# DecisionTree_gradu_ontime_overtime_DWF_CRDHRS3

jsparks3

2020-08-27

```
#################################
### Decision Tree ###############
#################################

#need packages party, rpart, rpart.plot
library(party)
```

```
## 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(rpart)
library(rpart.plot)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#import data set
library(readr)
Rtreedataset_M <- read_csv("C:/Users/jsparks3/Downloads/Rtreedataset_M.csv",
                           col_types = cols(X1 = col_skip(), Stu_ProgramCIPCodeChange = col_factor(levels = c()),
                                            Stu_CollegeChange = col_factor(levels = c()),
                                            Stu_DepartmentChange = col_factor(levels = c()),
                                            Stu_Department3 = col_factor(levels = c()),
                                            Stu_Gender = col_factor(levels = c()),
                                            Deg_CollegeCode = col_factor(levels = c()),
                                            Deg_Depar.Code = col_factor(levels = c()),
                                            Gradu_Code = col_factor(levels = c()))))
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
# characters are not supported, double or factor
Rtreedataset_M$Stu_Ethnicity <- as.factor(Rtreedataset_M$Stu_Ethnicity)


#Data partition
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

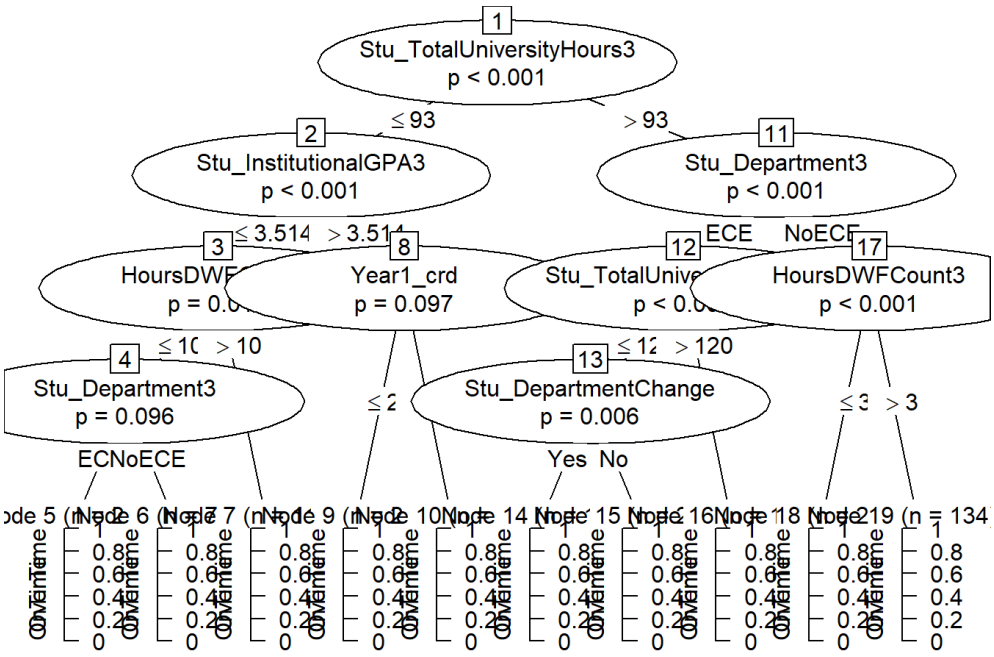
```
# Performs stratified random split of the data set
TrainingIndex <- createDataPartition(Rtreedataset_M$Gradu_Code, p=0.8, list = FALSE)
TrainingSet <- Rtreedataset_M[TrainingIndex,] # Training Set 677
```

```
## Warning: The `i` argument of ``[`()` can't be a matrix as of tibble 3.0.0.
## Convert to a vector.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
TestingSet <- Rtreedataset_M[-TrainingIndex,] # Test Set 166

#model 01 include Year 4 factors
treeGradu <- ctree(Gradu_Code ~ ., data=TrainingSet, controls = ctree_control(mincriterion = 0.9,minsplit = 10))
treeGradu
```

```
##
##   Conditional inference tree with 10 terminal nodes
##
## Response:  Gradu_Code
## Inputs:  Stu_ProgramCIPCodeChange, Stu_CollegeChange, Stu_TotalUniversityHoursBegin, Stu_DepartmentChange, Stu_TotalUnive
rsityHours1, Year1_crd, Stu_TotalUniversityHours2, Year2_crd, Stu_TotalUniversityHours3, Year3_crd, Stu_TotalInstHours1, Stu
_TotalInstHours2, Stu_TotalInstHours3, Stu_Department3, Stu_Gender, Stu_Ethnicity, Stu_InstitutionalGPA3, HoursDWFCount3, De
g_CollegeCode, Deg_Depar.Code
## Number of observations:  667
##
## 1) Stu_TotalUniversityHours3 <= 93; criterion = 1, statistic = 152.748
##   2) Stu_InstitutionalGPA3 <= 3.5144; criterion = 1, statistic = 24.366
##     3) HoursDWFCount3 <= 10; criterion = 0.982, statistic = 11.054
##       4) Stu_Department3 == {ECE}; criterion = 0.904, statistic = 7.87
##         5)*  weights = 27
##       4) Stu_Department3 == {NoECE}
##         6)*  weights = 79
##     3) HoursDWFCount3 > 10
##       7)*  weights = 116
##   2) Stu_InstitutionalGPA3 > 3.5144
##     8) Year1_crd <= 20; criterion = 0.903, statistic = 8.353
##       9)*  weights = 27
##     8) Year1_crd > 20
##       10)*  weights = 8
## 1) Stu_TotalUniversityHours3 > 93
##   11) Stu_Department3 == {ECE}; criterion = 1, statistic = 27.896
##     12) Stu_TotalUniversityHours3 <= 120; criterion = 1, statistic = 28.463
##       13) Stu_DepartmentChange == {Yes}; criterion = 0.994, statistic = 13.183
##         14)*  weights = 16
##       13) Stu_DepartmentChange == {No}
##         15)*  weights = 28
##     12) Stu_TotalUniversityHours3 > 120
##       16)*  weights = 17
##   11) Stu_Department3 == {NoECE}
##     17) HoursDWFCount3 <= 3; criterion = 1, statistic = 31.24
##       18)*  weights = 215
##     17) HoursDWFCount3 > 3
##       19)*  weights = 134
```

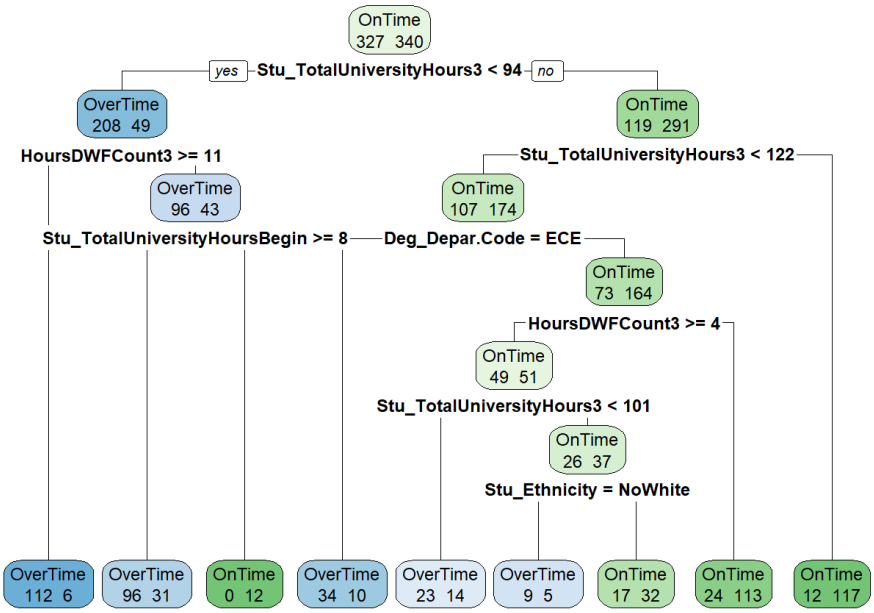
```
plot(treeGradu)
```



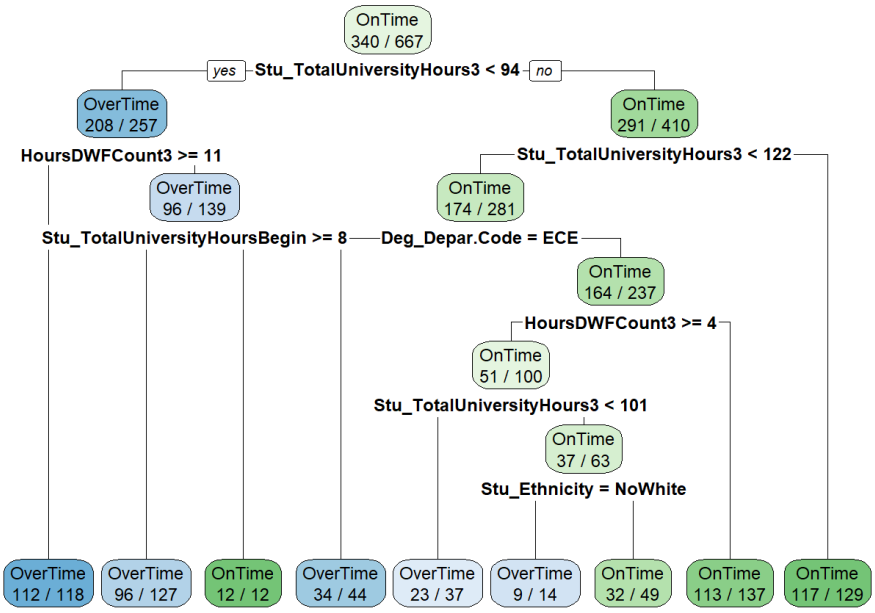

```
#confusion matrix
library(caret)
library(e1071)
confusionMatrix(predict(treeGradu), TrainingSet$Gradu_Code)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction OverTime OnTime
##   OverTime      243     50
##   OnTime         84    290
##
##                Accuracy : 0.7991
##                  95% CI : (0.7666, 0.8289)
##     No Information Rate : 0.5097
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5972
##
##  Mcnemar's Test P-Value : 0.004361
##
##             Sensitivity : 0.7431
##             Specificity : 0.8529
##          Pos Pred Value : 0.8294
##          Neg Pred Value : 0.7754
##              Prevalence : 0.4903
##          Detection Rate : 0.3643
##    Detection Prevalence : 0.4393
##       Balanced Accuracy : 0.7980
##
##        'Positive' Class : OverTime
##
```

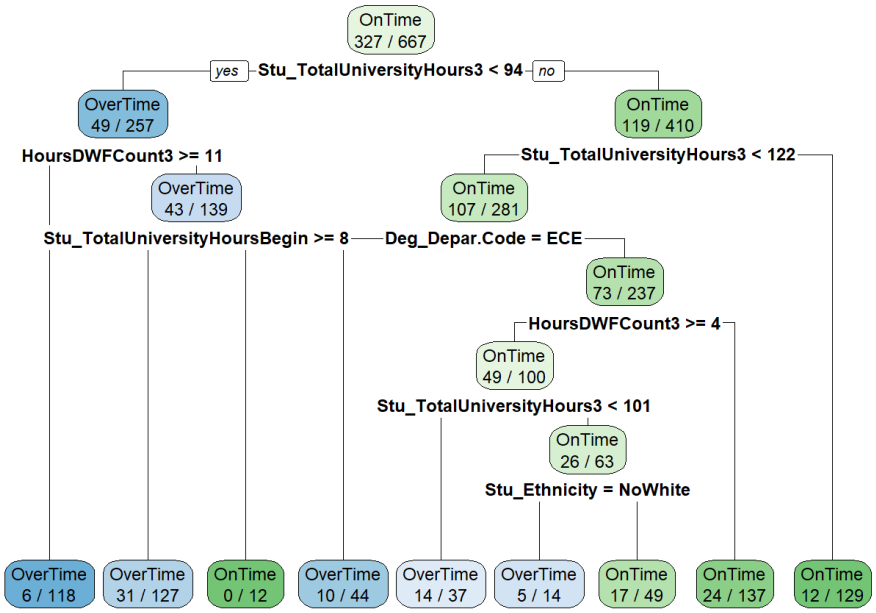
```
#export the results


# method is "class" since y is factor

tree1 <- rpart(Gradu_Code ~ Stu_ProgramCIPCodeChange+Stu_CollegeChange+Stu_TotalUniversityHoursBegin+Stu_DepartmentChange
               +Stu_TotalUniversityHours3+Stu_TotalInstHours3+Stu_Ethnicity+HoursDWFCount3+Stu_Gender+Deg_Depar.Code
               , data=TrainingSet, control = rpart.control(minsplit=20,cp=0.01))
rpart.plot(tree1, extra = 1)
```



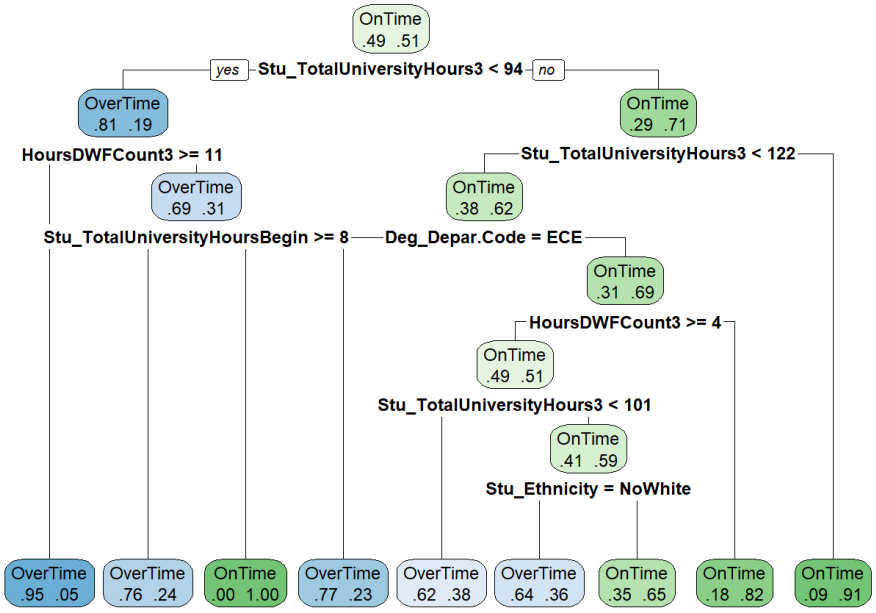

```
rpart.plot(tree1, extra = 2)
```




```
rpart.plot(tree1, extra = 3)
```




```
rpart.plot(tree1, extra = 4)
```




```
tree1$variable.importance
```

```
##     Stu_TotalUniversityHours3 Stu_TotalUniversityHoursBegin
##                     105.5366317                    54.6258818
##                HoursDWFCount3           Stu_TotalInstHours3
##                     43.2317985                    36.3393262
##                Deg_Depar.Code             Stu_CollegeChange
##                     16.0283742                     4.2102724
##                 Stu_Ethnicity      Stu_ProgramCIPCodeChange
##                      3.2321063                     0.4337116
```

```
varImp(tree1)
```

```
##                                 Overall
## Deg_Depar.Code                56.2154322
## HoursDWFCount3                95.0188414
## Stu_DepartmentChange           3.9955770
## Stu_Ethnicity                  1.9070295
## Stu_Gender                     0.5194525
## Stu_ProgramCIPCodeChange       2.0409768
## Stu_TotalInstHours3           51.3658828
## Stu_TotalUniversityHours3    116.4972221
## Stu_TotalUniversityHoursBegin 65.0142052
## Stu_CollegeChange              0.0000000
```