DS_Kaggle_BikeShare_Prediction_Models

Divya Sriram

6/10/2017

Decision trees are particulary nice to use when predicting continous outcome variables.

```
# cat("\014")
setwd("~/Desktop/MIDS/DivyaGitHub/TpT-BikeShareKaggle")

#libraries

library(rpart) #for tree
library(Metrics) #for rmsle
```

```
## Warning: package 'Metrics' was built under R version 3.3.2
```

```
library(party)
```

```
## Warning: package 'party' was built under R version 3.3.2

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

```
library(randomForest)
```

```
## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.
```

```r
train_data = read.csv("../TpT-BikeShareKaggle/FeatureEng_Data_Files/train_data.csv", sep = ',')
dev_data = read.csv("../TpT-BikeShareKaggle/FeatureEng_Data_Files/dev_data.csv", sep = ',')
test_data = read.csv("../TpT-BikeShareKaggle/FeatureEng_Data_Files/test_data.csv", sep = ',')

summary(train_data)
```

```
##        X                       datetime        season
##  Min.   :    0   2011-01-01 00:00:00:   1   Min.   :1.000
##  1st Qu.: 2736   2011-01-01 01:00:00:   1   1st Qu.:2.000
##  Median : 5476   2011-01-01 02:00:00:   1   Median :3.000
##  Mean   : 5456   2011-01-01 03:00:00:   1   Mean   :2.505
##  3rd Qu.: 8172   2011-01-01 04:00:00:   1   3rd Qu.:4.000
##  Max.   :10885   2011-01-01 06:00:00:   1   Max.   :4.000
##                  (Other)            :8702
##     holiday          workingday        weather           temp
##  Min.   :0.00000   Min.   :0.0000   Min.   :1.000   Min.   : 0.82
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:13.94
##  Median :0.00000   Median :1.0000   Median :1.000   Median :20.50
##  Mean   :0.03009   Mean   :0.6782   Mean   :1.419   Mean   :20.22
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:26.24
##  Max.   :1.00000   Max.   :1.0000   Max.   :3.000   Max.   :39.36
##
##      atemp          humidity       windspeed          casual
##  Min.   : 0.76   Min.   :  0.0   Min.   : 0.000   Min.   :  0.00
##  1st Qu.:16.66   1st Qu.: 47.0   1st Qu.: 7.002   1st Qu.:  4.00
##  Median :24.24   Median : 62.0   Median :12.998   Median : 17.00
##  Mean   :23.64   Mean   : 61.8   Mean   :12.878   Mean   : 36.11
##  3rd Qu.:31.06   3rd Qu.: 77.0   3rd Qu.:16.998   3rd Qu.: 49.00
##  Max.   :45.45   Max.   :100.0   Max.   :56.997   Max.   :367.00
##
##    registered        count            year          month
##  Min.   :  0.0   Min.   :  1.0   Min.   :2011   Min.   : 1.00
##  1st Qu.: 35.0   1st Qu.: 41.0   1st Qu.:2011   1st Qu.: 4.00
##  Median :118.0   Median :145.0   Median :2012   Median : 7.00
##  Mean   :155.3   Mean   :191.4   Mean   :2012   Mean   : 6.52
##  3rd Qu.:223.0   3rd Qu.:285.0   3rd Qu.:2012   3rd Qu.:10.00
##  Max.   :857.0   Max.   :970.0   Max.   :2012   Max.   :12.00
##
##       day              hour         dayofweek
##  Min.   : 1.000   Min.   : 0.00   Min.   :0.000
##  1st Qu.: 5.000   1st Qu.: 6.00   1st Qu.:1.000
##  Median :10.000   Median :12.00   Median :3.000
##  Mean   : 9.965   Mean   :11.54   Mean   :3.016
##  3rd Qu.:15.000   3rd Qu.:18.00   3rd Qu.:5.000
##  Max.   :19.000   Max.   :23.00   Max.   :6.000
##
```

(1) RPART MODEL

Using rpart (recursive partitioning and regression trees)

(1a) RPART Train Data

Let's try use the rpart model to train with our train_data set.

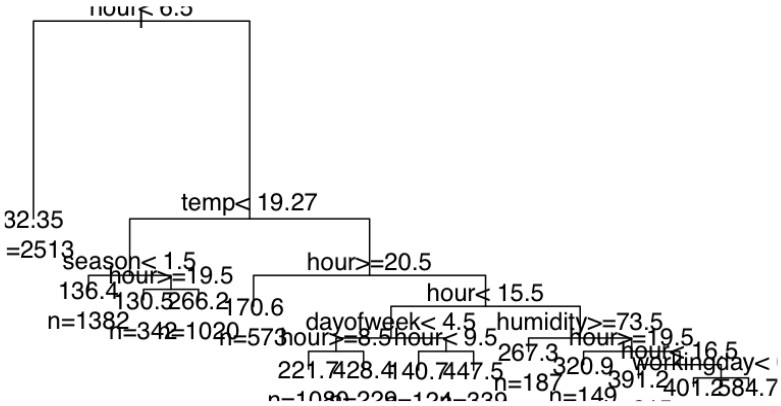
```
# choosing the variables to include in the model
formula_rpart = count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + work

# fitting forumula to the model
fit_rpart = rpart(formula_rpart, data=train_data)

# tells us the importance of each variable in the model
fit_rpart
```

```
## n= 8708
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##   1) root 8708 286020200.0 191.43130
##     2) hour< 6.5 2513   3902504.0  32.34819 *
##     3) hour>=6.5 6195 192721800.0 255.96340
##       6) temp< 19.27 2744  53889700.0 183.88670
##        12) season< 1.5 1382  16000440.0 136.35460 *
##        13) season>=1.5 1362  31598710.0 232.11670
##          26) hour>=19.5 342   1450887.0 130.54970 *
##          27) hour< 19.5 1020  25436870.0 266.17160 *
##       7) temp>=19.27 3451 113242100.0 313.27380
##        14) hour>=20.5 573   3615372.0 170.61260 *
##        15) hour< 20.5 2878  95643040.0 341.67720
##          30) hour< 15.5 1772  42756000.0 285.91200
##            60) dayofweek< 4.5 1309  23166980.0 257.82890
##             120) hour>=8.5 1080   7976795.0 221.65930 *
##             121) hour< 8.5 229   7113827.0 428.41050 *
##            61) dayofweek>=4.5 463  15637970.0 365.30890
##             122) hour< 9.5 124    955036.1 140.65320 *
##             123) hour>=9.5 339   6135463.0 447.48380 *
##          31) hour>=15.5 1106  38547790.0 431.02260
##            62) humidity>=73.5 187   3440661.0 267.32620 *
##            63) humidity< 73.5 919  29076540.0 464.33190
##             126) hour>=19.5 149   1318482.0 320.89260 *
##             127) hour< 19.5 770  24099190.0 492.08830
##               254) hour< 16.5 215   3093781.0 391.18600 *
##               255) hour>=16.5 555  17968460.0 531.17660
##                 510) workingday< 0.5 162   2080535.0 401.21600 *
##                 511) workingday>=0.5 393  12023910.0 584.74810 *
```

```
plot(fit_rpart)
text(fit_rpart, use.n=TRUE)
```



According to this model, the most important factor is hour (biggest split).

(1b) RPART Predict With Dev Data Set

Let's try use the rpart model to predict with our dev_data set. And then we can calculate rmsle to evaluate our model.

```
#dev_data
predict_rpart_dev = predict(fit_rpart, dev_data)

# putting our predictions + hours into dataframe
submit_rpart_dev = data.frame(datetime = dev_data$datetime, count=predict_rpart_dev)

#checking root mean squared log error (like the evaluation in kaggle)
rmsle(dev_data$count, abs(predict_rpart_dev))
```

```
## [1] 0.8758091
```

(1c) RPART Predict With Test Data Set

Let's try use the rpart model to predict with our test_data set. We'll save the predictions for the test_data set along with the datetime column as a dataframe and convert and save that into a csv file to upload to kaggle.

```
#test_data
predict_rpart_test = predict(fit_rpart, test_data)

# putting our predictions + hours into dataframe
submit_rpart_test = data.frame(datetime = test_data$datetime, count=predict_rpart_test)

# writing the dataframe to a csv file --> submit to kaggle
write.csv(submit_rpart_test, file="../TpT-BikeShareKaggle/Submission_Files/rpart/submit_rpart_test_v3.c
```

(2) PARTY MODEL

(2a) PARTY Train Data

Let's try use the party model to train with our train_data set.

Using party (recursive partitioning and regression trees)

```
# choosing the variables to include in the model
formula_ctree = count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + work

#fitting forumula to the model
fit_ctree = ctree(formula_ctree, data=train_data)

#tells us the importance of each variable in the model
fit_ctree
```

```
##
##   Conditional inference tree with 132 terminal nodes
##
## Response:  count
## Inputs:  hour, temp, humidity, season, weather, dayofweek, windspeed, month, workingday
## Number of observations:  8708
##
## 1) hour <= 6; criterion = 1, statistic = 1405.269
##   2) temp <= 17.22; criterion = 1, statistic = 124.829
##     3) season <= 2; criterion = 1, statistic = 65.455
##       4) temp <= 10.66; criterion = 1, statistic = 29.95
##         5)*  weights = 358
##       4) temp > 10.66
##         6)*  weights = 341
##     3) season > 2
##       7)*  weights = 465
##   2) temp > 17.22
##     8) workingday <= 0; criterion = 1, statistic = 58.19
##       9) hour <= 2; criterion = 1, statistic = 230.329
##         10) hour <= 0; criterion = 1, statistic = 55.326
##           11) temp <= 21.32; criterion = 0.985, statistic = 9.845
##             12)*  weights = 16
##           11) temp > 21.32
##             13)*  weights = 44
##         10) hour > 0
##           14) hour <= 1; criterion = 0.999, statistic = 14.765
##             15)*  weights = 62
##           14) hour > 1
##             16) dayofweek <= 5; criterion = 0.997, statistic = 13.136
##               17)*  weights = 34
##             16) dayofweek > 5
##               18)*  weights = 27
##       9) hour > 2
##         19)*  weights = 211
##     8) workingday > 0
##       20) hour <= 5; criterion = 1, statistic = 161.559
##         21) hour <= 0; criterion = 1, statistic = 84.748
```

```
##           22) dayofweek <= 2; criterion = 1, statistic = 29.388
##             23)*  weights = 89
##           22) dayofweek > 2
##             24)*  weights = 55
##         21) hour > 0
##           25) hour <= 4; criterion = 1, statistic = 23.648
##             26) hour <= 1; criterion = 1, statistic = 156.395
##               27) dayofweek <= 3; criterion = 0.998, statistic = 13.967
##                 28) weather <= 1; criterion = 0.981, statistic = 9.402
##                   29)*  weights = 72
##                 28) weather > 1
##                   30) weather <= 2; criterion = 0.961, statistic = 8.112
##                     31) dayofweek <= 2; criterion = 0.965, statistic = 8.301
##                       32)*  weights = 29
##                     31) dayofweek > 2
##                       33)*  weights = 8
##                   30) weather > 2
##                     34)*  weights = 13
##               27) dayofweek > 3
##                 35)*  weights = 22
##             26) hour > 1
##               36) hour <= 2; criterion = 1, statistic = 48.187
##                 37) dayofweek <= 2; criterion = 0.999, statistic = 14.885
##                   38)*  weights = 95
##                 37) dayofweek > 2
##                   39)*  weights = 54
##               36) hour > 2
##                 40) windspeed <= 7.0015; criterion = 0.983, statistic = 9.654
##                   41)*  weights = 98
##                 40) windspeed > 7.0015
##                   42)*  weights = 162
##           25) hour > 4
##             43) month <= 5; criterion = 0.972, statistic = 8.718
##               44)*  weights = 26
##             43) month > 5
##               45)*  weights = 104
##       20) hour > 5
##         46) weather <= 2; criterion = 1, statistic = 26.451
##           47)*  weights = 108
##         46) weather > 2
##           48)*  weights = 20
## 1) hour > 6
##   49) temp <= 18.86; criterion = 1, statistic = 1089.67
##     50) season <= 1; criterion = 1, statistic = 340.82
##       51) temp <= 13.12; criterion = 1, statistic = 131.234
##         52) hour <= 19; criterion = 1, statistic = 96.531
##           53) workingday <= 0; criterion = 1, statistic = 18.76
##             54) temp <= 9.84; criterion = 1, statistic = 32.191
##               55) hour <= 8; criterion = 1, statistic = 24.278
##                 56) hour <= 7; criterion = 0.998, statistic = 13.464
##                   57)*  weights = 20
##                 56) hour > 7
##                   58)*  weights = 15
##               55) hour > 8
```

```
##                 59) temp <= 8.2; criterion = 0.996, statistic = 12.399
##                   60) humidity <= 59; criterion = 0.979, statistic = 9.23
##                     61)*  weights = 54
##                   60) humidity > 59
##                     62)*  weights = 7
##                 59) temp > 8.2
##                   63) weather <= 1; criterion = 0.962, statistic = 8.142
##                     64)*  weights = 22
##                   63) weather > 1
##                     65)*  weights = 13
##             54) temp > 9.84
##               66)*  weights = 88
##           53) workingday > 0
##             67) hour <= 9; criterion = 1, statistic = 16.432
##               68) temp <= 7.38; criterion = 0.997, statistic = 13.032
##                 69)*  weights = 60
##               68) temp > 7.38
##                 70)*  weights = 86
##             67) hour > 9
##               71) hour <= 16; criterion = 1, statistic = 84.196
##                 72) temp <= 9.84; criterion = 1, statistic = 37.268
##                   73) weather <= 1; criterion = 0.956, statistic = 7.865
##                     74)*  weights = 76
##                   73) weather > 1
##                     75)*  weights = 38
##                 72) temp > 9.84
##                   76)*  weights = 95
##               71) hour > 16
##                 77) hour <= 18; criterion = 0.99, statistic = 10.57
##                   78) temp <= 9.84; criterion = 0.958, statistic = 7.97
##                     79)*  weights = 34
##                   78) temp > 9.84
##                     80)*  weights = 14
##                 77) hour > 18
##                   81)*  weights = 31
##         52) hour > 19
##           82) hour <= 20; criterion = 1, statistic = 59.514
##             83)*  weights = 54
##           82) hour > 20
##             84) temp <= 9.02; criterion = 1, statistic = 32.776
##               85) hour <= 21; criterion = 1, statistic = 24.493
##                 86)*  weights = 23
##               85) hour > 21
##                 87) hour <= 22; criterion = 0.982, statistic = 9.547
##                   88)*  weights = 33
##                 87) hour > 22
##                   89)*  weights = 35
##             84) temp > 9.02
##               90) hour <= 22; criterion = 0.993, statistic = 11.41
##                 91)*  weights = 57
##               90) hour > 22
##                 92)*  weights = 31
##       51) temp > 13.12
##         93) weather <= 1; criterion = 1, statistic = 31.471
```

```
##           94) temp <= 16.4; criterion = 0.977, statistic = 9.116
##             95) windspeed <= 19.0012; criterion = 0.982, statistic = 9.575
##               96) humidity <= 43; criterion = 0.996, statistic = 12.259
##                 97)*  weights = 69
##               96) humidity > 43
##                 98)*  weights = 63
##             95) windspeed > 19.0012
##               99)*  weights = 72
##           94) temp > 16.4
##             100)*  weights = 122
##         93) weather > 1
##           101) hour <= 19; criterion = 0.993, statistic = 11.245
##             102) weather <= 2; criterion = 0.993, statistic = 11.182
##               103)*  weights = 100
##             102) weather > 2
##               104) windspeed <= 8.9981; criterion = 0.971, statistic = 8.66
##                 105)*  weights = 12
##               104) windspeed > 8.9981
##                 106)*  weights = 27
##           101) hour > 19
##             107)*  weights = 31
##     50) season > 1
##       108) humidity <= 69; criterion = 1, statistic = 82.452
##         109) temp <= 14.76; criterion = 1, statistic = 40.23
##           110) hour <= 19; criterion = 1, statistic = 35.968
##             111) workingday <= 0; criterion = 1, statistic = 34.349
##               112) temp <= 10.66; criterion = 1, statistic = 36.729
##                 113) hour <= 9; criterion = 0.979, statistic = 9.256
##                   114)*  weights = 15
##                 113) hour > 9
##                   115)*  weights = 12
##               112) temp > 10.66
##                 116) month <= 10; criterion = 1, statistic = 19.479
##                   117)*  weights = 10
##                 116) month > 10
##                   118) temp <= 13.94; criterion = 1, statistic = 19.146
##                     119)*  weights = 46
##                   118) temp > 13.94
##                     120)*  weights = 21
##             111) workingday > 0
##               121)*  weights = 230
##           110) hour > 19
##             122) hour <= 21; criterion = 1, statistic = 58.812
##               123) workingday <= 0; criterion = 1, statistic = 17.114
##                 124)*  weights = 20
##               123) workingday > 0
##                 125)*  weights = 48
##             122) hour > 21
##               126) hour <= 22; criterion = 0.999, statistic = 14.34
##                 127)*  weights = 39
##               126) hour > 22
##                 128)*  weights = 29
##         109) temp > 14.76
##           129) season <= 3; criterion = 1, statistic = 26.895
```

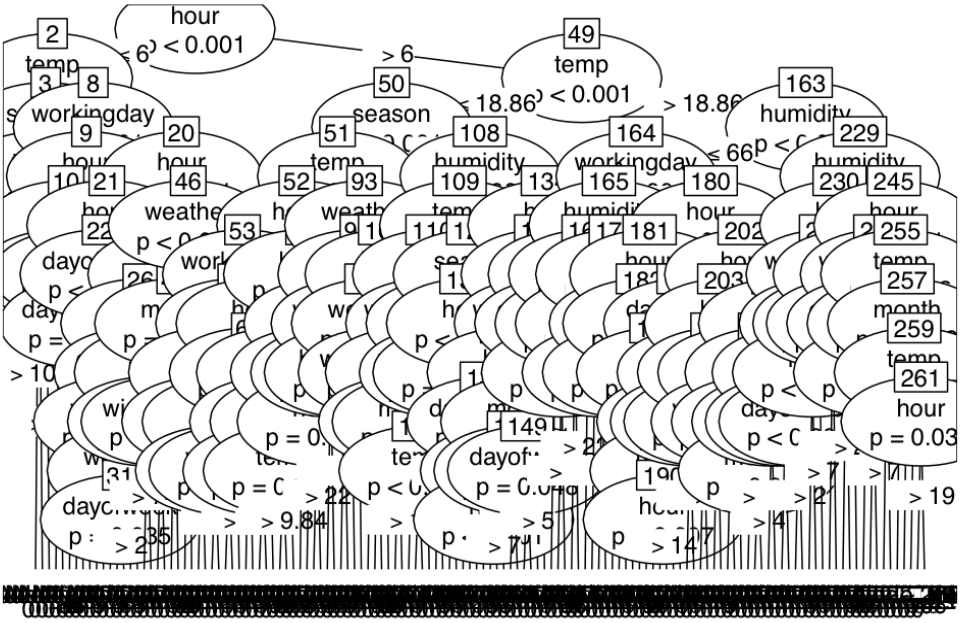
```
##             130) hour <= 20; criterion = 1, statistic = 16.909
##               131)*  weights = 106
##             130) hour > 20
##               132)*  weights = 28
##           129) season > 3
##             133)*  weights = 248
##       108) humidity > 69
##         134) hour <= 19; criterion = 1, statistic = 39.432
##           135) weather <= 2; criterion = 1, statistic = 35.686
##             136) workingday <= 0; criterion = 1, statistic = 29.463
##               137) hour <= 9; criterion = 1, statistic = 26.224
##                 138) dayofweek <= 4; criterion = 1, statistic = 26.032
##                   139)*  weights = 7
##                 138) dayofweek > 4
##                   140) hour <= 8; criterion = 1, statistic = 37.441
##                     141) hour <= 7; criterion = 1, statistic = 24.777
##                       142)*  weights = 27
##                     141) hour > 7
##                       143)*  weights = 19
##                   140) hour > 8
##                     144)*  weights = 13
##               137) hour > 9
##                 145) month <= 11; criterion = 0.989, statistic = 10.527
##                   146) humidity <= 72; criterion = 0.995, statistic = 11.809
##                     147)*  weights = 10
##                   146) humidity > 72
##                     148)*  weights = 23
##                 145) month > 11
##                   149) dayofweek <= 5; criterion = 0.952, statistic = 7.741
##                     150)*  weights = 16
##                   149) dayofweek > 5
##                     151)*  weights = 22
##             136) workingday > 0
##               152)*  weights = 149
##           135) weather > 2
##             153)*  weights = 84
##         134) hour > 19
##           154) hour <= 21; criterion = 1, statistic = 32.844
##             155) dayofweek <= 3; criterion = 0.999, statistic = 15.991
##               156)*  weights = 27
##             155) dayofweek > 3
##               157) month <= 10; criterion = 0.953, statistic = 7.767
##                 158)*  weights = 14
##               157) month > 10
##                 159)*  weights = 16
##           154) hour > 21
##             160) hour <= 22; criterion = 0.958, statistic = 7.959
##               161)*  weights = 40
##             160) hour > 22
##               162)*  weights = 43
##   49) temp > 18.86
##     163) humidity <= 66; criterion = 1, statistic = 328.281
##       164) workingday <= 0; criterion = 1, statistic = 43.599
##         165) humidity <= 47; criterion = 1, statistic = 80.781
```

```
##           166) hour <= 19; criterion = 1, statistic = 47.293
##             167)*  weights = 296
##           166) hour > 19
##             168) temp <= 21.32; criterion = 0.999, statistic = 14.266
##               169)*  weights = 13
##             168) temp > 21.32
##               170) hour <= 20; criterion = 0.965, statistic = 8.315
##                 171)*  weights = 15
##               170) hour > 20
##                 172)*  weights = 9
##         165) humidity > 47
##           173) temp <= 23.78; criterion = 0.988, statistic = 10.355
##             174)*  weights = 84
##           173) temp > 23.78
##             175) humidity <= 55; criterion = 0.995, statistic = 11.869
##               176) hour <= 19; criterion = 0.982, statistic = 9.527
##                 177)*  weights = 89
##               176) hour > 19
##                 178)*  weights = 21
##             175) humidity > 55
##               179)*  weights = 158
##       164) workingday > 0
##         180) hour <= 15; criterion = 1, statistic = 48.639
##           181) hour <= 8; criterion = 1, statistic = 74.412
##             182) hour <= 7; criterion = 1, statistic = 30.687
##               183)*  weights = 30
##             182) hour > 7
##               184)*  weights = 45
##           181) hour > 8
##             185) dayofweek <= 3; criterion = 1, statistic = 69.347
##               186) month <= 8; criterion = 0.998, statistic = 13.722
##                 187)*  weights = 481
##               186) month > 8
##                 188) humidity <= 45; criterion = 0.999, statistic = 14.817
##                   189) month <= 10; criterion = 0.956, statistic = 7.877
##                     190) hour <= 14; criterion = 0.993, statistic = 11.298
##                       191)*  weights = 31
##                     190) hour > 14
##                       192)*  weights = 9
##                   189) month > 10
##                     193)*  weights = 7
##                 188) humidity > 45
##                   194)*  weights = 106
##             185) dayofweek > 3
##               195) humidity <= 42; criterion = 0.988, statistic = 10.307
##                 196) hour <= 11; criterion = 0.953, statistic = 7.76
##                   197)*  weights = 16
##                 196) hour > 11
##                   198)*  weights = 59
##               195) humidity > 42
##                 199) month <= 8; criterion = 0.974, statistic = 8.838
##                   200)*  weights = 55
##                 199) month > 8
##                   201)*  weights = 27
```

```
##         180) hour > 15
##           202) hour <= 19; criterion = 1, statistic = 243.066
##             203) hour <= 16; criterion = 1, statistic = 27.303
##               204) month <= 4; criterion = 0.968, statistic = 8.448
##                 205)*  weights = 27
##               204) month > 4
##                 206)*  weights = 118
##             203) hour > 16
##               207) hour <= 18; criterion = 1, statistic = 65.484
##                 208) windspeed <= 16.9979; criterion = 0.996, statistic = 12.468
##                   209)*  weights = 162
##                 208) windspeed > 16.9979
##                   210)*  weights = 94
##               207) hour > 18
##                 211)*  weights = 100
##           202) hour > 19
##             212) hour <= 21; criterion = 1, statistic = 160.143
##               213) hour <= 20; criterion = 1, statistic = 39.636
##                 214) temp <= 22.14; criterion = 0.997, statistic = 13.013
##                   215)*  weights = 15
##                 214) temp > 22.14
##                   216) month <= 4; criterion = 0.981, statistic = 9.418
##                     217)*  weights = 13
##                   216) month > 4
##                     218)*  weights = 65
##               213) hour > 20
##                 219) temp <= 23.78; criterion = 0.984, statistic = 9.731
##                   220)*  weights = 24
##                 219) temp > 23.78
##                   221)*  weights = 50
##             212) hour > 21
##               222) hour <= 22; criterion = 1, statistic = 29.835
##                 223) dayofweek <= 1; criterion = 0.996, statistic = 12.492
##                   224)*  weights = 29
##                 223) dayofweek > 1
##                   225)*  weights = 39
##               222) hour > 22
##                 226) dayofweek <= 2; criterion = 1, statistic = 16.436
##                   227)*  weights = 26
##                 226) dayofweek > 2
##                   228)*  weights = 25
##     163) humidity > 66
##       229) humidity <= 81; criterion = 1, statistic = 72.575
##         230) hour <= 20; criterion = 1, statistic = 31.303
##           231) workingday <= 0; criterion = 1, statistic = 17.308
##             232) hour <= 8; criterion = 1, statistic = 30.273
##               233) hour <= 7; criterion = 1, statistic = 20.658
##                 234)*  weights = 23
##               233) hour > 7
##                 235)*  weights = 14
##             232) hour > 8
##               236)*  weights = 136
##           231) workingday > 0
##             237) windspeed <= 19.9995; criterion = 0.983, statistic = 9.646
```

```
##               238)*  weights = 280
##             237) windspeed > 19.9995
##               239)*  weights = 37
##           230) hour > 20
##             240) hour <= 22; criterion = 1, statistic = 76.78
##               241) hour <= 21; criterion = 1, statistic = 17.037
##                 242)*  weights = 59
##               241) hour > 21
##                 243)*  weights = 61
##             240) hour > 22
##               244)*  weights = 62
##         229) humidity > 81
##           245) hour <= 9; criterion = 1, statistic = 41.987
##             246) workingday <= 0; criterion = 1, statistic = 31.446
##               247) dayofweek <= 0; criterion = 0.952, statistic = 7.739
##                 248)*  weights = 7
##               247) dayofweek > 0
##                 249) hour <= 7; criterion = 0.998, statistic = 13.272
##                   250)*  weights = 13
##                 249) hour > 7
##                   251)*  weights = 11
##             246) workingday > 0
##               252) weather <= 2; criterion = 1, statistic = 26.739
##                 253)*  weights = 75
##               252) weather > 2
##                 254)*  weights = 23
##           245) hour > 9
##             255) temp <= 26.24; criterion = 0.998, statistic = 13.666
##               256)*  weights = 232
##             255) temp > 26.24
##               257) month <= 6; criterion = 0.981, statistic = 9.41
##                 258)*  weights = 12
##               257) month > 6
##                 259) temp <= 27.88; criterion = 0.953, statistic = 7.752
##                   260)*  weights = 48
##                 259) temp > 27.88
##                   261) hour <= 19; criterion = 0.969, statistic = 8.51
##                     262)*  weights = 7
##                   261) hour > 19
##                     263)*  weights = 13
```

```
plot(fit_ctree)
```



According to this model, the most important factor is temp (biggest split).

(2b) PARTY Predict With Dev Data Set

Let's try use the party model to predict with our dev_data set. And then we can calculate rmsle to evaluate our model.

```
#dev_data
predict_ctree_dev = predict(fit_ctree, dev_data)

# putting our predictions + hours into dataframe
submit_ctree_dev = data.frame(datetime = dev_data$datetime, count=predict_ctree_dev)

#checking root mean squared log error (like the evaluation in kaggle)
rmsle(dev_data$count, abs(predict_ctree_dev))
```

```
## [1] 0.6697017
```

(2c) PARTY Predict With Test Data Set

Let's try use the party model to predict with our test_data set. We'll save the predictions for the test_data set along with the datetime column as a dataframe and convert and save that into a csv file to upload to kaggle.

```
#test_data
predict_ctree_test = predict(fit_ctree, test_data)
```

```
# putting our predictions + hours into dataframe
submit_ctree_test = data.frame(datetime = test_data$datetime, count=predict_ctree_test)

# writing the dataframe to a csv file --> submit to kaggle
write.csv(submit_ctree_test, file="../TpT-BikeShareKaggle/Submission_Files/party/submit_ctree_test_chan
```

(3) RANDOM FORESTS MODEL

Random Forests Train Data

```
# choosing the variables to include in the model
formula_rf = count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + workin

# fitting forumula to the model
rf_model = randomForest(formula_rf, data=train_data, ntree = 250)

# tells us the importance of each variable in the model
print(rf_model)
```

```
##
## Call:
##  randomForest(formula = formula_rf, data = train_data, ntree = 250)
##                Type of random forest: regression
##                      Number of trees: 250
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 4708.352
##                     % Var explained: 85.67
```

Random Forests Predict With Dev Data Set

```
#dev_data
predict_rf_dev = predict(rf_model, dev_data)

# putting our predictions + hours into dataframe
submit_rf_dev = data.frame(datetime = dev_data$datetime, count=predict_rf_dev)

#checking root mean squared log error (like the evaluation in kaggle)
rmsle(dev_data$count, abs(predict_rf_dev))
```

```
## [1] 0.4718153
```

Random Forests Predict With Test Data Set

```r
#test_data
predict_rf_test = predict(rf_model, test_data)

# putting our predictions + hours into dataframe
submit_rf_test = data.frame(datetime = test_data$datetime, count=predict_rf_test)

# writing the dataframe to a csv file --> submit to kaggle
write.csv(submit_rf_test, file="../TpT-BikeShareKaggle/Submission_Files/randomforest/submit_rf_test_250
```

SUBMISSIONS RECORDS:

1. submit_rpart_test_v1.csv: 0.90215

- variables all (count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + workingday)
- unchanged seasons

2. submit_ctree_test_changedseasonsfewvariables.csv: 0.67175

- not all variables
- changed seasons

3. submit_ctree_test_changedseasons.csv: 0.63706

- variables all (count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + workingday)
- changed seasons

4. submit_rf_test_250trees3var.csv: 0.60693

- variables all (count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + workingday)
- 250 trees
- 3 variables
- changed seasons

5. submit_rf_test_250trees3var_removedwindspeed.csv : 0.60693

- variables (no windspeed) (count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + workingday)
- 250 trees
- 3 variables
- changes seasons

6. submit_rf_test_250trees3var_UNchangedseasons.csv : (0.59960)

- variables all (count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + workingday)
- 250 trees
- 3 variables
- UNchanged seasons