More Data using Machine Learning Methods

Alison Jing Huang

5/6/2018

Load the data

```
##   X Y month day FFMC  DMC    DC  ISI temp RH wind rain area
## 1 7 5     8   1 86.2 26.2  94.3  5.1  8.2 51  6.7  0.0    0
## 2 7 4    11   6 90.6 35.4 669.1  6.7 18.0 33  0.9  0.0    0
## 3 7 4    11   3 90.6 43.7 686.9  6.7 14.6 33  1.3  0.0    0
## 4 8 6     8   1 91.7 33.3  77.5  9.0  8.3 97  4.0  0.2    0
## 5 8 6     8   4 89.3 51.3 102.2  9.6 11.4 99  1.8  0.0    0
## 6 8 6     2   4 92.3 85.3 488.0 14.7 22.2 29  5.4  0.0    0
```

Create the training and test dataset with 70/30 Rule

```
set.seed(517)
trainRowNumbers <- createDataPartition(df$area, p = 0.7, list = FALSE)
trainData <- df[trainRowNumbers,]
testData <- df[-trainRowNumbers,]

x = trainData[, 1:12]
y = trainData$area
```

Validate 70/30 rule by checking size of training datset vs. test data

```
# Dimension of trainning set
dim(trainData)
```

```
## [1] 364  13
```

```
# Dimension of test set
dim(testData)
```

```
## [1] 153  13
```

Preprocessing & Trainning

First check to see the data contains any missing values by using anyNA() method. “FALSE” indicates the the dataset contains all the values.

```
anyNA(df)
```

```
## [1] FALSE
```

10-Fold Cross Validation

```
trctrl <- trainControl(method='repeatedcv', number = 10, verboseIter = TRUE)
set.seed(1034)
dtree_fit1 <- train(area ~., data= trainData, method ="rpart", parms = list(split = "information"), trC
```

```
## + Fold01.Rep1: cp=0.0006389
## - Fold01.Rep1: cp=0.0006389
## + Fold02.Rep1: cp=0.0006389
## - Fold02.Rep1: cp=0.0006389
## + Fold03.Rep1: cp=0.0006389
## - Fold03.Rep1: cp=0.0006389
## + Fold04.Rep1: cp=0.0006389
## - Fold04.Rep1: cp=0.0006389
## + Fold05.Rep1: cp=0.0006389
## - Fold05.Rep1: cp=0.0006389
## + Fold06.Rep1: cp=0.0006389
## - Fold06.Rep1: cp=0.0006389
## + Fold07.Rep1: cp=0.0006389
## - Fold07.Rep1: cp=0.0006389
## + Fold08.Rep1: cp=0.0006389
## - Fold08.Rep1: cp=0.0006389
## + Fold09.Rep1: cp=0.0006389
## - Fold09.Rep1: cp=0.0006389
## + Fold10.Rep1: cp=0.0006389
## - Fold10.Rep1: cp=0.0006389
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.
```

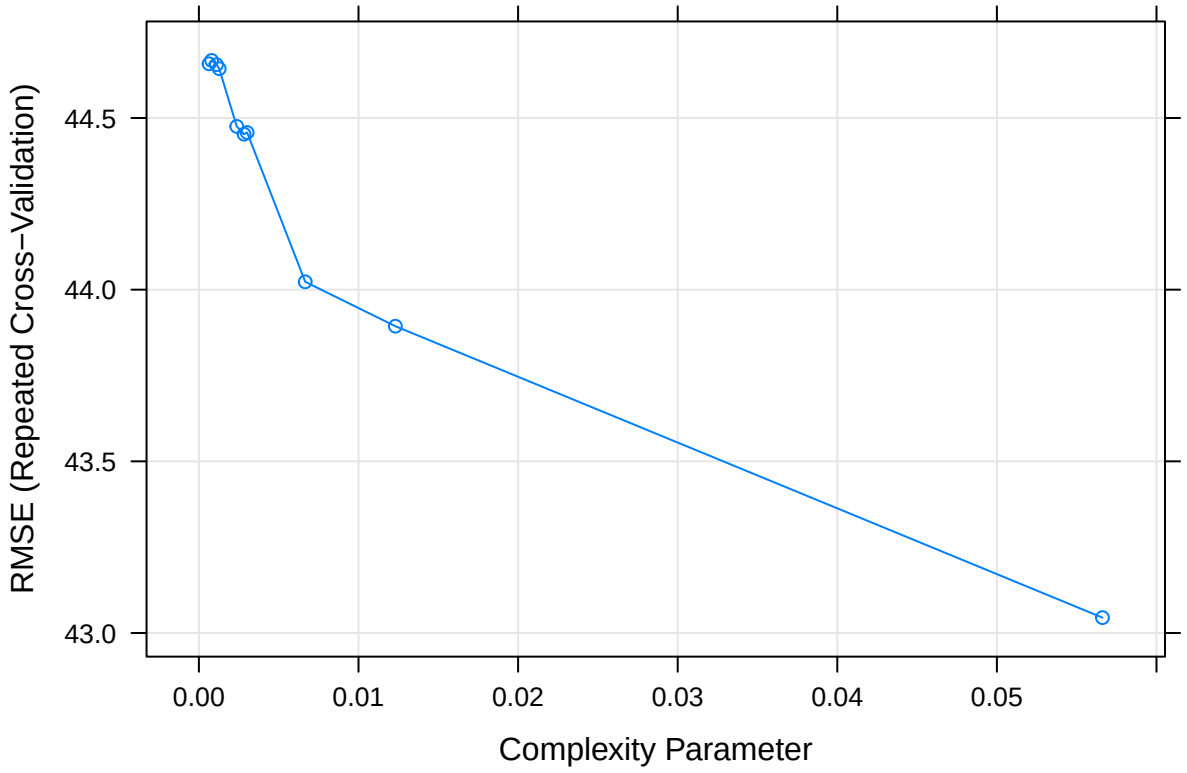
```
## Aggregating results
## Selecting tuning parameters
## Fitting cp = 0.0566 on full training set
```

```
dtree_fit1
```

```
## CART
##
## 364 samples
##  12 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 328, 328, 326, 328, 327, 328, ...
## Resampling results across tuning parameters:
##
##   cp            RMSE      Rsquared     MAE
##   0.0006389014  44.65793  0.007186379  18.16086
##   0.0007990614  44.66766  0.007308091  18.22733
##   0.0011021913  44.65528  0.008071924  18.17416
##   0.0012725506  44.64357  0.008110482  18.21701
##   0.0023726900  44.47553  0.007502064  18.02801
##   0.0028315286  44.45286  0.007590684  18.03090
##   0.0030231196  44.45773  0.007617076  18.04684
##   0.0066635890  44.02287  0.006482179  17.78441
```

```
##   0.0123115695  43.89363  0.004912405  17.78639
##   0.0566147526  43.04478  0.008654396  17.43427
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.05661475.
```

```
plot(dtree_fit1)
```



5 X 5 Fold Cross-Validation

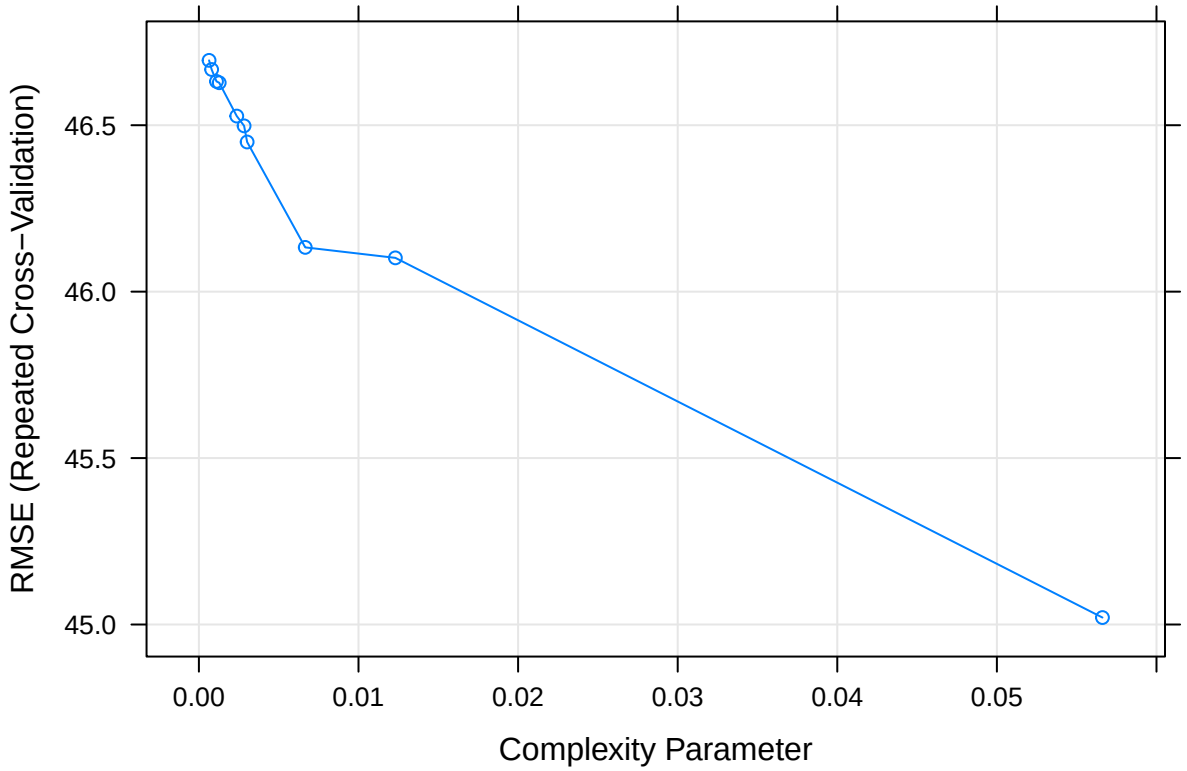
```
trctrl <- trainControl(method='repeatedcv', number = 5, repeats = 5, verboseIter = TRUE)
set.seed(1034)
dtree_fit2 <- train(area ~., data= trainData, method ="rpart", parms = list(split = "information"), trC
```

```
## + Fold1.Rep1: cp=0.0006389
## - Fold1.Rep1: cp=0.0006389
## + Fold2.Rep1: cp=0.0006389
## - Fold2.Rep1: cp=0.0006389
## + Fold3.Rep1: cp=0.0006389
## - Fold3.Rep1: cp=0.0006389
## + Fold4.Rep1: cp=0.0006389
## - Fold4.Rep1: cp=0.0006389
## + Fold5.Rep1: cp=0.0006389
## - Fold5.Rep1: cp=0.0006389
## + Fold1.Rep2: cp=0.0006389
## - Fold1.Rep2: cp=0.0006389
## + Fold2.Rep2: cp=0.0006389
## - Fold2.Rep2: cp=0.0006389
```

```
## + Fold3.Rep2: cp=0.0006389
## - Fold3.Rep2: cp=0.0006389
## + Fold4.Rep2: cp=0.0006389
## - Fold4.Rep2: cp=0.0006389
## + Fold5.Rep2: cp=0.0006389
## - Fold5.Rep2: cp=0.0006389
## + Fold1.Rep3: cp=0.0006389
## - Fold1.Rep3: cp=0.0006389
## + Fold2.Rep3: cp=0.0006389
## - Fold2.Rep3: cp=0.0006389
## + Fold3.Rep3: cp=0.0006389
## - Fold3.Rep3: cp=0.0006389
## + Fold4.Rep3: cp=0.0006389
## - Fold4.Rep3: cp=0.0006389
## + Fold5.Rep3: cp=0.0006389
## - Fold5.Rep3: cp=0.0006389
## + Fold1.Rep4: cp=0.0006389
## - Fold1.Rep4: cp=0.0006389
## + Fold2.Rep4: cp=0.0006389
## - Fold2.Rep4: cp=0.0006389
## + Fold3.Rep4: cp=0.0006389
## - Fold3.Rep4: cp=0.0006389
## + Fold4.Rep4: cp=0.0006389
## - Fold4.Rep4: cp=0.0006389
## + Fold5.Rep4: cp=0.0006389
## - Fold5.Rep4: cp=0.0006389
## + Fold1.Rep5: cp=0.0006389
## - Fold1.Rep5: cp=0.0006389
## + Fold2.Rep5: cp=0.0006389
## - Fold2.Rep5: cp=0.0006389
## + Fold3.Rep5: cp=0.0006389
## - Fold3.Rep5: cp=0.0006389
## + Fold4.Rep5: cp=0.0006389
## - Fold4.Rep5: cp=0.0006389
## + Fold5.Rep5: cp=0.0006389
## - Fold5.Rep5: cp=0.0006389

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info =
## trainInfo, : There were missing values in resampled performance measures.

## Aggregating results
## Selecting tuning parameters
## Fitting cp = 0.0566 on full training set
dtree_fit2

## CART
##
## 364 samples
##  12 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold, repeated 5 times)
## Summary of sample sizes: 292, 290, 291, 291, 292, 292, ...
## Resampling results across tuning parameters:
##
```

```
##    cp            RMSE      Rsquared    MAE
##    0.0006389014  46.69479  0.01412200  17.70691
##    0.0007990614  46.66742  0.01406078  17.66687
##    0.0011021913  46.63163  0.01460787  17.64243
##    0.0012725506  46.62763  0.01478830  17.59299
##    0.0023726900  46.52748  0.01430154  17.54266
##    0.0028315286  46.49803  0.01406254  17.55108
##    0.0030231196  46.44955  0.01396697  17.48534
##    0.0066635890  46.13294  0.01475282  17.26142
##    0.0123115695  46.10152  0.01375737  17.11528
##    0.0566147526  45.02088  0.01485337  16.69526
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was cp = 0.05661475.
```

```
plot(dtree_fit2)
```



Recursive Feature Elimination(RFE) for choosing important features

```
set.seed(1034)
options(warn=-1)

subsets <- c(1:12)
ctrl <- rfeControl(functions=rfFuncs, method='repeatedcv', repeats= 5, verbose = FALSE)

lmProfile <- rfe(x=trainData[, 1:12], y=trainData$area, size=subsets, rfeControl = ctrl)

lmProfile
```

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold, repeated 5 times)
##
## Resampling performance over subset size:
##
##  Variables  RMSE Rsquared   MAE RMSESD RsquaredSD MAESD Selected
##          1 40.60  0.01974 17.96  31.35    0.02890 7.370
##          2 40.65  0.02660 17.61  31.14    0.05245 6.883
##          3 40.97  0.02067 17.81  31.05    0.03271 6.802
##          4 39.37  0.01540 17.42  31.10    0.02162 6.635
##          5 38.55  0.01374 17.11  31.35    0.01643 6.603        *
##          6 40.49  0.02100 17.80  30.90    0.02917 6.732
##          7 40.28  0.02276 17.76  30.90    0.03497 6.651
##          8 40.13  0.01867 17.77  30.80    0.03522 6.449
##          9 40.82  0.02027 18.28  30.60    0.03752 6.452
##         10 40.55  0.02310 18.22  30.56    0.04167 6.468
##         11 39.76  0.02236 18.07  30.52    0.03937 6.417
##         12 39.72  0.02307 18.30  30.72    0.04041 6.678
##
## The top 5 variables (out of 5):
##    DC, RH, month, wind, DMC
```

Based on above RFE method result, the 5 most important features are identified as **DMC**, **DC**, **temp**, **month** and **ISI**.

Decision Tree Model

Model 1

1. Use all the factor variables with respect to the resposne varialble **AREA**

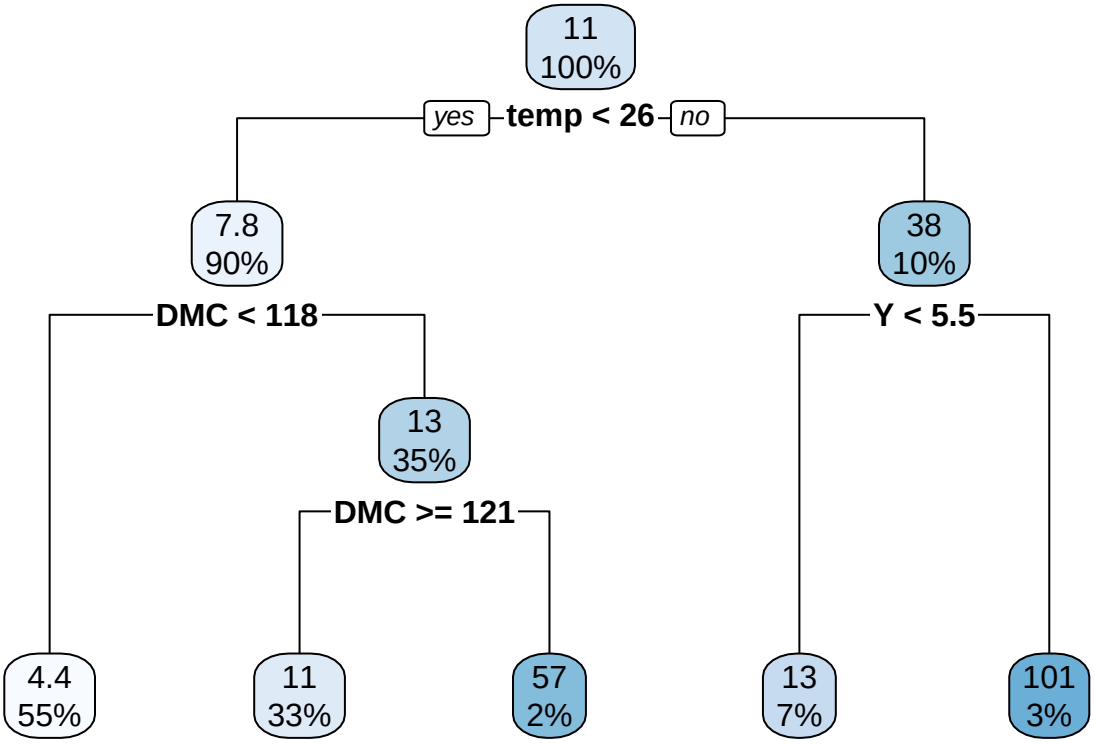
```
model1 <- rpart(area~ X + Y + month + day + FFMC + DMC + DC + ISI + temp + RH + wind + rain, data = tra
model1$terms <- eval(model1$call$formula)
summary(model1)
```

```
## Call:
## rpart(formula = area ~ X + Y + month + day + FFMC + DMC + DC +
##     ISI + temp + RH + wind + rain, data = trainData)
##   n= 364
##
##           CP nsplit rel error   xerror      xstd
## 1 0.05661475      0 1.0000000 1.004430 0.6772672
## 2 0.01231157      2 0.8867705 1.160355 0.6791242
## 3 0.01000000      4 0.8621474 1.170527 0.6726874
##
## Variable importance
##     Y  temp     X   DMC  FFMC   ISI    DC month
##    34    18    15    14     8     6     4     1
##
## Node number 1: 364 observations,    complexity param=0.05661475
##   mean=11.02681, MSE=2220.233
##   left son=2 (326 obs) right son=3 (38 obs)
```

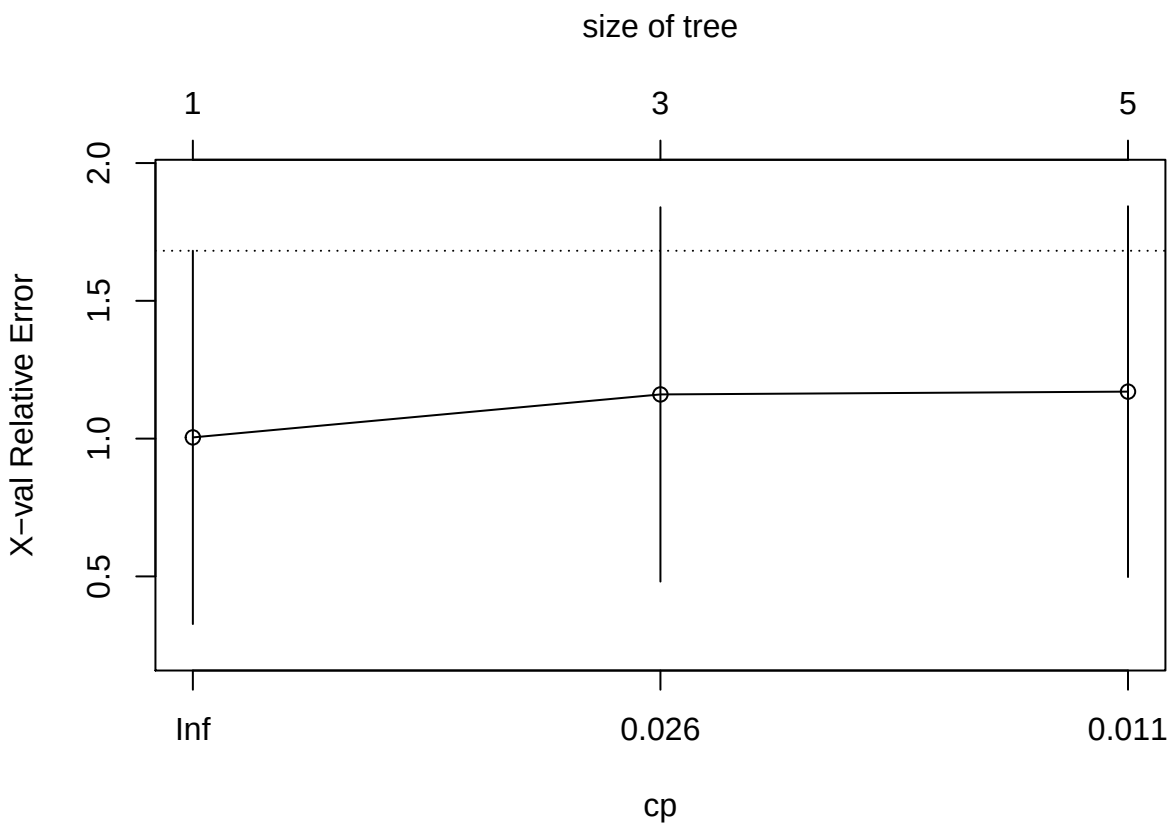
```
##   Primary splits:
##       temp < 26.15  to the left,  improve=0.03914212, (0 missing)
##       DMC  < 220.05 to the left,  improve=0.02649120, (0 missing)
##       X    < 7.5    to the left,  improve=0.01805966, (0 missing)
##       Y    < 5.5    to the left,  improve=0.01626588, (0 missing)
##       FFMC < 94.7   to the left,  improve=0.01540449, (0 missing)
##   Surrogate splits:
##       FFMC < 95.65  to the left,  agree=0.907, adj=0.105, (0 split)
##
## Node number 2: 326 observations,    complexity param=0.01231157
##   mean=7.844049, MSE=636.5423
##   left son=4 (199 obs) right son=5 (127 obs)
##   Primary splits:
##       DMC   < 118.45 to the left,  improve=0.02839467, (0 missing)
##       Y     < 2.5    to the right, improve=0.02248241, (0 missing)
##       DC    < 674.1  to the left,  improve=0.01142663, (0 missing)
##       month < 11.5   to the left,  improve=0.01065027, (0 missing)
##       wind  < 2.45   to the right, improve=0.01031726, (0 missing)
##   Surrogate splits:
##       month < 2.5    to the right, agree=0.684, adj=0.189, (0 split)
##       DC    < 780.3  to the left,  agree=0.681, adj=0.181, (0 split)
##       FFMC  < 92.7   to the left,  agree=0.666, adj=0.142, (0 split)
##       RH    < 55.5   to the left,  agree=0.644, adj=0.087, (0 split)
##       ISI   < 12.6   to the left,  agree=0.635, adj=0.063, (0 split)
##
## Node number 3: 38 observations,    complexity param=0.05661475
##   mean=38.33158, MSE=14974.18
##   left son=6 (27 obs) right son=7 (11 obs)
##   Primary splits:
##       Y    < 5.5    to the left,  improve=0.10522470, (0 missing)
##       DMC  < 186.25 to the left,  improve=0.09248318, (0 missing)
##       wind < 4.7    to the left,  improve=0.08008441, (0 missing)
##       DC   < 693.85 to the left,  improve=0.07722254, (0 missing)
##       X    < 7.5    to the left,  improve=0.06406626, (0 missing)
##   Surrogate splits:
##       X    < 7.5    to the left,  agree=0.842, adj=0.455, (0 split)
##       FFMC < 91.75  to the right, agree=0.763, adj=0.182, (0 split)
##       ISI  < 7.95   to the right, agree=0.763, adj=0.182, (0 split)
##       DMC  < 186.25 to the left,  agree=0.737, adj=0.091, (0 split)
##       DC   < 725    to the left,  agree=0.737, adj=0.091, (0 split)
##
## Node number 4: 199 observations
##   mean=4.447739, MSE=86.86881
##
## Node number 5: 127 observations,    complexity param=0.01231157
##   mean=13.16583, MSE=1451.446
##   left son=10 (120 obs) right son=11 (7 obs)
##   Primary splits:
##       DMC   < 121.15 to the right, improve=0.07598884, (0 missing)
##       Y     < 2.5    to the right, improve=0.07130966, (0 missing)
##       RH    < 46.5   to the right, improve=0.03658832, (0 missing)
##       wind  < 2.45   to the right, improve=0.02494294, (0 missing)
##       month < 9.5    to the left,  improve=0.02450355, (0 missing)
##
```

```
## Node number 6: 27 observations
##   mean=12.99519, MSE=634.2235
##
## Node number 7: 11 observations
##   mean=100.5209, MSE=44729.09
##
## Node number 10: 120 observations
##   mean=10.62933, MSE=1060.177
##
## Node number 11: 7 observations
##   mean=56.64857, MSE=6157.876
```

```
rpart.plot(model1)
```




```
plotcp(model1)
```



Model 2

2. Use the five important attribute variables namely **DMC**, **DC**, **temp**, **month** and **ISI** with respect to the response varialble **AREA**

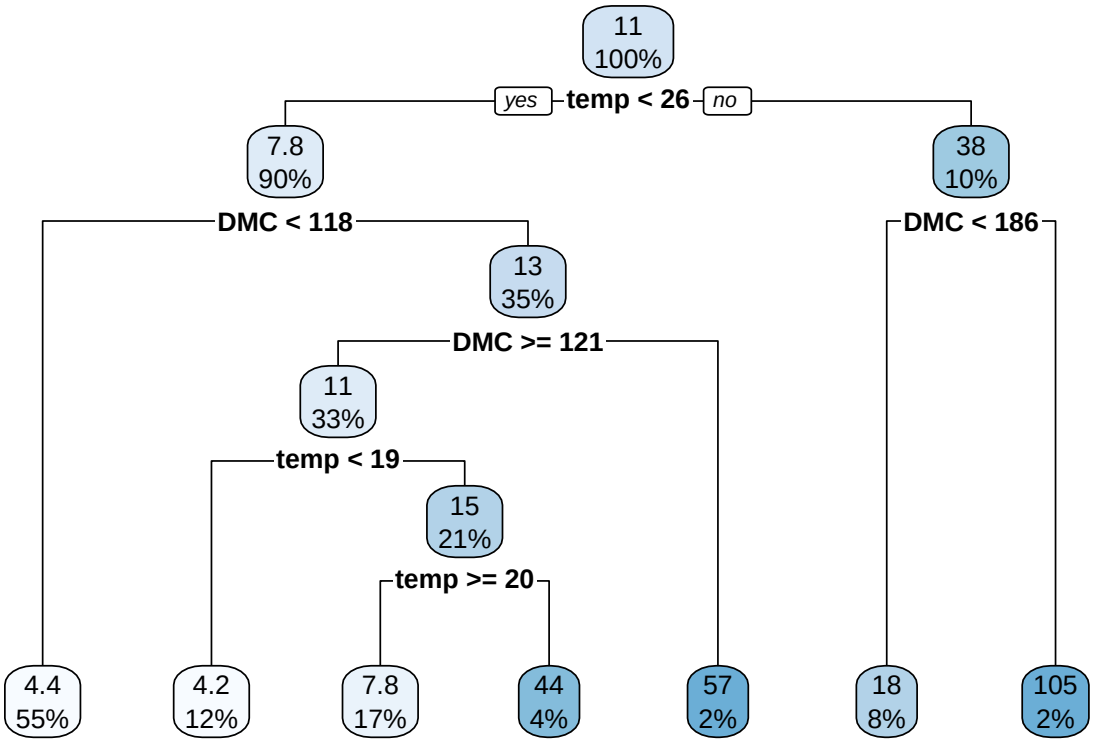
```
model2 <- rpart(area~ DMC + DC + temp + month + ISI, data = trainData)
model2$terms <- eval(model2$call$formula)
summary(model2)
```

```
## Call:
## rpart(formula = area ~ DMC + DC + temp + month + ISI, data = trainData)
##   n= 364
##
##           CP nsplit rel error   xerror      xstd
## 1 0.05212919      0 1.0000000 1.008180 0.6791478
## 2 0.01231157      2 0.8957416 1.158850 0.6813600
## 3 0.01089243      4 0.8711185 1.177379 0.6706523
## 4 0.01000000      6 0.8493336 1.178379 0.6706517
##
## Variable importance
##   DMC  temp   ISI    DC month
##    51    43     5     1     1
##
## Node number 1: 364 observations,    complexity param=0.05212919
##   mean=11.02681, MSE=2220.233
##   left son=2 (326 obs) right son=3 (38 obs)
##   Primary splits:
```

```
##       temp  < 26.15  to the left,  improve=0.039142120, (0 missing)
##       DMC   < 220.05 to the left,  improve=0.026491200, (0 missing)
##       DC    < 674.1  to the left,  improve=0.010879010, (0 missing)
##       ISI   < 13.85  to the left,  improve=0.007559234, (0 missing)
##       month < 3.5    to the right, improve=0.001785507, (0 missing)
##
## Node number 2: 326 observations,    complexity param=0.01231157
##   mean=7.844049, MSE=636.5423
##   left son=4 (199 obs) right son=5 (127 obs)
##   Primary splits:
##       DMC   < 118.45 to the left,  improve=0.028394670, (0 missing)
##       DC    < 674.1  to the left,  improve=0.011426630, (0 missing)
##       month < 11.5   to the left,  improve=0.010650270, (0 missing)
##       ISI   < 8.65   to the right, improve=0.010100260, (0 missing)
##       temp  < 17.65  to the left,  improve=0.008770793, (0 missing)
##   Surrogate splits:
##       month < 2.5    to the right, agree=0.684, adj=0.189, (0 split)
##       DC    < 780.3  to the left,  agree=0.681, adj=0.181, (0 split)
##       ISI   < 12.6   to the left,  agree=0.635, adj=0.063, (0 split)
##       temp  < 21.45  to the left,  agree=0.632, adj=0.055, (0 split)
##
## Node number 3: 38 observations,    complexity param=0.05212919
##   mean=38.33158, MSE=14974.18
##   left son=6 (29 obs) right son=7 (9 obs)
##   Primary splits:
##       DMC   < 186.25 to the left,  improve=0.092483180, (0 missing)
##       DC    < 693.85 to the left,  improve=0.077222540, (0 missing)
##       temp  < 27.65  to the right, improve=0.062398920, (0 missing)
##       ISI   < 13.85  to the left,  improve=0.031885150, (0 missing)
##       month < 4      to the right, improve=0.003375592, (0 missing)
##   Surrogate splits:
##       temp < 26.25  to the right, agree=0.816, adj=0.222, (0 split)
##       ISI  < 8.65   to the right, agree=0.789, adj=0.111, (0 split)
##
## Node number 4: 199 observations
##   mean=4.447739, MSE=86.86881
##
## Node number 5: 127 observations,    complexity param=0.01231157
##   mean=13.16583, MSE=1451.446
##   left son=10 (120 obs) right son=11 (7 obs)
##   Primary splits:
##       DMC   < 121.15 to the right, improve=0.07598884, (0 missing)
##       month < 9.5    to the left,  improve=0.02450355, (0 missing)
##       temp  < 19.7   to the right, improve=0.02224183, (0 missing)
##       ISI   < 10.65  to the right, improve=0.02013627, (0 missing)
##       DC    < 673.5  to the left,  improve=0.01088544, (0 missing)
##
## Node number 6: 29 observations
##   mean=17.60034, MSE=839.9484
##
## Node number 7: 9 observations
##   mean=105.1322, MSE=54670.63
##
## Node number 10: 120 observations,    complexity param=0.01089243
```

```
##   mean=10.62933, MSE=1060.177
##   left son=20 (45 obs) right son=21 (75 obs)
##   Primary splits:
##       temp  < 18.6   to the left,  improve=0.023668120, (0 missing)
##       ISI   < 10.65  to the right, improve=0.015296430, (0 missing)
##       DMC   < 143.9  to the right, improve=0.014795980, (0 missing)
##       month < 9.5    to the left,  improve=0.009563861, (0 missing)
##       DC    < 692.45 to the left,  improve=0.009188408, (0 missing)
##   Surrogate splits:
##       DC    < 729.6  to the right, agree=0.717, adj=0.244, (0 split)
##       DMC   < 234.65 to the right, agree=0.675, adj=0.133, (0 split)
##       ISI   < 7.55   to the left,  agree=0.675, adj=0.133, (0 split)
##       month < 9.5    to the right, agree=0.642, adj=0.044, (0 split)
##
## Node number 11: 7 observations
##   mean=56.64857, MSE=6157.876
##
## Node number 20: 45 observations
##   mean=4.162444, MSE=48.89202
##
## Node number 21: 75 observations,    complexity param=0.01089243
##   mean=14.50947, MSE=1626.8
##   left son=42 (61 obs) right son=43 (14 obs)
##   Primary splits:
##       temp  < 19.7   to the right, improve=0.11961860, (0 missing)
##       ISI   < 10.8   to the right, improve=0.04098495, (0 missing)
##       DMC   < 130.2  to the right, improve=0.03666081, (0 missing)
##       month < 9.5    to the left,  improve=0.02937278, (0 missing)
##       DC    < 692.45 to the left,  improve=0.02538321, (0 missing)
##
## Node number 42: 61 observations
##   mean=7.826557, MSE=592.2001
##
## Node number 43: 14 observations
##   mean=43.62786, MSE=5092.221
```

```
rpart.plot(model2)
```




```
plotcp(model2)
```