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.
```

```
#train = read.csv("../TpT-BikeShareKaggle/Orig_Data_Files/train.csv", sep = ',')
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.: 2753   2011-01-01 02:00:00:   1   1st Qu.:2.000
##  Median : 5450   2011-01-01 03:00:00:   1   Median :3.000
##  Mean   : 5443   2011-01-01 04:00:00:   1   Mean   :2.503
##  3rd Qu.: 8150   2011-01-01 05:00:00:   1   3rd Qu.:3.000
##  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   :0.02905   Mean   :0.6811   Mean   :1.422   Mean   :20.18
##  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.   :41.00
##
##      atemp          humidity        windspeed          casual
##  Min.   : 0.76   Min.   :  0.00   Min.   : 0.000   Min.   :  0.00
##  1st Qu.:16.66   1st Qu.: 47.00   1st Qu.: 7.002   1st Qu.:  4.00
##  Median :24.24   Median : 62.00   Median :12.998   Median : 16.00
##  Mean   :23.59   Mean   : 61.92   Mean   :12.737   Mean   : 36.05
##  3rd Qu.:31.06   3rd Qu.: 77.00   3rd Qu.:16.998   3rd Qu.: 49.00
##  Max.   :45.45   Max.   :100.00   Max.   :56.997   Max.   :367.00
##
##    registered        count            year          month
##  Min.   :  0.0   Min.   :  1.0   Min.   :2011   Min.   : 1.000
##  1st Qu.: 37.0   1st Qu.: 43.0   1st Qu.:2011   1st Qu.: 4.000
##  Median :118.0   Median :145.0   Median :2012   Median : 7.000
##  Mean   :155.5   Mean   :191.6   Mean   :2012   Mean   : 6.523
##  3rd Qu.:222.0   3rd Qu.:283.0   3rd Qu.:2012   3rd Qu.:10.000
##  Max.   :886.0   Max.   :977.0   Max.   :2012   Max.   :12.000
##
##       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   : 9.976   Mean   :11.53   Mean   :3.017
##  3rd Qu.:15.000   3rd Qu.:17.25   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.

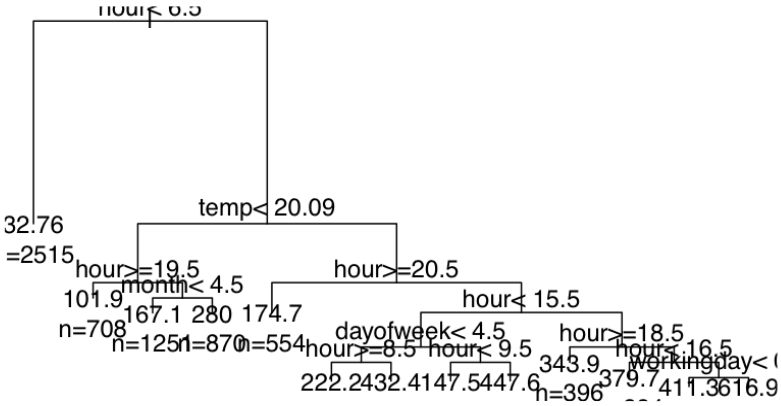
```
# choosing the variables to include in the model
formula_rpart = count ~ hour + temp + humidity + season + weather + dayofweek + windspeed + month + work

# 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 286886900.0 191.58100
##     2) hour< 6.5 2515   3936066.0  32.75865 *
##     3) hour>=6.5 6193 193748100.0 256.07930
##       6) temp< 20.09 2829  56345820.0 185.48070
##        12) hour>=19.5 708   2904246.0 101.86720 *
##        13) hour< 19.5 2121  46839530.0 213.39130
##          26) month< 4.5 1251  18420470.0 167.05200 *
##          27) month>=4.5 870  21870000.0 280.02410 *
##       7) temp>=20.09 3364 111444300.0 315.45010
##        14) hour>=20.5 554   3479361.0 174.69490 *
##        15) hour< 20.5 2810  94825210.0 343.20040
##          30) hour< 15.5 1747  42023440.0 285.85800
##            60) dayofweek< 4.5 1298  23623170.0 258.80820
##             120) hour>=8.5 1072   7952175.0 222.20240 *
##             121) hour< 8.5 226   7420890.0 432.44250 *
##            61) dayofweek>=4.5 449  14704960.0 364.05570
##             122) hour< 9.5 125    932235.2 147.50400 *
##             123) hour>=9.5 324   5649384.0 447.60190 *
##          31) hour>=15.5 1063  37616710.0 437.44030
##            62) hour>=18.5 396   6611300.0 343.85860 *
##            63) hour< 18.5 667  25478480.0 493.00000
##             126) hour< 16.5 234   3842868.0 379.74790 *
##             127) hour>=16.5 433  17012370.0 554.20320
##               254) workingday< 0.5 132   2341758.0 411.27270 *
##               255) workingday>=0.5 301  10791380.0 616.88370 *
```

```
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] 0.8877354
```

(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 156 terminal nodes
##
## Response:  count
## Inputs:  hour, temp, humidity, season, weather, dayofweek, windspeed, month, workingday
## Number of observations:  8708
##
## 1) hour <= 6; criterion = 1, statistic = 1387.954
##   2) temp <= 11.48; criterion = 1, statistic = 117.933
##     3) month <= 4; criterion = 1, statistic = 42.231
##       4) hour <= 5; criterion = 0.993, statistic = 11.275
##         5) workingday <= 0; criterion = 1, statistic = 61.292
##           6) hour <= 2; criterion = 1, statistic = 61.801
##             7)*  weights = 51
##           6) hour > 2
##             8) hour <= 3; criterion = 0.999, statistic = 15.838
##               9)*  weights = 20
##             8) hour > 3
##               10)*  weights = 47
##         5) workingday > 0
##           11) temp <= 10.66; criterion = 0.984, statistic = 9.708
##             12)*  weights = 200
##           11) temp > 10.66
##             13)*  weights = 22
##       4) hour > 5
##         14) workingday <= 0; criterion = 1, statistic = 30.263
##           15) month <= 1; criterion = 0.983, statistic = 10.928
##             16)*  weights = 9
##           15) month > 1
##             17)*  weights = 11
##         14) workingday > 0
##           18)*  weights = 48
##     3) month > 4
##       19)*  weights = 183
##   2) temp > 11.48
##     20) workingday <= 0; criterion = 1, statistic = 52.631
##       21) hour <= 2; criterion = 1, statistic = 309.392
```

```
##         22) hour <= 1; criterion = 1, statistic = 56.718
##           23) temp <= 21.32; criterion = 1, statistic = 43.029
##             24) month <= 4; criterion = 0.999, statistic = 15.082
##               25)*  weights = 41
##             24) month > 4
##               26) dayofweek <= 4; criterion = 0.997, statistic = 12.848
##                 27)*  weights = 7
##               26) dayofweek > 4
##                 28)*  weights = 47
##           23) temp > 21.32
##             29) hour <= 0; criterion = 1, statistic = 20.758
##               30)*  weights = 43
##             29) hour > 0
##               31)*  weights = 44
##         22) hour > 1
##           32) temp <= 19.68; criterion = 1, statistic = 19.919
##             33) dayofweek <= 5; criterion = 0.995, statistic = 11.744
##               34)*  weights = 23
##             33) dayofweek > 5
##               35)*  weights = 19
##           32) temp > 19.68
##             36) dayofweek <= 5; criterion = 0.966, statistic = 8.351
##               37)*  weights = 25
##             36) dayofweek > 5
##               38)*  weights = 22
##       21) hour > 2
##         39) season <= 3; criterion = 0.999, statistic = 15.097
##           40)*  weights = 289
##         39) season > 3
##           41)*  weights = 39
##     20) workingday > 0
##       42) hour <= 5; criterion = 1, statistic = 207.878
##         43) hour <= 0; criterion = 1, statistic = 103.676
##           44) dayofweek <= 3; criterion = 1, statistic = 20.372
##             45)*  weights = 156
##           44) dayofweek > 3
##             46) temp <= 18.04; criterion = 0.974, statistic = 8.885
##               47)*  weights = 13
##             46) temp > 18.04
##               48)*  weights = 25
##         43) hour > 0
##           49) hour <= 4; criterion = 1, statistic = 26.322
##             50) hour <= 1; criterion = 1, statistic = 205.521
##               51) dayofweek <= 3; criterion = 1, statistic = 18.8
##                 52) temp <= 17.22; criterion = 0.987, statistic = 10.187
##                   53) month <= 10; criterion = 0.979, statistic = 9.229
##                     54)*  weights = 17
##                   53) month > 10
##                     55)*  weights = 15
##                 52) temp > 17.22
##                   56)*  weights = 122
##               51) dayofweek > 3
##                 57)*  weights = 41
##             50) hour > 1
```

```
##               58) hour <= 2; criterion = 1, statistic = 49.329
##                 59) dayofweek <= 3; criterion = 1, statistic = 17.918
##                   60) temp <= 16.4; criterion = 0.954, statistic = 7.794
##                     61)*  weights = 30
##                   60) temp > 16.4
##                     62)*  weights = 108
##                 59) dayofweek > 3
##                   63)*  weights = 38
##               58) hour > 2
##                 64) month <= 4; criterion = 1, statistic = 21.685
##                   65)*  weights = 70
##                 64) month > 4
##                   66) weather <= 2; criterion = 0.997, statistic = 12.828
##                     67)*  weights = 278
##                   66) weather > 2
##                     68)*  weights = 32
##           49) hour > 4
##             69) month <= 5; criterion = 1, statistic = 24.617
##               70)*  weights = 64
##             69) month > 5
##               71)*  weights = 123
##       42) hour > 5
##         72) weather <= 2; criterion = 1, statistic = 32.727
##           73) temp <= 18.04; criterion = 1, statistic = 26.145
##             74) month <= 9; criterion = 0.999, statistic = 15.395
##               75)*  weights = 37
##             74) month > 9
##               76)*  weights = 31
##           73) temp > 18.04
##             77)*  weights = 97
##         72) weather > 2
##           78)*  weights = 28
## 1) hour > 6
##   79) temp <= 19.68; criterion = 1, statistic = 1049.957
##     80) month <= 4; criterion = 1, statistic = 343.026
##       81) temp <= 13.12; criterion = 1, statistic = 183.674
##         82) hour <= 19; criterion = 1, statistic = 101.915
##           83) temp <= 9.02; criterion = 1, statistic = 26.516
##             84) workingday <= 0; criterion = 1, statistic = 21.514
##               85) humidity <= 47; criterion = 1, statistic = 17.79
##                 86)*  weights = 55
##               85) humidity > 47
##                 87) hour <= 8; criterion = 0.999, statistic = 14.972
##                   88)*  weights = 23
##                 87) hour > 8
##                   89) windspeed <= 15.0013; criterion = 0.992, statistic = 10.956
##                     90)*  weights = 21
##                   89) windspeed > 15.0013
##                     91)*  weights = 7
##             84) workingday > 0
##               92) hour <= 9; criterion = 0.982, statistic = 9.559
##                 93)*  weights = 94
##               92) hour > 9
##                 94) hour <= 16; criterion = 1, statistic = 55.336
```

```
##                 95)*  weights = 92
##               94) hour > 16
##                 96) humidity <= 41; criterion = 0.966, statistic = 8.35
##                   97)*  weights = 20
##                 96) humidity > 41
##                   98)*  weights = 17
##           83) temp > 9.02
##             99)*  weights = 352
##         82) hour > 19
##           100) hour <= 21; criterion = 1, statistic = 62.79
##             101) workingday <= 0; criterion = 0.999, statistic = 15.918
##               102)*  weights = 43
##             101) workingday > 0
##               103) temp <= 9.84; criterion = 0.994, statistic = 11.619
##                 104)*  weights = 47
##               103) temp > 9.84
##                 105)*  weights = 30
##           100) hour > 21
##             106) hour <= 22; criterion = 1, statistic = 21.123
##               107) temp <= 9.02; criterion = 0.971, statistic = 8.64
##                 108)*  weights = 23
##               107) temp > 9.02
##                 109)*  weights = 33
##             106) hour > 22
##               110) temp <= 9.02; criterion = 0.996, statistic = 12.135
##                 111)*  weights = 31
##               110) temp > 9.02
##                 112)*  weights = 33
##       81) temp > 13.12
##         113) weather <= 1; criterion = 1, statistic = 41.365
##           114) hour <= 19; criterion = 1, statistic = 28.417
##             115) temp <= 16.4; criterion = 1, statistic = 18.258
##               116)*  weights = 188
##             115) temp > 16.4
##               117)*  weights = 148
##           114) hour > 19
##             118) hour <= 21; criterion = 1, statistic = 42.851
##               119) workingday <= 0; criterion = 0.984, statistic = 9.701
##                 120)*  weights = 15
##               119) workingday > 0
##                 121)*  weights = 44
##             118) hour > 21
##               122) hour <= 22; criterion = 0.998, statistic = 13.491
##                 123) workingday <= 0; criterion = 0.964, statistic = 8.27
##                   124)*  weights = 7
##                 123) workingday > 0
##                   125)*  weights = 20
##               122) hour > 22
##                 126)*  weights = 33
##         113) weather > 1
##           127) windspeed <= 12.998; criterion = 0.988, statistic = 10.239
##             128) temp <= 16.4; criterion = 0.999, statistic = 15.669
##               129)*  weights = 88
##             128) temp > 16.4
```

```
##               130) weather <= 2; criterion = 0.967, statistic = 8.405
##                 131)*  weights = 46
##               130) weather > 2
##                 132)*  weights = 11
##           127) windspeed > 12.998
##             133)*  weights = 146
##      80) month > 4
##        134) humidity <= 62; criterion = 1, statistic = 73.99
##          135) temp <= 14.76; criterion = 1, statistic = 38.972
##            136) hour <= 19; criterion = 1, statistic = 22.614
##              137) workingday <= 0; criterion = 1, statistic = 20.126
##                138) temp <= 13.12; criterion = 1, statistic = 42.149
##                  139) temp <= 10.66; criterion = 0.998, statistic = 13.563
##                    140)*  weights = 17
##                  139) temp > 10.66
##                    141) month <= 11; criterion = 0.953, statistic = 7.758
##                      142)*  weights = 12
##                    141) month > 11
##                      143)*  weights = 20
##                138) temp > 13.12
##                  144)*  weights = 36
##              137) workingday > 0
##                145) month <= 11; criterion = 0.972, statistic = 8.723
##                  146)*  weights = 84
##                145) month > 11
##                  147)*  weights = 100
##            136) hour > 19
##              148) hour <= 21; criterion = 1, statistic = 36.226
##                149) workingday <= 0; criterion = 1, statistic = 16.391
##                  150)*  weights = 15
##                149) workingday > 0
##                  151) hour <= 20; criterion = 0.972, statistic = 8.73
##                    152)*  weights = 16
##                  151) hour > 20
##                    153)*  weights = 16
##              148) hour > 21
##                154) hour <= 22; criterion = 0.966, statistic = 8.385
##                  155)*  weights = 26
##                154) hour > 22
##                  156)*  weights = 21
##          135) temp > 14.76
##            157)*  weights = 248
##        134) humidity > 62
##          158) hour <= 19; criterion = 1, statistic = 63.629
##            159) weather <= 2; criterion = 1, statistic = 31.569
##              160) workingday <= 0; criterion = 1, statistic = 26.034
##                161) hour <= 9; criterion = 1, statistic = 21.911
##                  162) dayofweek <= 5; criterion = 1, statistic = 22.868
##                    163)*  weights = 30
##                  162) dayofweek > 5
##                    164) hour <= 8; criterion = 0.999, statistic = 16.157
##                      165)*  weights = 18
##                    164) hour > 8
##                      166)*  weights = 10
```

```
##               161) hour > 9
##                 167) month <= 11; criterion = 0.968, statistic = 8.445
##                   168) humidity <= 71; criterion = 0.995, statistic = 11.898
##                     169)*  weights = 19
##                   168) humidity > 71
##                     170)*  weights = 13
##                 167) month > 11
##                   171) dayofweek <= 5; criterion = 0.953, statistic = 7.772
##                     172)*  weights = 14
##                   171) dayofweek > 5
##                     173)*  weights = 25
##             160) workingday > 0
##               174)*  weights = 190
##           159) weather > 2
##             175) month <= 10; criterion = 0.952, statistic = 7.742
##               176)*  weights = 22
##             175) month > 10
##               177) workingday <= 0; criterion = 0.955, statistic = 7.828
##                 178)*  weights = 9
##               177) workingday > 0
##                 179)*  weights = 39
##         158) hour > 19
##           180) hour <= 21; criterion = 1, statistic = 46.227
##             181) workingday <= 0; criterion = 0.985, statistic = 9.93
##               182)*  weights = 27
##             181) workingday > 0
##               183) windspeed <= 12.998; criterion = 0.987, statistic = 10.203
##                 184) hour <= 20; criterion = 0.991, statistic = 10.883
##                   185)*  weights = 10
##                 184) hour > 20
##                   186)*  weights = 19
##               183) windspeed > 12.998
##                 187)*  weights = 8
##           180) hour > 21
##             188) hour <= 22; criterion = 0.997, statistic = 12.82
##               189)*  weights = 41
##             188) hour > 22
##               190)*  weights = 57
##   79) temp > 19.68
##     191) humidity <= 75; criterion = 1, statistic = 282.936
##       192) humidity <= 42; criterion = 1, statistic = 62.129
##         193) workingday <= 0; criterion = 1, statistic = 65.293
##           194) hour <= 19; criterion = 1, statistic = 26.001
##             195) season <= 1; criterion = 0.996, statistic = 12.464
##               196)*  weights = 102
##             195) season > 1
##               197) humidity <= 31; criterion = 0.994, statistic = 11.628
##                 198)*  weights = 29
##               197) humidity > 31
##                 199) temp <= 31.98; criterion = 0.99, statistic = 10.639
##                   200) season <= 2; criterion = 0.998, statistic = 14.162
##                     201)*  weights = 28
##                   200) season > 2
##                     202)*  weights = 23
```

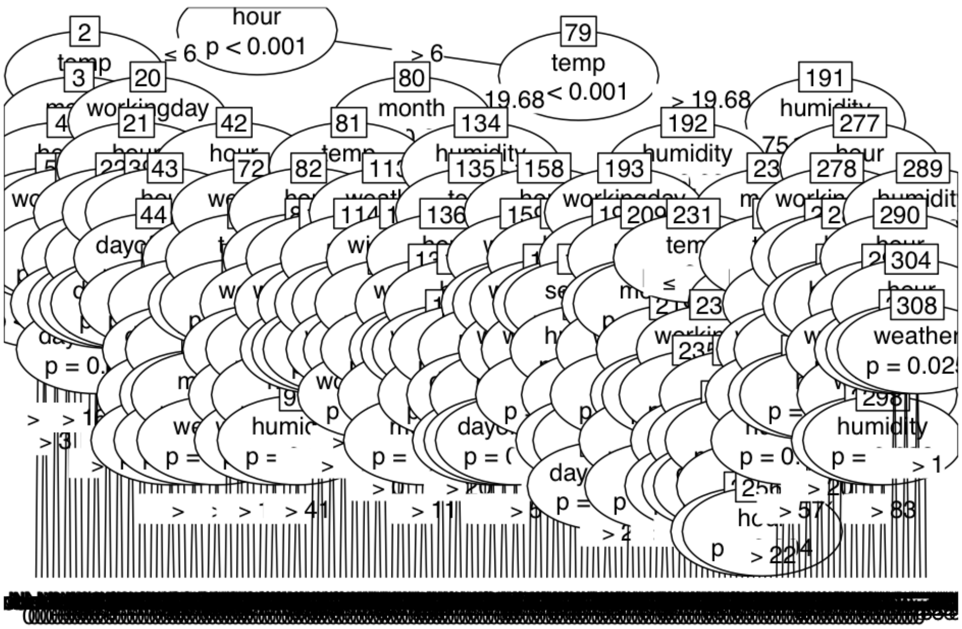
```
##                 199) temp > 31.98
##                   203) month <= 7; criterion = 0.965, statistic = 8.312
##                     204) dayofweek <= 2; criterion = 0.959, statistic = 8.024
##                       205)*  weights = 8
##                     204) dayofweek > 2
##                       206)*  weights = 22
##                   203) month > 7
##                     207)*  weights = 11
##             194) hour > 19
##               208)*  weights = 17
##           193) workingday > 0
##             209) hour <= 15; criterion = 1, statistic = 77.525
##               210) dayofweek <= 3; criterion = 1, statistic = 40.777
##                 211) weather <= 1; criterion = 0.956, statistic = 7.879
##                   212) month <= 8; criterion = 0.988, statistic = 10.268
##                     213)*  weights = 188
##                   212) month > 8
##                     214)*  weights = 23
##                 211) weather > 1
##                   215)*  weights = 21
##               210) dayofweek > 3
##                 216)*  weights = 70
##             209) hour > 15
##               217) month <= 2; criterion = 1, statistic = 26.923
##                 218)*  weights = 12
##               217) month > 2
##                 219) hour <= 19; criterion = 1, statistic = 23.764
##                   220) month <= 8; criterion = 0.999, statistic = 14.487
##                     221) hour <= 16; criterion = 0.998, statistic = 13.57
##                       222)*  weights = 60
##                     221) hour > 16
##                       223) hour <= 18; criterion = 1, statistic = 23.108
##                         224)*  weights = 93
##                       223) hour > 18
##                         225)*  weights = 30
##                   220) month > 8
##                     226)*  weights = 19
##                 219) hour > 19
##                   227) hour <= 21; criterion = 1, statistic = 23.779
##                     228)*  weights = 35
##                   227) hour > 21
##                     229)*  weights = 11
##         192) humidity > 42
##           230) month <= 8; criterion = 0.999, statistic = 16.069
##             231) temp <= 22.14; criterion = 1, statistic = 18.976
##               232)*  weights = 140
##             231) temp > 22.14
##               233) humidity <= 55; criterion = 0.994, statistic = 11.441
##                 234) workingday <= 0; criterion = 0.97, statistic = 8.566
##                   235) hour <= 19; criterion = 0.999, statistic = 15.864
##                     236)*  weights = 116
##                   235) hour > 19
##                     237) hour <= 21; criterion = 0.997, statistic = 12.832
##                       238) temp <= 27.88; criterion = 0.975, statistic = 8.923
```

```
##                     239)*  weights = 12
##                   238) temp > 27.88
##                     240)*  weights = 10
##                 237) hour > 21
##                   241)*  weights = 15
##             234) workingday > 0
##               242) hour <= 15; criterion = 0.979, statistic = 9.26
##                 243) hour <= 8; criterion = 1, statistic = 32.282
##                   244) hour <= 7; criterion = 0.999, statistic = 15.971
##                     245)*  weights = 13
##                   244) hour > 7
##                     246)*  weights = 14
##                 243) hour > 8
##                   247) dayofweek <= 1; criterion = 1, statistic = 16.273
##                     248)*  weights = 80
##                   247) dayofweek > 1
##                     249)*  weights = 110
##               242) hour > 15
##                 250) hour <= 19; criterion = 1, statistic = 74.689
##                   251)*  weights = 117
##                 250) hour > 19
##                   252) hour <= 21; criterion = 1, statistic = 49.922
##                     253) hour <= 20; criterion = 0.971, statistic = 8.628
##                       254)*  weights = 21
##                     253) hour > 20
##                       255)*  weights = 19
##                   252) hour > 21
##                     256) hour <= 22; criterion = 0.996, statistic = 12.19
##                       257)*  weights = 18
##                     256) hour > 22
##                       258)*  weights = 20
##           233) humidity > 55
##             259)*  weights = 626
##         230) month > 8
##           260) temp <= 25.42; criterion = 0.999, statistic = 14.441
##             261)*  weights = 342
##           260) temp > 25.42
##             262) hour <= 15; criterion = 0.972, statistic = 8.695
##               263) workingday <= 0; criterion = 1, statistic = 38.545
##                 264) hour <= 9; criterion = 0.997, statistic = 12.842
##                   265)*  weights = 7
##                 264) hour > 9
##                   266) humidity <= 57; criterion = 0.976, statistic = 8.992
##                     267)*  weights = 18
##                   266) humidity > 57
##                     268)*  weights = 33
##               263) workingday > 0
##                 269)*  weights = 80
##             262) hour > 15
##               270) hour <= 19; criterion = 1, statistic = 25.283
##                 271) workingday <= 0; criterion = 0.998, statistic = 13.253
##                   272)*  weights = 28
##                 271) workingday > 0
##                   273)*  weights = 51
```

```
##               270) hour > 19
##                 274) hour <= 20; criterion = 0.989, statistic = 10.483
##                   275)*  weights = 10
##                 274) hour > 20
##                   276)*  weights = 18
##     191) humidity > 75
##       277) hour <= 8; criterion = 1, statistic = 63.608
##         278) workingday <= 0; criterion = 1, statistic = 56.917
##           279) hour <= 7; criterion = 0.984, statistic = 9.748
##             280)*  weights = 19
##           279) hour > 7
##             281)*  weights = 18
##         278) workingday > 0
##           282) hour <= 7; criterion = 1, statistic = 27.564
##             283) weather <= 2; criterion = 1, statistic = 18.341
##               284)*  weights = 46
##             283) weather > 2
##               285)*  weights = 13
##           282) hour > 7
##             286) humidity <= 89; criterion = 0.97, statistic = 8.583
##               287)*  weights = 38
##             286) humidity > 89
##               288)*  weights = 8
##       277) hour > 8
##         289) humidity <= 89; criterion = 1, statistic = 40.984
##           290) hour <= 20; criterion = 0.999, statistic = 14.9
##             291) hour <= 15; criterion = 0.996, statistic = 12.579
##               292) workingday <= 0; criterion = 1, statistic = 20.86
##                 293)*  weights = 29
##               292) workingday > 0
##                 294) weather <= 2; criterion = 0.999, statistic = 16.21
##                   295) hour <= 9; criterion = 0.999, statistic = 14.978
##                     296)*  weights = 27
##                   295) hour > 9
##                     297)*  weights = 49
##                 294) weather > 2
##                   298) humidity <= 83; criterion = 0.974, statistic = 8.826
##                     299)*  weights = 13
##                   298) humidity > 83
##                     300)*  weights = 21
##             291) hour > 15
##               301) workingday <= 0; criterion = 0.98, statistic = 9.327
##                 302)*  weights = 48
##               301) workingday > 0
##                 303)*  weights = 76
##           290) hour > 20
##             304) hour <= 22; criterion = 1, statistic = 37.947
##               305) humidity <= 79; criterion = 0.998, statistic = 13.421
##                 306)*  weights = 41
##               305) humidity > 79
##                 307)*  weights = 61
##             304) hour > 22
##               308) weather <= 1; criterion = 0.975, statistic = 8.941
##                 309)*  weights = 25
```

```
##               308) weather > 1
##                 310)*  weights = 33
##         289) humidity > 89
##           311)*  weights = 79
```

```r
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.

```r
#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] 0.5832768
```

(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_rpart, data=train_data, ntree = 250)

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

```
##
## Call:
##  randomForest(formula = formula_rpart, 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: 4696.426
##                     % Var explained: 85.74
```

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] 0.4782102
```

Random Forests Predict With Test Data Set

```
#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

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