R Decision Tree for ELPAC Data

Oscar Gil

11/13/2022

```
# R Libraries
library(caret)
library(AppliedPredictiveModeling)
#library(Hmisc)
library(dplyr)
library(tidyverse)
library(ggplot2)
library(corrplot)
library(MASS)
library(ISLR)
library(rpart)
library(partykit)
library(randomForestSRC)
library(earth)
library(MARSS)
library(e1071)
library(summarytools)
library(grid)
library(MLeval)
library(pROC)
```

Load the ELPAC data set from GitHub

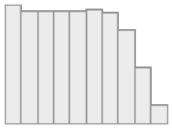
```
df <- read.csv(
  url("https://raw.githubusercontent.com/OscarG-DataSci/ADS-599B/main/Data%20Folder/elpac.csv")
                       , header = TRUE)
```

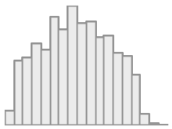
Data Summary

```
# use the view function to view in R Studio
#view(
dfSummary(df,
          plain.ascii  = FALSE,
          style        = "grid",
          graph.magnif = 0.75,
          valid.col    = FALSE,
          tmp.img.dir  = "NA")
```

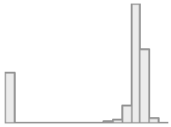
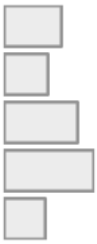
Data Frame Summary

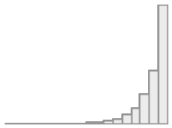
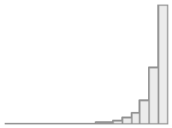
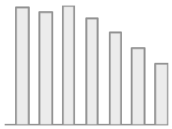
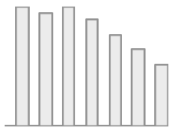
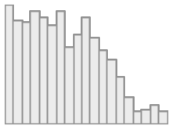
df Dimensions: 11628 x 24
Duplicates: 0

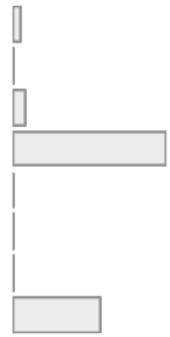
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	AcademicYear [character]	1. 2017-2018 2. 2018-2019 3. 2019-2020 4. 2020-2021 5. 2021-2022	2428 (20.9%) 2273 (19.5%) 2515 (21.6%) 2428 (20.9%) 1984 (17.1%)		0 (0.0%)
2	Stu_deID [integer]	Mean (sd) : 2138.8 (1271.3) min < med < max: 0 < 2120.5 < 4752 IQR (CV) : 2140.8 (0.6)	4519 distinct values		0 (0.0%)
3	School_deID [integer]	Mean (sd) : 4.2 (2.8) min < med < max: 0 < 4 < 9 IQR (CV) : 5 (0.7)	0 : 1359 (11.7%) 1 : 1205 (10.4%) 2 : 1472 (12.7%) 3 : 1306 (11.2%) 4 : 1132 (9.7%) 5 : 974 (8.4%) 6 : 1154 (9.9%) 7 : 1245 (10.7%) 8 : 772 (6.6%) 9 : 1009 (8.7%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
4	GradeLevel [integer]	Mean (sd) : 2.6 (1.9) min < med < max: 0 < 2 < 6 IQR (CV) : 3 (0.7)	0 : 2008 (17.3%) 1 : 1920 (16.5%) 2 : 2010 (17.3%) 3 : 1800 (15.5%) 4 : 1558 (13.4%) 5 : 1294 (11.1%) 6 : 1038 (8.9%)		0 (0.0%)
5	DOB [character]	1. 2010-08-21 2. 2012-12-12 3. 2011-01-13 4. 2011-04-01 5. 2011-05-26 6. 2012-04-11 7. 2012-08-15 8. 2010-10-19 9. 2010-10-24 10. 2010-11-15 [2585 others]	27 (0.2%) 24 (0.2%) 23 (0.2%) 22 (0.2%) 20 (0.2%) 20 (0.2%) 20 (0.2%) 19 (0.2%) 19 (0.2%) 19 (0.2%) 11415 (98.2%)		0 (0.0%)
6	TestAge [numeric]	Mean (sd) : 8.9 (1.9) min < med < max: 5.2 < 8.9 < 13.5 IQR (CV) : 2.9 (0.2)	2583 distinct values		2159 (18.6%)
7	StudentGender [character]	1. F 2. M 3. X	5504 (47.3%) 6114 (52.6%) 10 (0.1%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
8	StudentEthnicity [character]	1. Am Indian/Alskn Nat 2. Asian 3. Black/African Am 4. Filipino 5. Hispanic 6. Missing 7. Multiple 8. Nat Hwiin/Othr Pac Islndr 9. White	1 (0.0%) 167 (1.4%) 31 (0.3%) 785 (6.8%) 10436 (89.7%) 37 (0.3%) 53 (0.5%) 13 (0.1%) 105 (0.9%)		0 (0.0%)
9	Special_Education [character]	1. N 2. Y	10097 (86.8%) 1531 (13.2%)		0 (0.0%)
10	Homeless [character]	1. N 2. Y	10721 (92.2%) 907 (7.8%)		0 (0.0%)
11	SocioEconomically [character]	1. N 2. Y	2095 (18.0%) 9533 (82.0%)		0 (0.0%)
12	TestDayName [character]	1. (Empty string) 2. Friday 3. Monday 4. Saturday 5. Sunday 6. Thursday 7. Tuesday 8. Wednesday	2159 (18.6%) 1887 (16.2%) 1565 (13.5%) 23 (0.2%) 12 (0.1%) 1884 (16.2%) 1937 (16.7%) 2161 (18.6%)		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
13	TestDate [character]	1. (Empty string) 2. 2020-03-11 3. 2020-03-13 4. 2020-03-05 5. 2020-03-10 6. 2022-03-03 7. 2022-05-23 8. 2020-03-12 9. 2020-03-03 10. 2022-04-20 [315 others]	2159 (18.6%) 335 (2.9%) 175 (1.5%) 136 (1.2%) 104 (0.9%) 93 (0.8%) 91 (0.8%) 88 (0.8%) 81 (0.7%) 80 (0.7%) 8286 (71.3%)		0 (0.0%)
14	OverallScore [integer]	Mean (sd) : 1201.2 (577.8) min < med < max: 0 < 1462 < 1731 IQR (CV) : 114 (0.5)	408 distinct values		0 (0.0%)
15	OverallLevel [integer]	Mean (sd) : 2 (1.3) min < med < max: 0 < 2 < 4 IQR (CV) : 2 (0.6)	0 : 2166 (18.6%) 1 : 1687 (14.5%) 2 : 2782 (23.9%) 3 : 3432 (29.5%) 4 : 1561 (13.4%)		0 (0.0%)
16	ExpectedAttendanceDays [numeric]	Mean (sd) : 176.8 (9) min < med < max: 0 < 180 < 180 IQR (CV) : 0 (0.1)	93 distinct values		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
17	DaysAttended [numeric]	Mean (sd) : 164.8 (17.3) min < med < max: 0 < 171 < 180 IQR (CV) : 17 (0.1)	127 distinct values		0 (0.0%)
18	AttendedPct [numeric]	Mean (sd) : 0.9 (0.1) min < med < max: 0.1 < 1 < 1 IQR (CV) : 0.1 (0.1)	743 distinct values		3 (0.0%)
19	EnrolledPct [numeric]	Mean (sd) : 1 (0) min < med < max: 0 < 1 < 1 IQR (CV) : 0 (0.1)	93 distinct values		0 (0.0%)
20	GradeEnrolledPct [numeric]	Mean (sd) : 3.6 (1.9) min < med < max: 0.4 < 3 < 7 IQR (CV) : 3 (0.5)	339 distinct values		0 (0.0%)
21	GradeAttendedPct [numeric]	Mean (sd) : 3.5 (1.9) min < med < max: 0.3 < 3 < 7 IQR (CV) : 3 (0.5)	1535 distinct values		3 (0.0%)
22	TeacherGender [character]	1. F 2. M	10614 (91.3%) 1014 (8.7%)		0 (0.0%)
23	TeacherTotalYearsOfService [integer]	Mean (sd) : 14.3 (8.9) min < med < max: 1 < 13 < 38 IQR (CV) : 14 (0.6)	38 distinct values		0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
					
24	TeacherEthnicity [character]	1. Asian 2. Black/African Am 3. Filipino 4. Hispanic 5. Missing 6. Multiple 7. Nat Hwiin/Othr Pac Islndr 8. White	375 (3.2%) 55 (0.5%) 587 (5.0%) 6739 (58.0%) 24 (0.2%) 8 (0.1%) 16 (0.1%) 3824 (32.9%)		0 (0.0%)

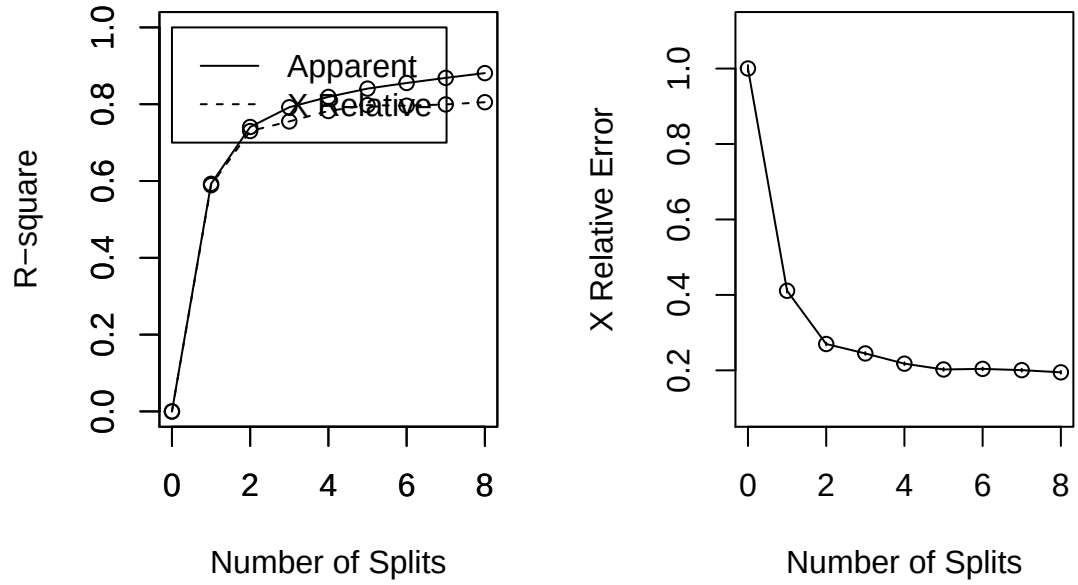
```
#    )
```

Decision Tree

```
# Convert target variable to factor to ensure proper interpretation by model
#rf_wine_train$quality <- as.factor(rf_wine_train$quality)

# Begin model...
rPartTree <- rpart(OverallLevel ~ ., data = df)

rpartTree2 <- as.party(rPartTree)

# R-Squared plot
par(mfrow=c(1,2))
rsq.rpart(rPartTree)
```

```
##
## Regression tree:
## rpart(formula = OverallLevel ~ ., data = df)
##
## Variables actually used in tree construction:
## [1] DOB          OverallScore TestDate
##
## Root node error: 20002/11628 = 1.7202
##
## n= 11628
##
##         CP nsplit rel error  xerror      xstd
## 1 0.592725      0   1.00000 1.00025 0.0088282
```

```
## 2 0.148054      1   0.40727 0.41096 0.0047565
## 3 0.051238      2   0.25922 0.26988 0.0035422
## 4 0.027032      3   0.20798 0.24483 0.0036265
## 5 0.021551      4   0.18095 0.21772 0.0036073
## 6 0.014677      5   0.15940 0.20237 0.0036193
## 7 0.013272      6   0.14472 0.20390 0.0037345
## 8 0.012456      7   0.13145 0.20059 0.0037243
## 9 0.010000      8   0.11900 0.19476 0.0036161
```

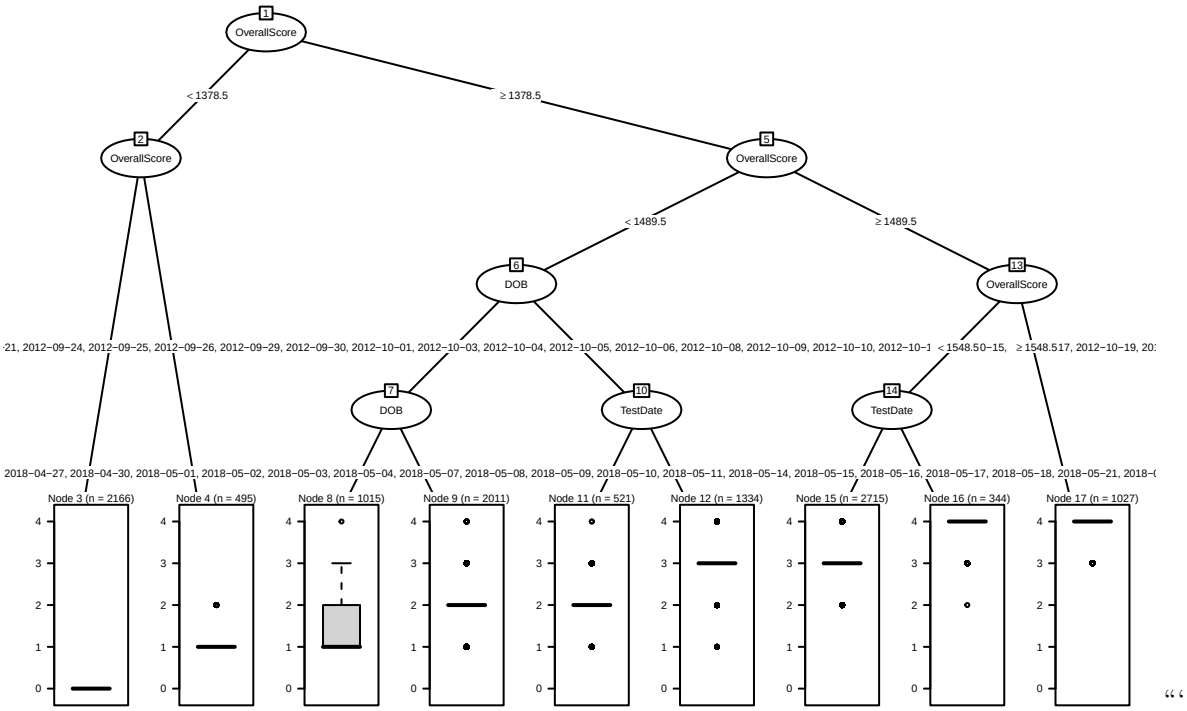



```r
# Results
rpartTree2
```

```
##
## Model formula:
## OverallLevel ~ AcademicYear + Stu_deID + School_deID + GradeLevel +
##     DOB + TestAge + StudentGender + StudentEthnicity + Special_Education +
##     Homeless + SocioEconomically + TestDayName + TestDate + OverallScore +
##     ExpectedAttendanceDays + DaysAttended + AttendedPct + EnrolledPct +
##     GradeEnrolledPct + GradeAttendedPct + TeacherGender + TeacherTotalYearsOfService +
##     TeacherEthnicity
##
## Fitted party:
## [1] root
## |   [2] OverallScore < 1378.5
## |   |   [3] OverallScore < 575: 0.000 (n = 2166, err = 0.0)
## |   |   [4] OverallScore >= 575: 1.034 (n = 495, err = 16.4)
## |   [5] OverallScore >= 1378.5
## |   |   [6] OverallScore < 1489.5
## |   |   |   [7] DOB in 2005-03-14, 2005-03-27, 2005-04-29, 2005-08-05, 2005-09-19, 2005-09-27, 2005-
## |   |   |   |   [8] DOB in 2005-03-27, 2005-08-05, 2005-09-19, 2005-10-13, 2005-12-03, 2006-01-04, 20
## |   |   |   |   [9] DOB in 2005-03-14, 2005-04-29, 2005-09-27, 2005-10-11, 2005-12-08, 2005-12-31, 20
## |   |   |   [10] DOB in 2009-02-06, 2009-09-03, 2009-09-05, 2009-09-08, 2009-09-10, 2009-09-14, 2009-
## |   |   |   |   [11] TestDate in 2018-02-14, 2019-02-08, 2019-02-11, 2019-02-20, 2019-02-21, 2019-02-
## |   |   |   |   [12] TestDate in 2018-02-20, 2018-03-09, 2018-04-03, 2018-04-04, 2018-04-05, 2018-04-
## |   |   [13] OverallScore >= 1489.5
## |   |   |   [14] OverallScore < 1548.5
```

```
## |   |   |   |   [15] TestDate in 2018-02-02, 2018-02-05, 2018-02-06, 2018-02-07, 2018-02-08, 2018-02-
## |   |   |   |   [16] TestDate in 2018-02-20, 2018-04-03, 2018-04-04, 2018-04-05, 2018-04-06, 2018-04-
## |   |   |   [17] OverallScore >= 1548.5: 3.852 (n = 1027, err = 129.5)
##
## Number of inner nodes:    8
## Number of terminal nodes: 9
```

```
plot(rpartTree2, gp = gpar(fontsize=4))
```



“