsuccessful free-throws	0.01	22.89
Steals	0.02	19.90	Turnovers	0.01	17.92
Blocks received	0.02	17.44	Steals	0.01	9.86
Successful 3-pt field-goal	0.01	8.63	Blocks made	0.01	9.60
Turnovers	0.01	5.97	Offensive rebounds	0.00	6.95
Offensive rebounds	0.00	3.11	Game location	0.00	0.49
	Estimation (SE)			Estimation (SE)	
Cross-validation	.280 (.013)		Cross-validation	.307 (.013)	

Case Study: Conclusions

- The CART analysis explained 72% of the total variance in the slower- and 69.3% in the faster-paced games.

- The results identified important variables for classification of winning or losing teams:
  - For fast-paced games: defensive rebounds, successful free-throws, assists, and fouls committed
  - For slow-paced games: successful free-throws, defensive rebounds, fouls committed, assists, successful 2- point and 3-point field-goals.

- The influence of situational variables was identified only for team quality in the slow-paced games.

The findings allow coaches for a better control of games and competition.

Gomez et al. The use of classification and regression tree when classifying winning and losing basketball teams. Kinesiology 49(2017)1:47-56

Your turn:

- Read the following journal article and answer the questions below:

Longo et al. (2016) Age of peak performance in Olympic sports. Journal of Human Sport & Exercise.

https://rua.ua.es/dspace/bitstream/10045/61889/1/jhse_Vol_11_N_1_31-41.pdf

1. What was the aim of the analysis?
2. Briefly describe the CART approach used in the analysis.
3. What software did the authors use to fit the CART model?
4. What were the main results obtained from the CART analysis?
5. Identify one advantage and one disadvantage of using CART for this problem.

Fitting CART in R: setup

Tutorial:
Cross.
Decision trees to predict NFL play outcomes
https://rpubs.com/jcross/nfl_trees

Dataset:
Trice.
NFLPlaybyPlay2015.csv
https://github.com/timtrice/datasets/find/master

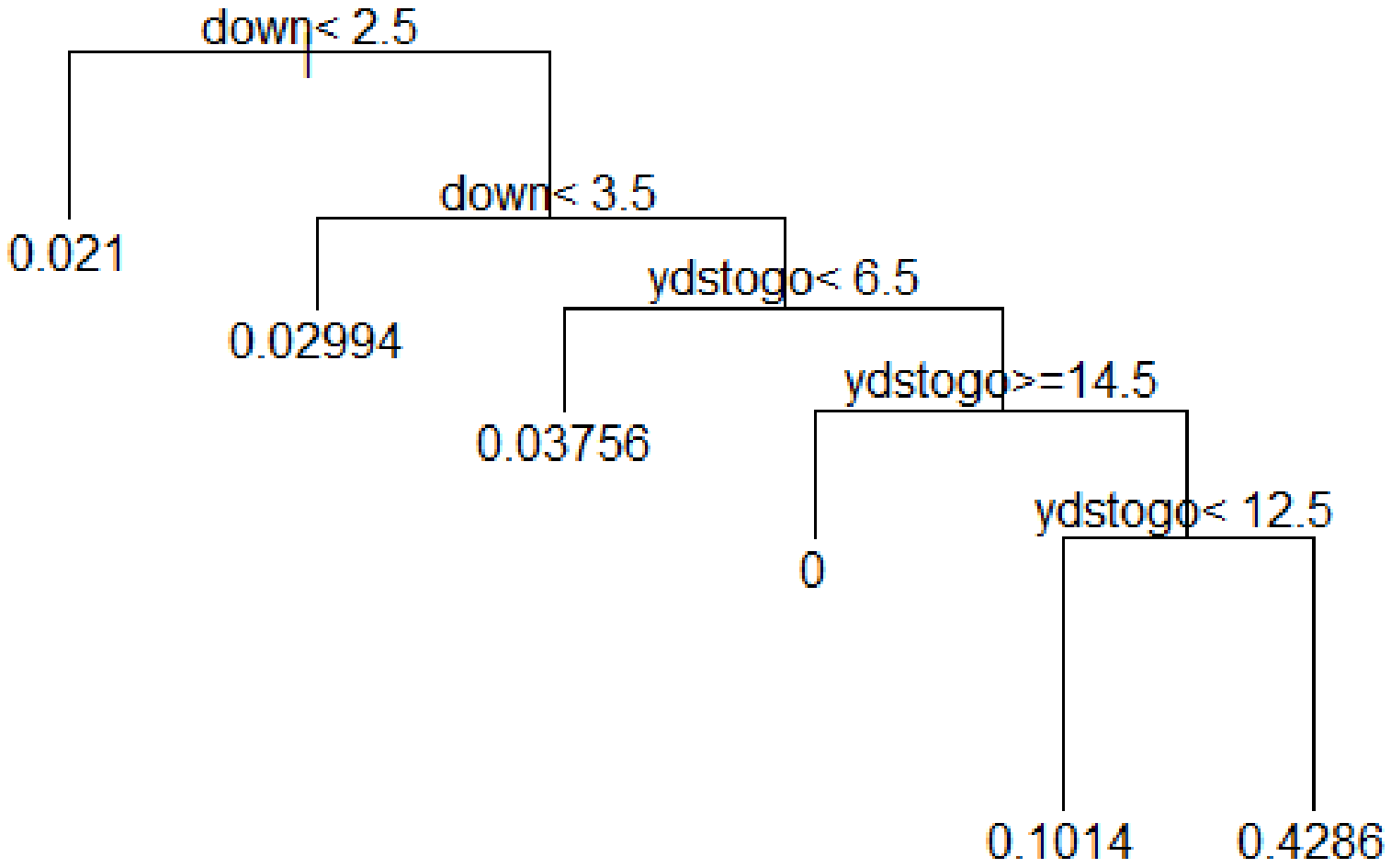
Fitting CART in R: setup

```
# set the working directory
> setwd("c://Work/Work18/courses/stats_in_sport_bne_0218")

# read the Excel (csv) data file
> nfl <- read.csv('NFLPlaybyPlay2015.csv')

# view the file within R
> View(nfl)

# install and load libraries
> install.packages(c("dplyr", "rpart", "rpart.plot"))
> library(dplyr);library(rpart); library(rpart.plot)
```

Fitting CART in R: analyse

```
# First, make a data frame of only pass plays.
> pass <- nfl %>% filter(PlayType %in% c("Pass"))

# Now predict “InterceptionThrown” using “down” and “distance”.
> fit <- rpart(InterceptionThrown ~ down+ydstogo,data=pass, cp=0.0005)

# One way to plot your results
> par(mfrow=c(1,1), xpd=TRUE)
> plot(fit)
> text(fit)

# Another way to plot your results
> prp(fit, type=1, fallen.leaves=TRUE, extra=1, cex=0.7)

# Display your results
> fit
```

Fitting CART in R: results



Fitting CART in R: results

```
n= 18323

node), split, n, deviance, yval
      * denotes terminal node

 1) root 18323 430.386000 0.02406811
   2) down< 2.5 12999 267.266600 0.02100162 *
   3) down>=2.5 5324 162.698700 0.03155522
     6) down< 3.5 5010 145.509000 0.02994012 *
     7) down>=3.5 314  16.968150 0.05732484
      14) ydstogo< 6.5 213   7.699531 0.03755869 *
      15) ydstogo>=6.5 101   9.009901 0.09900990
        30) ydstogo>=14.5 25   0.000000 0.00000000 *
        31) ydstogo< 14.5 76   8.684211 0.13157890
          62) ydstogo< 12.5 69   6.289855 0.10144930 *
          63) ydstogo>=12.5 7   1.714286 0.42857140 *
```

Fitting CART in R: interpret

Q1: Explain your results.

Q2: Try adding the quarter and score differential to your model ("qtr" and "ScoreDiff").
You may want to reduce the complexity parameter.
What does your model show?

Fitting CART in R: analyse

Say we want to predict whether the upcoming play is a run or a pass.
First, eliminate special teams plays and
create a column **run** = 1 if the team ran the ball and 0 otherwise

```
> nfl.run.or.pass <- nfl %>% filter(PlayType %in% c("Run", "Pass")) %>%
mutate(run = ifelse(PlayType=="Run", 1, 0))
```

> Now predict whether a team will run the ball using the down and distance.

```
> fit <- rpart(run ~ down+ydstogo,data=nfl.run.or.pass, cp=0.01)
> prp(fit, type=1, fallen.leaves=TRUE, extra=1, cex=0.7)
```

Fitting CART in R: interpret

Q1: Explain your results.

Q2: Try adding the quarter and score differential to your model ("qtr" and "ScoreDiff").
You may want to reduce the complexity parameter.
What does your model show?

Comments on CART

- Advantages:
  - Popular “off-the-shelf” procedure for data mining
  - No need to scale or transform variables (although sometimes you may want to in practice)
  - Robust to inclusion of irrelevant features
  - Produces inspectable models.
- Drawbacks:
  - They are seldom accurate!
  - Control parameters can manage bias and variance (overfitting and accuracy) to some extent, but…
  - Trees that are grown very deep tend to overfit their training sets, i.e. have low bias but very high variance.

https://en.wikipedia.org/wiki/Random_forest

Improving CART

- Boosting

build consecutive sets of trees using misclassified observations from the previous tree

- Bagging

build multiple trees based on samples of the data; predict from all the trees and average the predictions

- Random forests

build many shallow trees; predict from all the trees and average the predictions

- Combine with other methods

e.g. CART + Logistic regression

Fitting BRT in R: setup and analyse

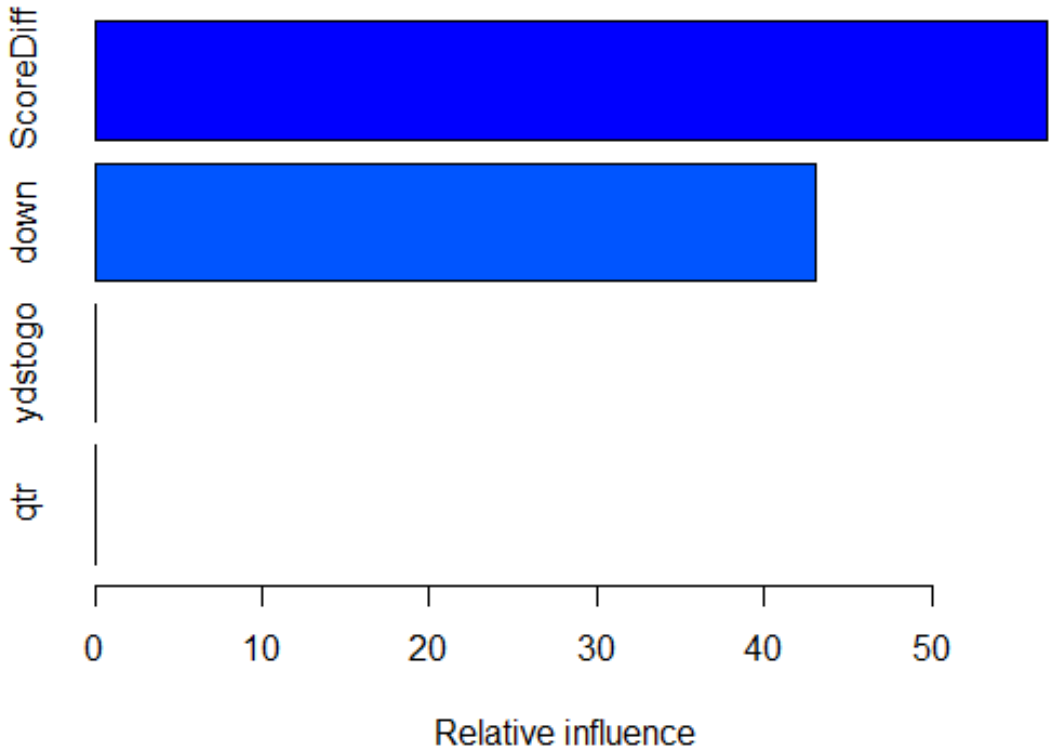
```
# install and load the package "gbm"
> install.packages("gbm")
> library(gbm)

# fit a boosted regression tree to predict "InterceptionThrown"
> fit <- gbm(InterceptionThrown ~ down+ydstogo+qtr+ScoreDiff,data=pass)

# display results
> summary(fit)

# plot results
> plot(fit)
```

Excellent article on BRTs: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2656.2008.01390.x/pdf

Fitting BRT in R: results

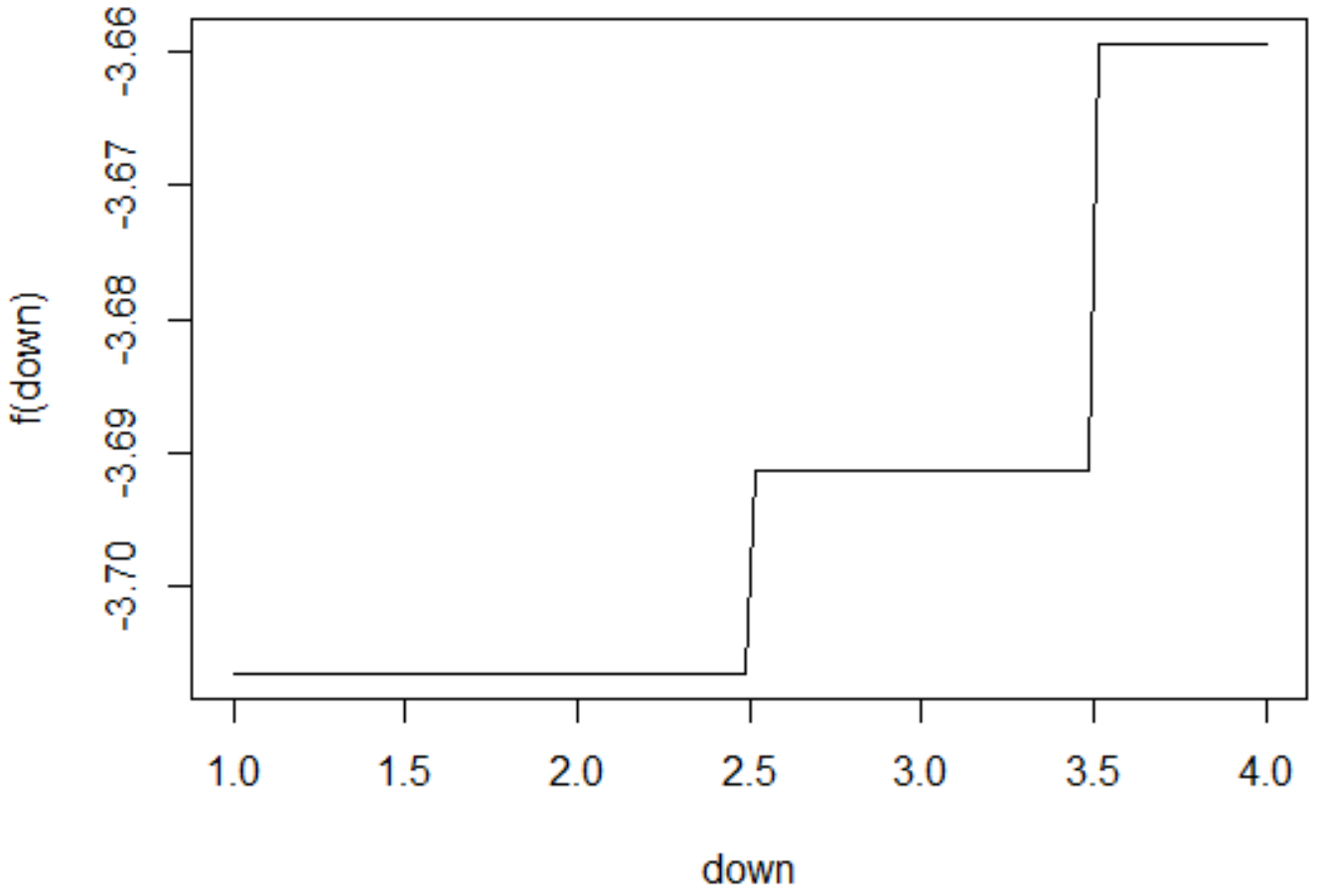
```
> summary(fit)
```

var	rel.inf
ScoreDiff	56.90018
down	43.09982
ydstogo	0.00000
qtr	0.00000



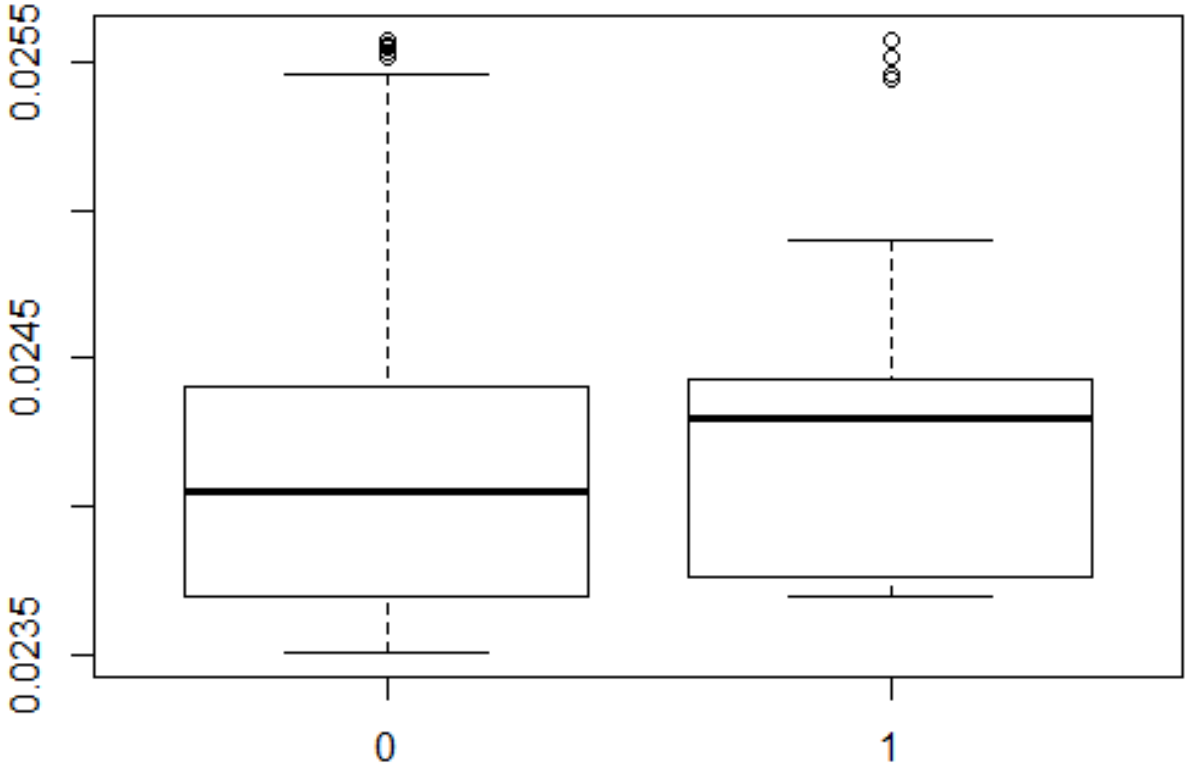
Fitting BRT in R: results

```
> plot(fit)
```



Fitting BRT in R: results

```
> fit.p <- predict.gbm(fit, newdata=pass, n.trees=100, type="response")

> boxplot(fit.predict~pass$InterceptionThrown)
```



No. observations

0	1
17882	441

Further reading: BRT and RF

Excellent article on BRTs:
http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2656.2008.01390.x/pdf

Lock and Nettleton (2014) Using random forests to estimate win probability before each play of an NFL game. JQAS.

http://homepage.divms.uiowa.edu/~dzimmer/sports-statistics/nettletonandlock.pdf

Tat (2017) Seeing the random forest from the decision trees: An explanation of Random Forest. (with R code)

https://towardsdatascience.com/seeing-the-random-forest-from-the-decision-trees-an-intuitive-explanation-of-random-forest-beaa2d6a0d80

Further reading

- Comparison of models: Logistic regression and CART

Jayalath (2017) A machine learning approach to analyze ODI cricket predictors. Journal of Sports Analytics.

- Comparison of models: Logistic regression and Random Forest

Wundersitz et al. (2015) Classification of team sport activities using a single wearable tracking device. Journal of Biomechanics.

- Other methods: Neural Networks

Bunker and Thabtah (2017) A machine learning framework for sport result prediction. Applied Computing and Informatics.

https://www.youtube.com/watch?v=aircAruvnKk&list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi&index=1