Predicting College Student Success Using a Classification Tree Model

Jae Hak Jung, Ph. D. Senior Manager, Institutional Research, Lone Star College
Kwanghee Jung, Ph.D., Associate Professor, Texas Tech University
Terrance Youngblood, Ph.D. candidate, Texas Tech University

.

Purpose

1. To demonstrate how to build a predictive model to identify college student success using educational data mining

2. To build a Power BI Dashboard to identify students at risk using the predictive model

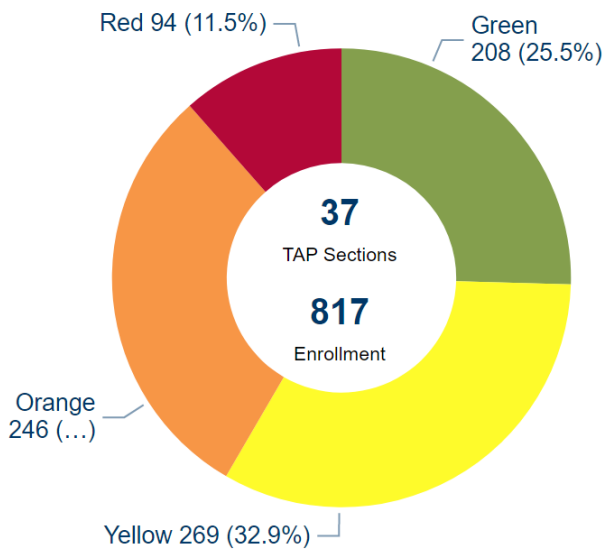
LSC Early Alert Power BI

- Received request from LSC Leadership and Faculty to develop an Early Alert system for At-risk students
- Asked to create an Early Alert model to predict student performance in a classroom
- Conducted regression analysis to identify significant predictors of course success
- Developed a beta version of the Early Alert PBI dashboard using identified predictors

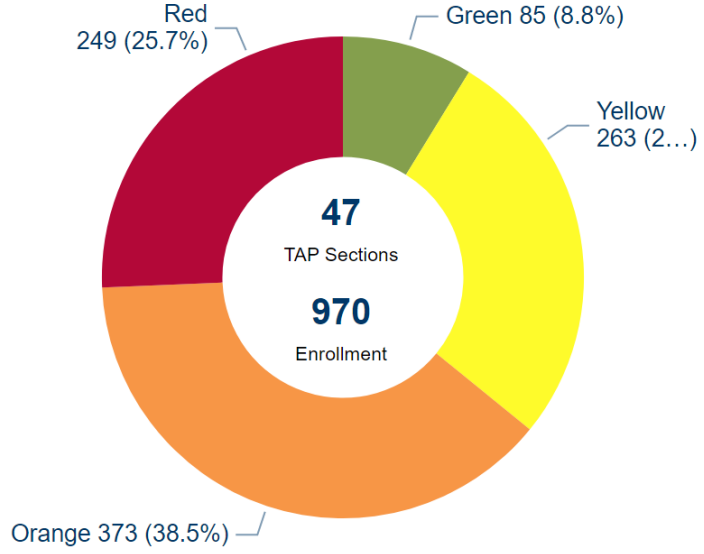
This visual does not support exporting

Term	Session	College	Location	Course	Section	Modality	Alert Level / Score	TAP Sections	Enrollment
Spring 2023	All	All	All	All	All	All	All	126	2,627

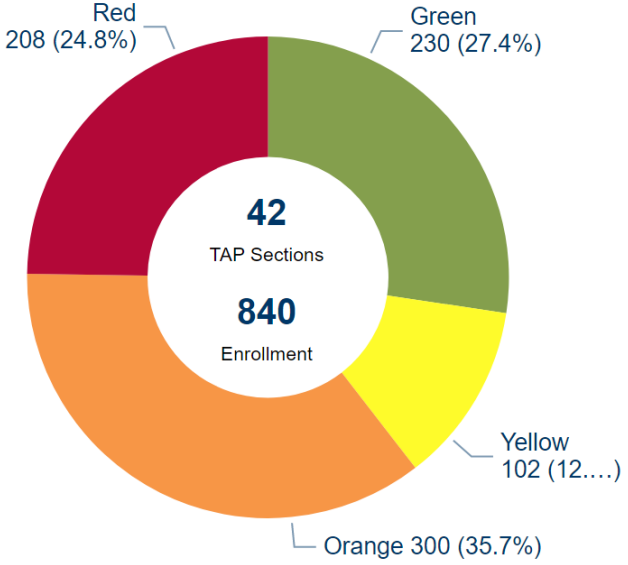
MATH 1314 Alert Level Summary



EDUC 1300 Alert Level Summary



ENGL 1301 Alert Level Summary



Alert Level Color Coding

Alert levels are based on students' course-specific alert score tiering. Alert scores range from 0 (most likely to succeed) to 10 (least likely to succeed) based on course-specific predictive measures based on regression analysis of historical student performance. Tiering is based on the following alert score ranges.

- **Green**: 0 - 1
- **Yellow**: 2 - 3
- **Orange**: 4 - 6
- **Red**: 7 - 10

Please view the report notes and data definitions within the Appendix.

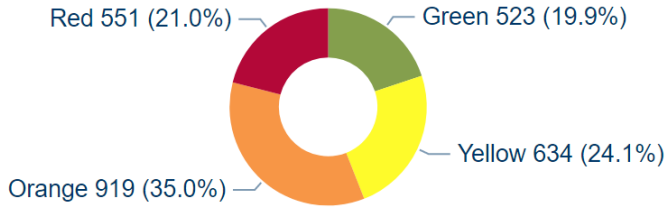
AIR-Approved

Last Updated
2/15/2023

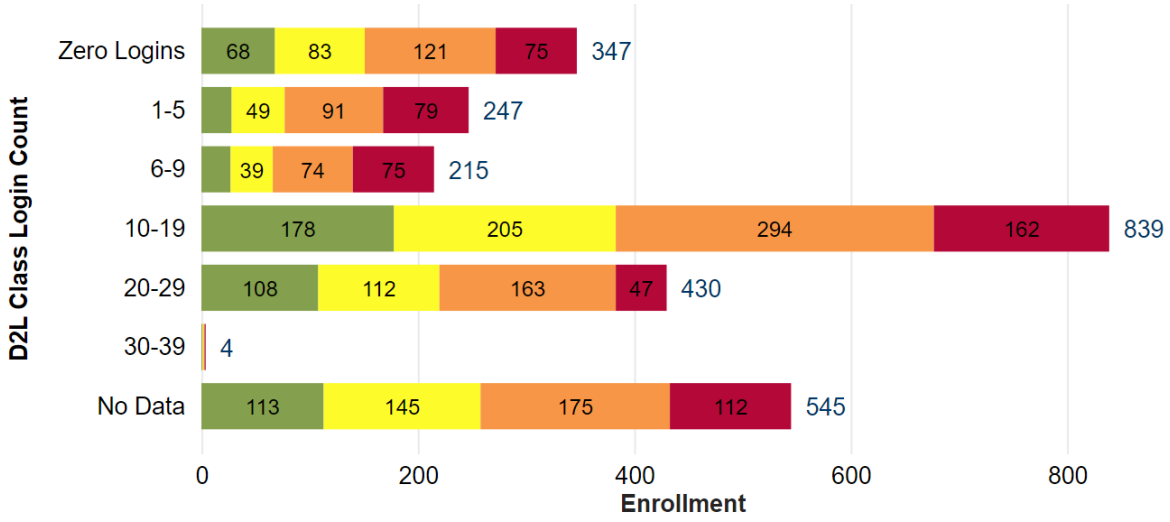
This visual does not support exporting

Term	Session	College	Location	Course	Section	Modality	Alert Level / Score	TAP Sections	Enrollment
Spring 2023	All	All	All	All	All	All	All	126	2,627

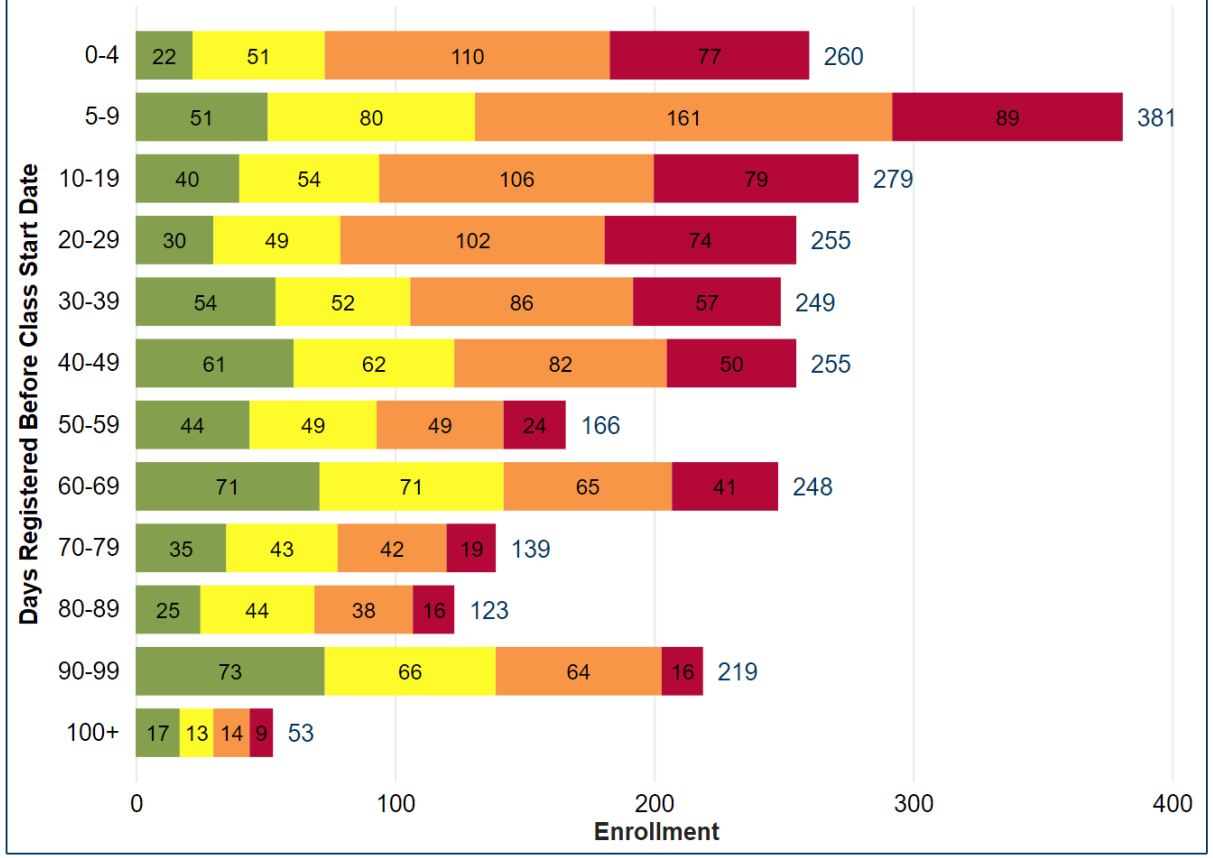
Students by Alert Groups



Students by D2L Class Login Count



Students by Days Registered Before Class Start



Note: 'No Data' for D2L class logins indicates that there is no D2L login data found for the specified class section and/or student. This may reflect students who have withdrawn. Entire classes missing may reflect effects of consolidating multiple sections via roster synchronization as the data model searches for the original D2L course shell only. 'Zero Logins' for D2L class login counts indicates that data exists for the class and student, but the student has not logged into the specified D2L class section.

Please view the report notes and data definitions within the Appendix.

AIR-Approved

This visual does not support exporting

Last Updated
2/15/2023

Analytics & Institutional Research (AIR)

Clear Filters & Slicers

This visual does not support exporting

126 TAP Sections

2,627 Enrollment

Term	Session	College	Location	Course	Section	Modality	Alert Level / Score
Spring 2023	All	All	All	All	All	All	All

Alert Level by Course

Alert Level by D2L Login & Registration Timing

Student & Section Information

Appendix

Alert & Class Interaction

- [ ] Select all
- [x] Select none
- [ ] Alert Score (Class Start)
- [ ] Days Reg Before Class Start
- [ ] D2L Class Logins
- [ ] Last D2L Class Access
- [ ] Final Course Grade

Student Contact Information

- [ ] Select all
- [x] Select none
- [ ] Student ID
- [ ] Student Name
- [ ] Street Line 1
- [ ] Street Line 2
- [ ] City
- [ ] State
- [ ] Zip Code
- [ ] LSC Email
- [ ] Personal Email
- [ ] Cell Phone
- [ ] Home Phone

Section Information

- [ ] Select all
- [x] Select none
- [ ] College
- [ ] Location
- [ ] Modality
- [ ] Corequisite Section Flag
- [ ] Session Code
- [ ] Faculty Name
- [ ] Faculty Email
- [ ] Course ID
- [ ] Term Code
- [ ] Class Number
- [ ] Class Start Date

Student and Section Information for EDUC 1300 & ENGL 1301 & MATH 1314 Spring 2023

Term	Course	Section	Alert Level	Enrollment
Spring 2023	EDUC 1300	1029	Red	5
Spring 2023	EDUC 1300	1030	Red	3
Spring 2023	EDUC 1300	1046	Red	5
Spring 2023	EDUC 1300	1047	Red	6
Spring 2023	EDUC 1300	1049	Red	8
Spring 2023	EDUC 1300	1050	Red	8
Spring 2023	EDUC 1300	1058	Red	5
Spring 2023	EDUC 1300	2801	Red	3
Spring 2023	EDUC 1300	2802	Red	10
Spring 2023	EDUC 1300	2804	Red	3
Spring 2023	EDUC 1300	2808	Red	5
Spring 2023	EDUC 1300	3E01	Red	3
Spring 2023	EDUC 1300	3E02	Red	6
Spring 2023	EDUC 1300	3E03	Red	5
Spring 2023	EDUC 1300	3E04	Red	8
Spring 2023	EDUC 1300	3E06	Red	6
Spring 2023	EDUC 1300	4W01	Red	6
Spring 2023	EDUC 1300	4W02	Red	7
Spring 2023	EDUC 1300	4W03	Red	4
Spring 2023	EDUC 1300	4W06	Red	10
Spring 2023	EDUC 1300	4W08	Red	7
Spring 2023	EDUC 1300	4W10	Red	6
Spring 2023	EDUC 1300	5033	Red	4
Spring 2023	EDUC 1300	5034	Red	6
Total				2,627

Please view the report notes and data definitions within the Appendix.

AIR-Approved

This visual does not support exporting

Last Updated

2/15/2023

Implication

- Early Alert Power BI can help college administrators and faculty identify at-risk students early on in the semester or academic year.
- By identifying at-risk students early, interventions and support services can be put in place to help these students succeed and avoid dropping out of college.
- Early Alert Power BI can provide a comprehensive view of a student's academic performance and behavior, which can help identify patterns and potential issues that may contribute to their risk status.
- Early Alert Power BI can also help colleges and universities meet accountability and reporting requirements related to student success and retention.

Limitation and implication

- Regression analysis may not be easily applicable to the Early Alert Power BI, as it may be challenging to interpret and integrate the regression results into this system.
- Selecting a cut point for an important predictor in a regression analysis can be subjective and may not accurately capture the underlying relationships in the data.
- Regression analysis may struggle to categorize students accurately, particularly if the data is complex or multidimensional.
- Meeting all of the assumptions of regression analysis can be challenging, particularly with skewed data, which may lead to biased or inaccurate results.

Lone Star College Data

- From Fall 2017 to Spring 2022 semester database
- Students who enrolled in Online MATH 1314.
- 22,182 students

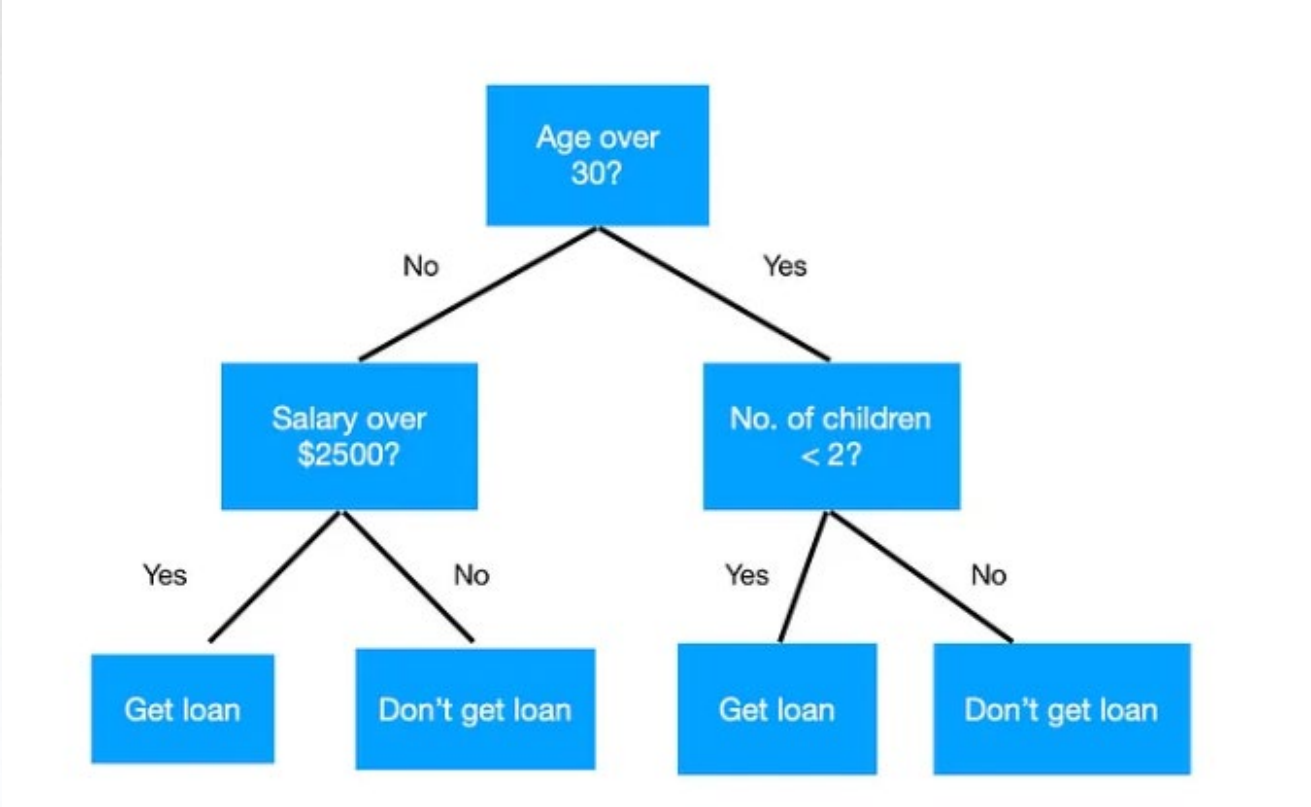
Predictors of MATH 1314 Success

1) Cumulative GPA
2) Previous term credit hours
3) Met TSI Math Milestone
4) Ratio between Credits Earned and Credits Attempted
5) Enrolled Full-Time or More
6) Financial aid
7) Registration Time: # Days prior to start of term (Early/Late Registration)
8) Age
9) Race/ethnicity
10) Gender
11) Others

Decision Tree Algorithm
: Classification and Regression Trees

Decision Trees

- Decision trees are a popular machine learning method that use a tree-like structure to model decisions and their possible consequences.



https://medium.com/@borcandumitrumarius/decision-tree-classifiers-explained-e47a5b68477a

Classification and Regression Trees (CART)

- CART is a type of decision tree algorithm that can be used for both classification and regression problems
- CART builds a decision tree by recursively splitting the data into smaller subsets, based on the features that best separate the classes or minimize the sum of squared errors in the case of regression.

Classification and Regression Trees (CART)

- CART has several strengths, including the ability to handle both categorical and continuous variables, the ability to handle missing values, and the ability to model nonlinear relationships between features and target variables.
- CART also produces interpretable models that can be easily visualized and understood by non-experts.
- The R implementation of the CART algorithm is called "RPART" (Recursive Partitioning And Regression Trees

http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/

Installing and Loading R packages

- "rpart" for computing decision tree models
- "rpart.plot" for visualizing decision tree models

```
# For decision tree model
install.packages("rpart")
library(rpart)

# For data visualization
install.packages("rpart.plot")
library(rpart.plot)
```

https://rpubs.com/camguild/803096

Basics and Visual Representation

- The algorithm of decision tree models works by repeatedly partitioning the data into multiple sub-spaces
- The decision rules generated by the CART predictive model are generally visualized as a binary tree

https://rpubs.com/camguild/803096

Lone Star College Data

- A tree model predicting student MATH 1314 success
- 22,182 students with 23 variables

GPA-LSCdata-v3-ftic.R | GPA-LSCdata-v2.R | subset_data

Filter

	gpab	repeatcours	cumgpab	curcredenroll	cumcredatt	cumcredcomp	regdaytpt	priordc	tsimath	tsiread	tsiwrite	alertsall
7	Non-success	1	2.18	12	42	27	28	0	0	1	1	1
10	Success	0	3.54	7	19	13	-13	0	1	1	1	0
12	Success	0	3.64	10	28	28	93	0	1	1	1	0
13	Non-success	0	3.71	3	90	86	71	0	1	1	1	0
14	Non-success	0	2.61	3	100	78	132	0	1	1	1	0
15	Success	1	2.61	3	103	78	8	0	1	1	1	0
16	Success	0	4.00	3	38	28	-38	0	1	1	1	0
17	Non-success	0	3.87	6	55	55	34	0	1	1	1	0
19	NA	1	2.70	3	42	27	0	0	1	1	1	0
21	Non-success	0	2.52	16	50	28	-51	0	0	1	1	0
22	Success	1	2.46	17	104	50	-1	0	0	1	1	1
23	Success	0	3.93	9	46	43	0	0	1	1	1	0
26	Success	0	4.00	6	75	62	-2	0	1	1	1	0
27	Success	0	3.21	13	119	87	68	0	0	1	1	2
30	Non-success	0	0.00	12	23	0	7	0	1	1	1	0
31	Non-success	1	3.29	9	53	21	38	0	1	1	1	1
32	Non-success	1	3.24	6	72	37	-8	0	1	1	1	1

Showing 1 to 17 of 22,182 entries, 24 total columns

Splitting Training and Test Sets

- A tree model predicting student MATH 1314 success

```
# Split the data into training and test set
set.seed(123)
sample_data = sample.split(sub_data$gpab, SplitRatio = 0.8)
train_data <- subset(sub_data, sample_data == TRUE)
test_data <- subset(sub_data, sample_data == FALSE)
```

Model Fitting

- A tree model predicting student MATH 1314 success

```
# Modeling fitting
fit.tree = rpart(gpab ~ ., data=dat2, method = "class", cp=0.005)
fit.tree

# Visualizing the unpruned tree
rpart.plot(fit.tree)
```

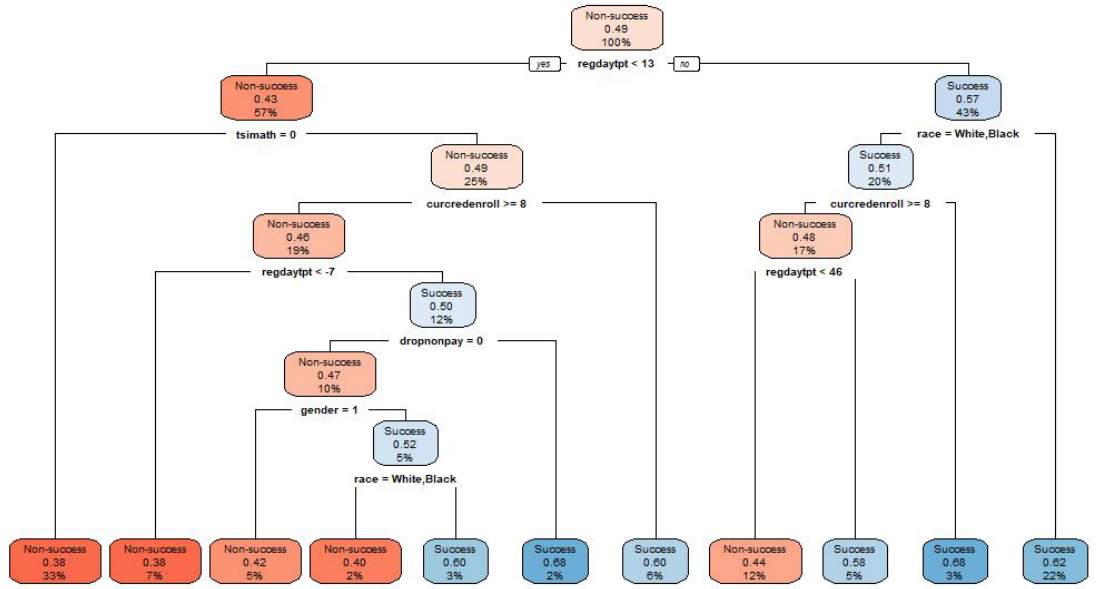
- cp (complexity parameter) determines how deep the tree will grow

RStudio Interface: Decision Tree

```
#select columns by index if necessary
lscdat <-dat          #[c(31,28,6,12,13,17,18,22,23,25,32,33,34,35,37,38)]
head(lscdat)
str(lscdat)

detach(lscdat)
attach(lscdat)

#table(race)
names(lscdat)

# > names(lscdat)
# [1] "gpab"         "repeatcours"   "cumgpab"       "curcredenroll" "cumcredatt"    "cumc
# [9] "tsimath"      "tsiread"       "tsiwrite"      "alertsall"     "alertcourse"   "apcr
# [17] "aid"          "gender"        "age"           "totaltermat"   "dropnonpay"    "rac

# Code **** as a factor variable
lscdat$gpab = as.factor(lscdat$gpab)
lscdat$repeatcours = as.factor(lscdat$repeatcours)
lscdat$priordc = as.factor(lscdat$priordc)
lscdat$tsimath = as.factor(lscdat$tsimath)
lscdat$tsiread = as.factor(lscdat$tsiread)
lscdat$tsiwrite = as.factor(lscdat$tsiwrite)
lscdat$alertsall = as.factor(lscdat$alertsall)
lscdat$alertcourse = as.factor(lscdat$alertcourse)
lscdat$veteran = as.factor(lscdat$veteran)
lscdat$aid = as.factor(lscdat$aid)
lscdat$gender = as.factor(lscdat$gender)
```

```
   2) regdaytpt< 12.5 1539  657 Non-success (0.5730994 0.4269006)
     4) tsimath=0 877  333 Non-success (0.6202965 0.3797035) *
     5) tsimath=1 662  324 Non-success (0.5105740 0.4894260)
      10) curcredenroll>=8 503  229 Non-success (0.5447316 0.4552684)
        20) regdaytpt< -7.5 195   74 Non-success (0.6205128 0.3794872) *
        21) regdaytpt>=-7.5 308  153 Success (0.4967532 0.5032468)
          42) dropnonpay=0 258  121 Non-success (0.5310078 0.4689922)
            84) gender=1 127   53 Non-success (0.5826772 0.4173228) *
            85) gender=0 131   63 Success (0.4809160 0.5190840)
             170) race=white,Black 53   21 Non-success (0.6037736 0.3962264) *
             171) race=Hispanic,Other 78   31 Success (0.3974359 0.6025641) *
          43) dropnonpay=1 50   16 Success (0.3200000 0.6800000) *
      11) curcredenroll< 8 159   64 Success (0.4025157 0.5974843) *
   3) regdaytpt>=12.5 1138  489 Success (0.4297012 0.5702988)
     6) race=White,Black 536  263 Success (0.4906716 0.5093284)
      12) curcredenroll>=7.5 459  221 Non-success (0.5185185 0.4814815)
        24) regdaytpt< 45.5 331  147 Non-success (0.5558912 0.4441088) *
        25) regdaytpt>=45.5 128   54 Success (0.4218750 0.5781250) *
      13) curcredenroll< 7.5 77   25 Success (0.3246753 0.6753247) *
     7) race=Hispanic,Other 602  226 Success (0.3754153 0.6245847) *
> # Visualizing the tree
> rpart.plot(fit.tree, box.palette="RdBu")
>
```

Data	
fit.tree	List of 15
lscdat	99791 obs. of 24 variables
sub_data	4260 obs. of 23 variables
subset_data	4260 obs. of 24 variables
test_data	1583 obs. of 23 variables
train_data	2677 obs. of 23 variables
Values	
GCtorture	FALSE
sample_data	logi [1:4260] TRUE TRUE TRUE FALSE TRUE FALSE ...

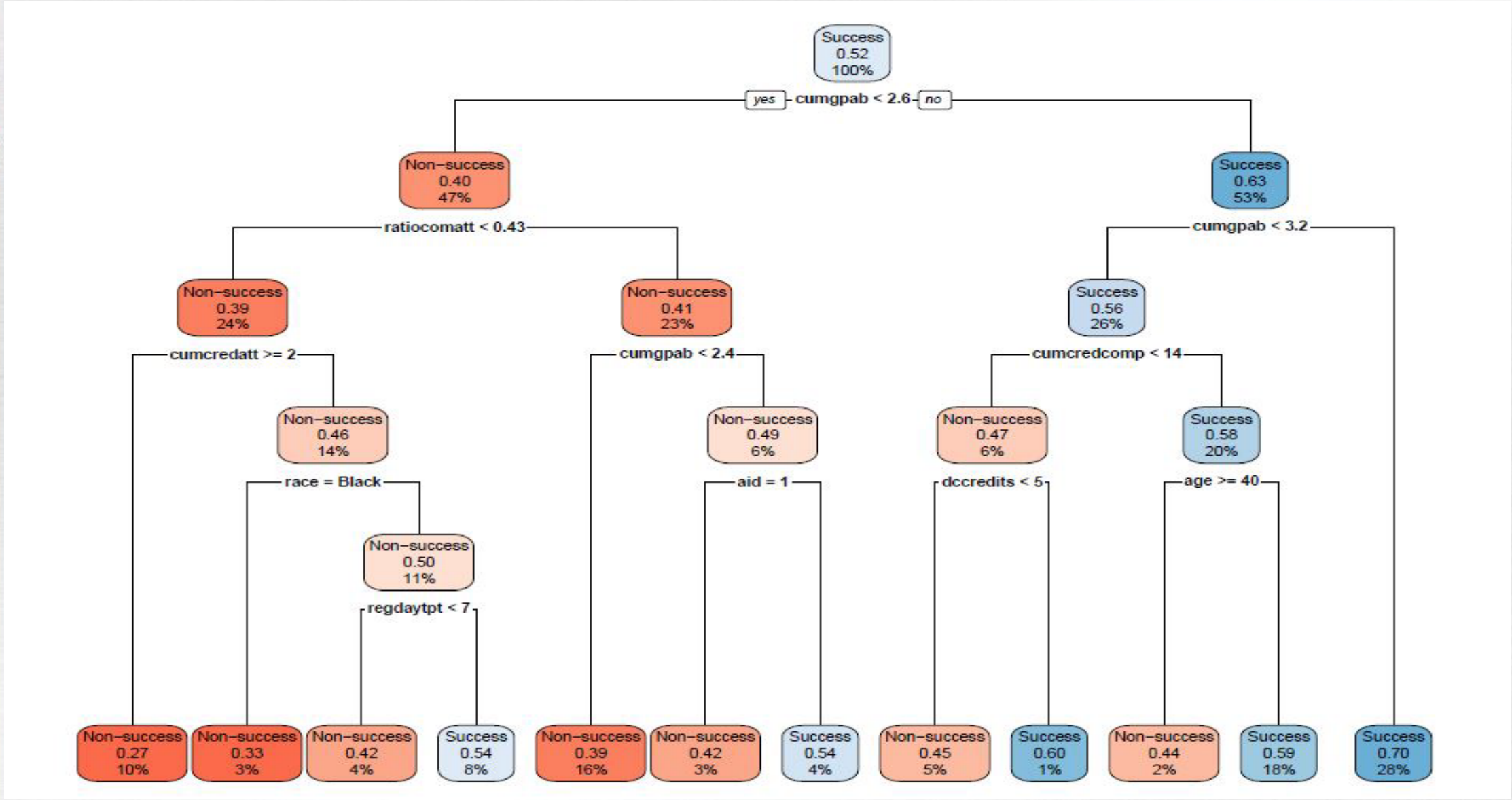


Decision Tree

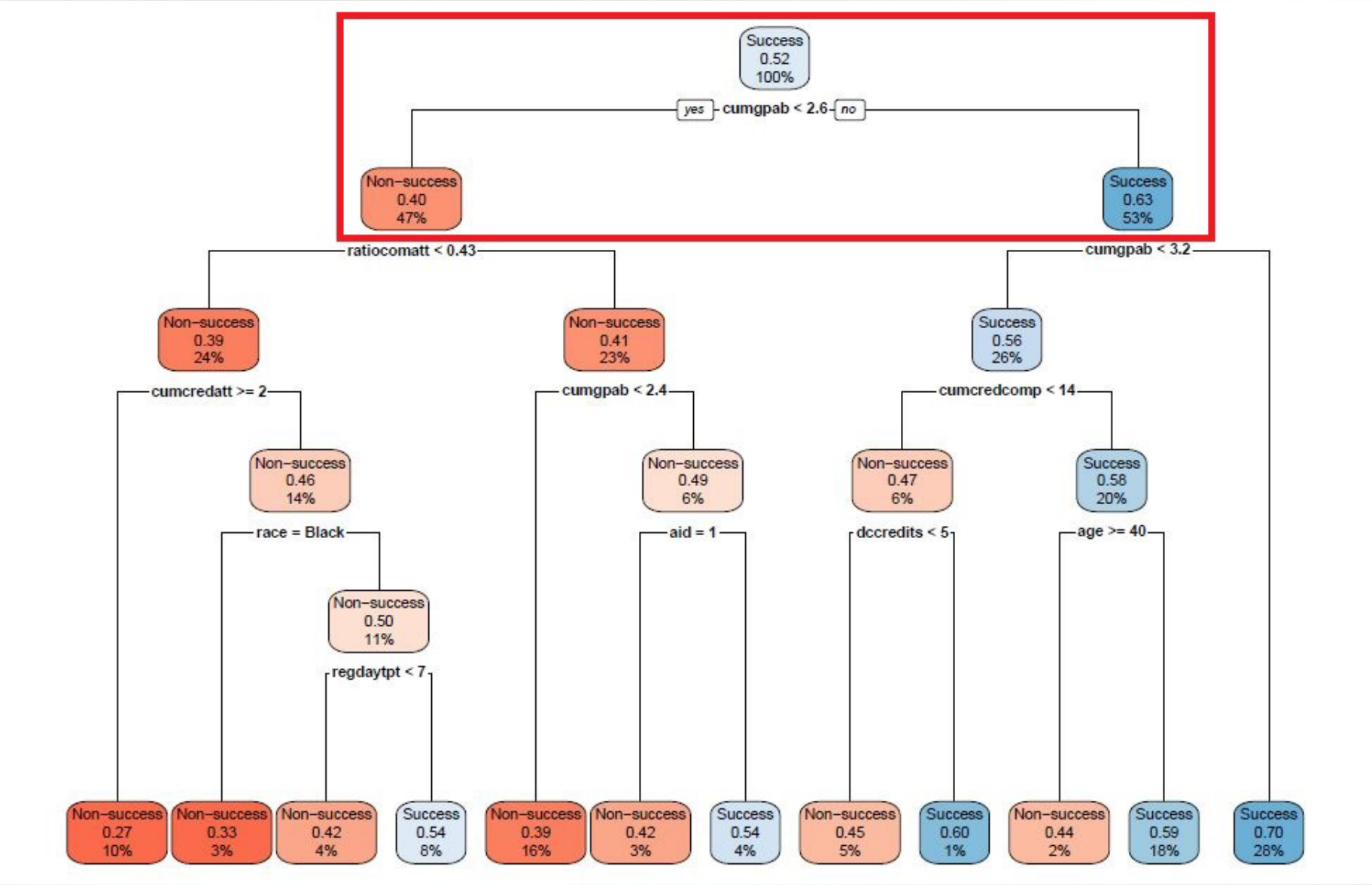
Decision Tree: MATH 1314 success: Non-FTIC

```
n= 15143

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 15143 7203 Success (0.4756653 0.5243347)
   2) cumgpab< 2.605 7067 2833 Non-success (0.5991227 0.4008773)
     4) ratiocomatt< 0.4263776 3653 1418 Non-success (0.6118259 0.3881741)
       8) cumcredatt>=1.5 1467  403 Non-success (0.7252897 0.2747103) *
       9) cumcredatt< 1.5 2186 1015 Non-success (0.5356816 0.4643184)
        18) race=Black 449  147 Non-success (0.6726058 0.3273942) *
        19) race=White,Hispanic,Other 1737  868 Non-success (0.5002879 0.4997121)
          38) regdaytpt< 6.5 565  239 Non-success (0.5769912 0.4230088) *
          39) regdaytpt>=6.5 1172  543 Success (0.4633106 0.5366894) *
     5) ratiocomatt>=0.4263776 3414 1415 Non-success (0.5855302 0.4144698)
      10) cumgpab< 2.425 2440  941 Non-success (0.6143443 0.3856557) *
      11) cumgpab>=2.425 974  474 Non-success (0.5133470 0.4866530)
        22) aid=1 421  177 Non-success (0.5795724 0.4204276) *
        23) aid=0 553  256 Success (0.4629295 0.5370705) *
   3) cumgpab>=2.605 8076 2969 Success (0.3676325 0.6323675)
     6) cumgpab< 3.225 3907 1732 Success (0.4433069 0.5566931)
      12) cumcredcomp< 13.5 873  412 Non-success (0.5280641 0.4719359)
        24) dccredits< 5 747  337 Non-success (0.5488621 0.4511379) *
        25) dccredits>=5 126   51 Success (0.4047619 0.5952381) *
      13) cumcredcomp>=13.5 3034 1271 Success (0.4189189 0.5810811)
        26) age>=39.5 237  104 Non-success (0.5611814 0.4388186) *
        27) age< 39.5 2797 1138 Success (0.4068645 0.5931355) *
     7) cumgpab>=3.225 4169 1237 Success (0.2967138 0.7032862) *
```

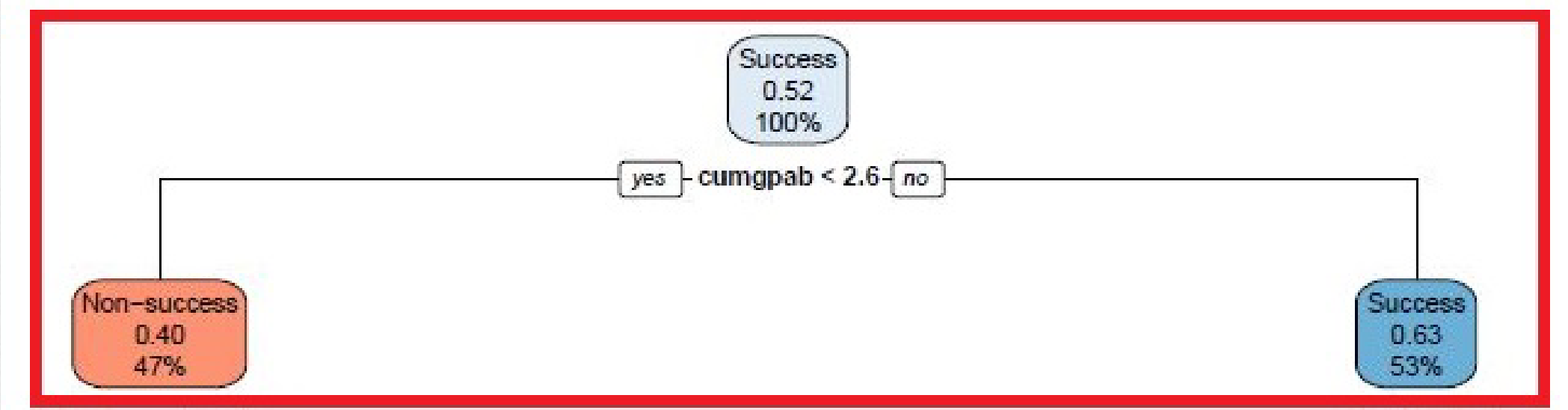
Decision Tree: MATH 1314 success: Non-FTIC



Decision Tree: Cumulative GPA



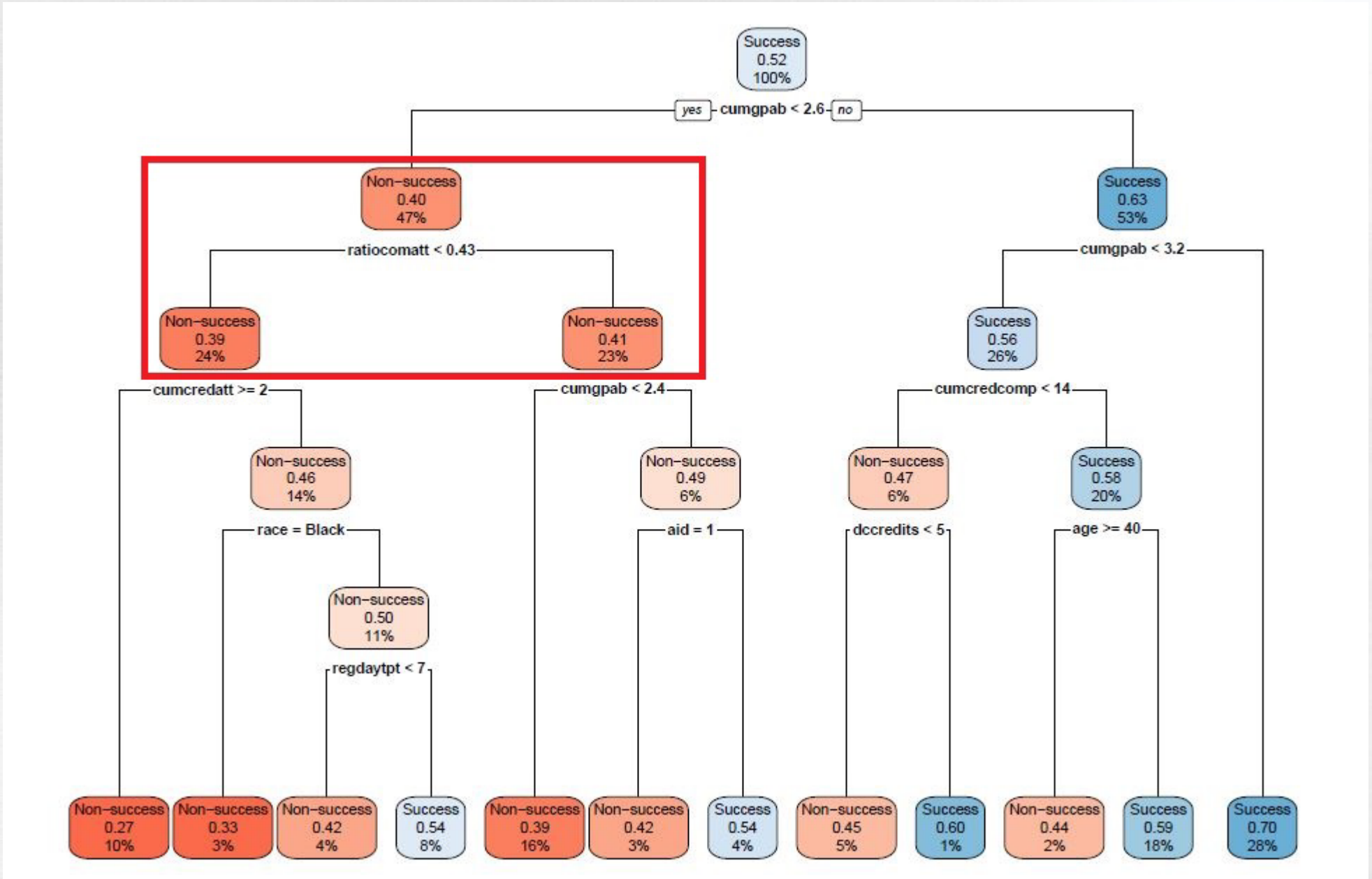
Decision Tree: Cumulative GPA



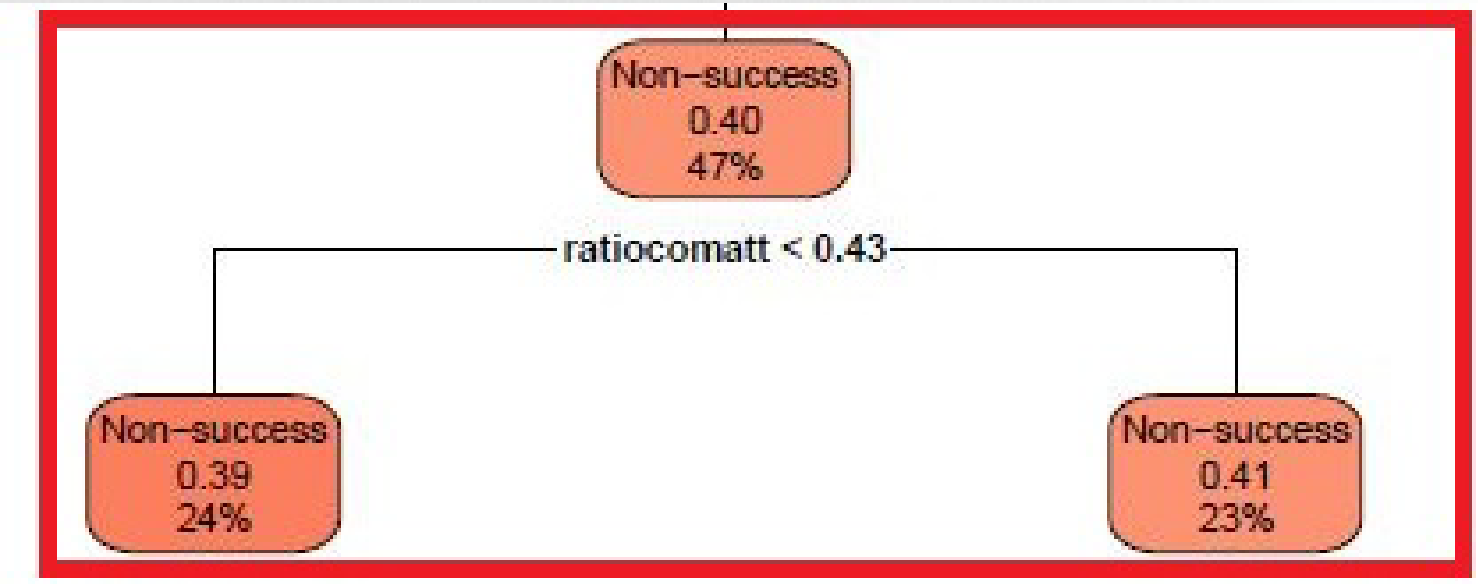
- Automatically detect the top predictor and find the cut score.

  (Cumulative GPA < 2.6)
- If students have less than a 2.6 on their cumulative GPA, only 40% of students could successfully complete MATH 1314.
- If students have higher than a 2.6 on their cumulative GPA, 63% of students could successfully complete MATH 1314.

Decision Tree: Ratio between CUM Credits Earned and CUM Credits Earned

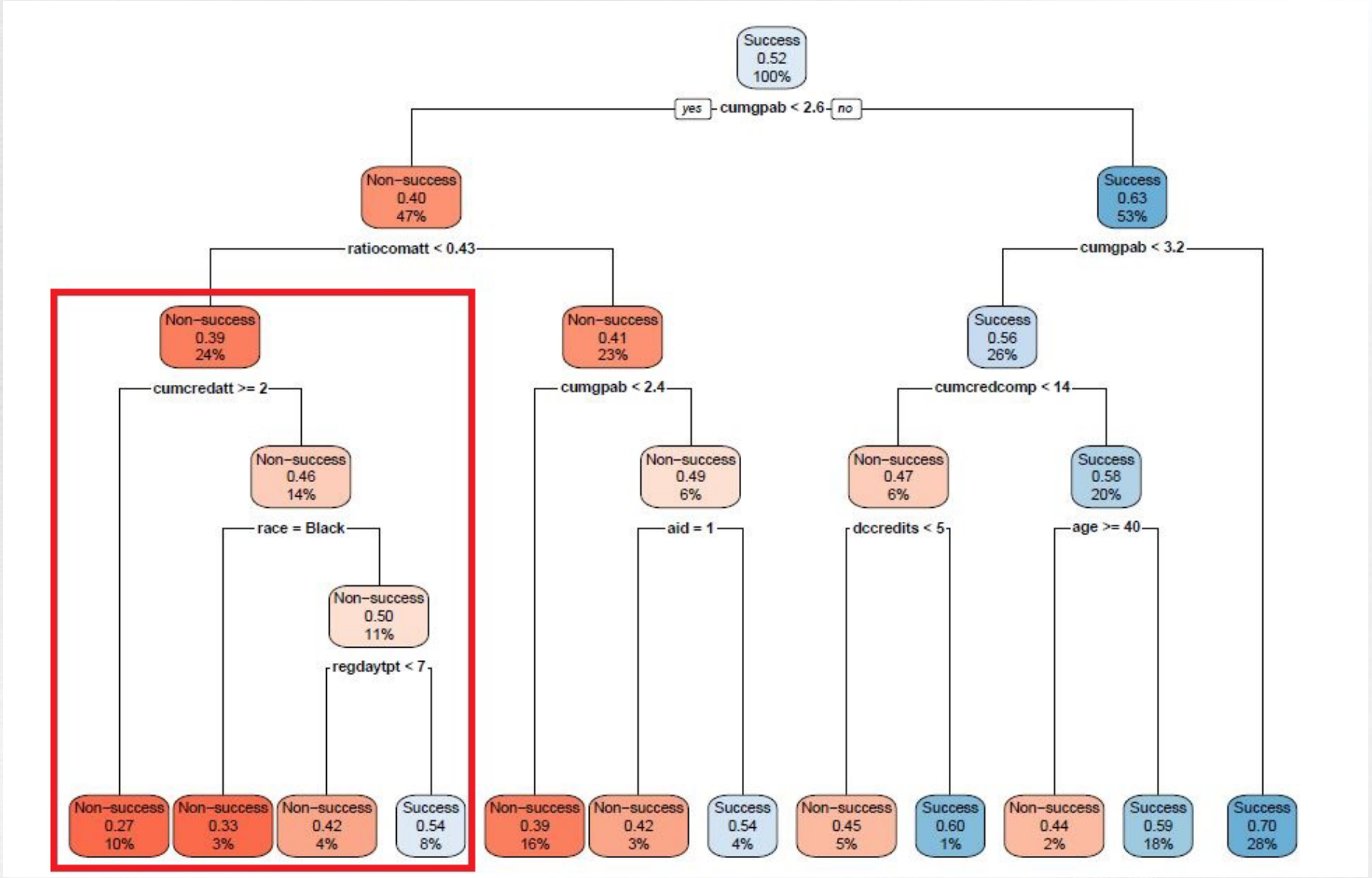


Decision Tree: Ratio between CUM Credits Earned and CUM Credits Earned

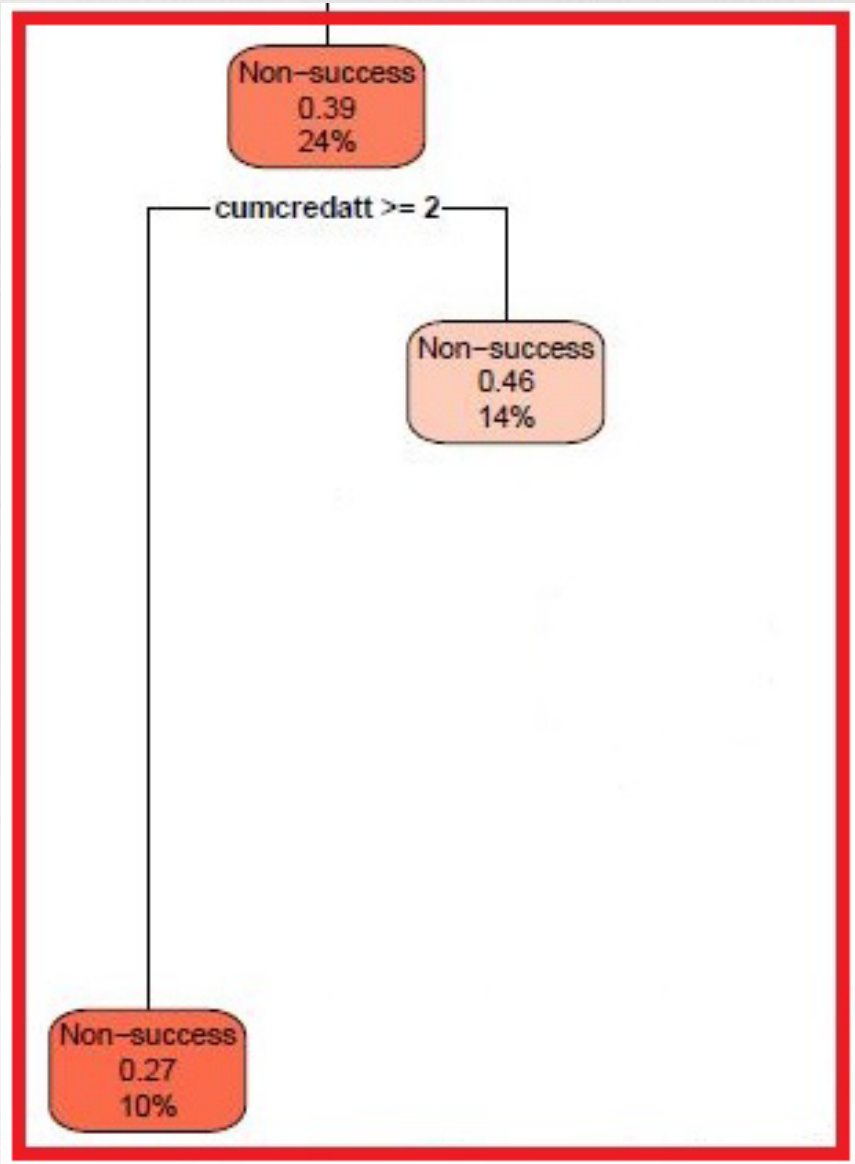


- Automatically detect the next top predictor and find the cut score. (Credits Earned / Credits Attempted < 0.43)
- If yes, only 39% of students could successfully complete MATH 1314.
- If no, only 41% of students could successfully complete MATH 1314.

Decision Tree: Cumulative Credit Attempted



Decision Tree: Cumulative Credit Attempted



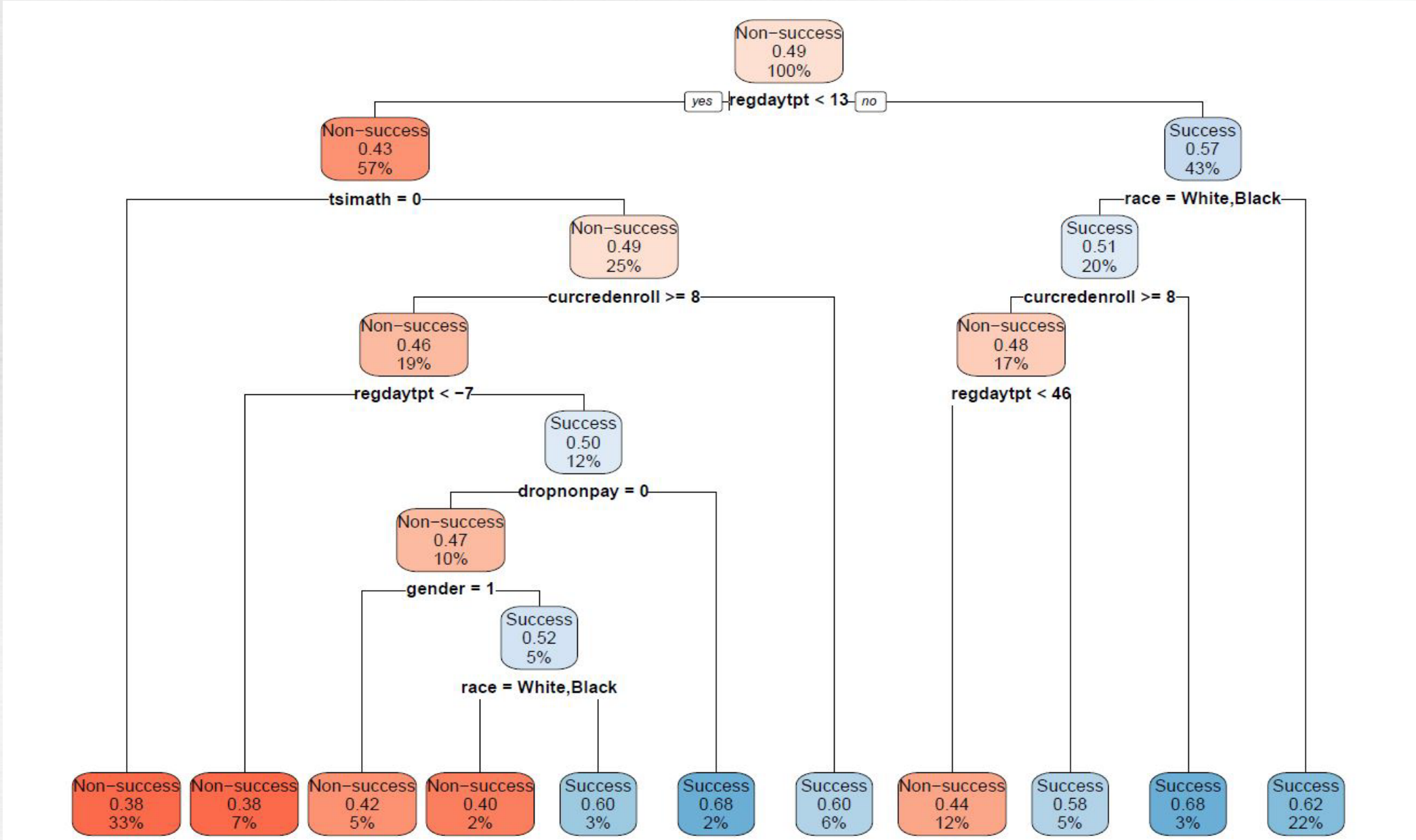
- Automatically detect the next top predictor and find the cut score.

  (Cumulative Credit Attempted >=2)
- If yes, only 27% of students could successfully complete MATH 1314.
- If no, 46% of students could successfully complete MATH 1314.

Decision Tree: MATH 1314 success: FTIC

```
> fit.tree
n= 2677

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 2677 1306 Non-success (0.5121405 0.4878595)
    2) regdaytpt< 12.5 1539  657 Non-success (0.5730994 0.4269006)
      4) tsimath=0 877  333 Non-success (0.6202965 0.3797035) *
      5) tsimath=1 662  324 Non-success (0.5105740 0.4894260)
       10) curcredenroll>=8 503  229 Non-success (0.5447316 0.4552684)
         20) regdaytpt< -7.5 195   74 Non-success (0.6205128 0.3794872) *
         21) regdaytpt>=-7.5 308  153 Success (0.4967532 0.5032468)
           42) dropnonpay=0 258  121 Non-success (0.5310078 0.4689922)
             84) gender=1 127   53 Non-success (0.5826772 0.4173228) *
             85) gender=0 131   63 Success (0.4809160 0.5190840)
              170) race=White,Black 53   21 Non-success (0.6037736 0.3962264) *
              171) race=Hispanic,Other 78   31 Success (0.3974359 0.6025641) *
           43) dropnonpay=1 50   16 Success (0.3200000 0.6800000) *
       11) curcredenroll< 8 159   64 Success (0.4025157 0.5974843) *
    3) regdaytpt>=12.5 1138  489 Success (0.4297012 0.5702988)
      6) race=White,Black 536  263 Success (0.4906716 0.5093284)
       12) curcredenroll>=7.5 459  221 Non-success (0.5185185 0.4814815)
         24) regdaytpt< 45.5 331  147 Non-success (0.5558912 0.4441088) *
         25) regdaytpt>=45.5 128   54 Success (0.4218750 0.5781250) *
       13) curcredenroll< 7.5 77   25 Success (0.3246753 0.6753247) *
      7) race=Hispanic,Other 602  226 Success (0.3754153 0.6245847) *
```

Decision Tree: MATH 1314 success: FTIC



Applications of Decision Tree Models

- Potential predictors (TSI scores, student non-cognitive factor, D2L data, Tutoring data)
- Apply to course success for each courses or subject
- Graduation Rates
- Retention/Persistence
- Student Purge
- Transfer

Future Task and Implication

- Create a Power BI report that incorporates Azure Machine Learning. This involves integrating Azure Machine Learning into your Power BI report by connecting to the machine learning models and datasets stored in Azure.
- By integrating Azure Machine Learning into the Power BI report, we can leverage the power of machine learning to generate new insights and findings and then share those insights with others in a way that is easy to understand and interact with.

Thank you

Any questions?