Predicting Student Success In STAT 119

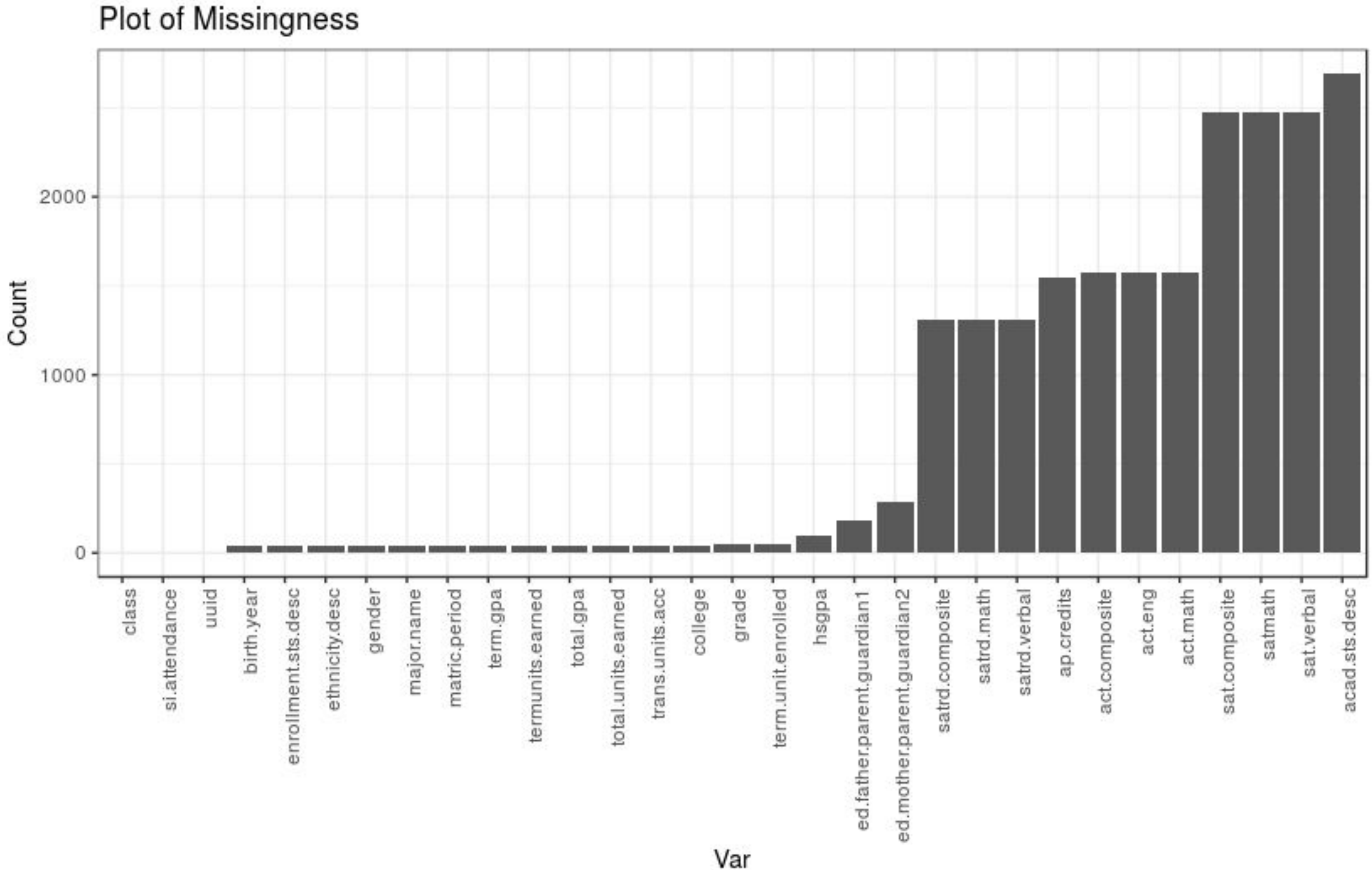
By Tristan Hillis, Travis Klipp, Ryan Tomiyama

Description Of Dataset and Problem

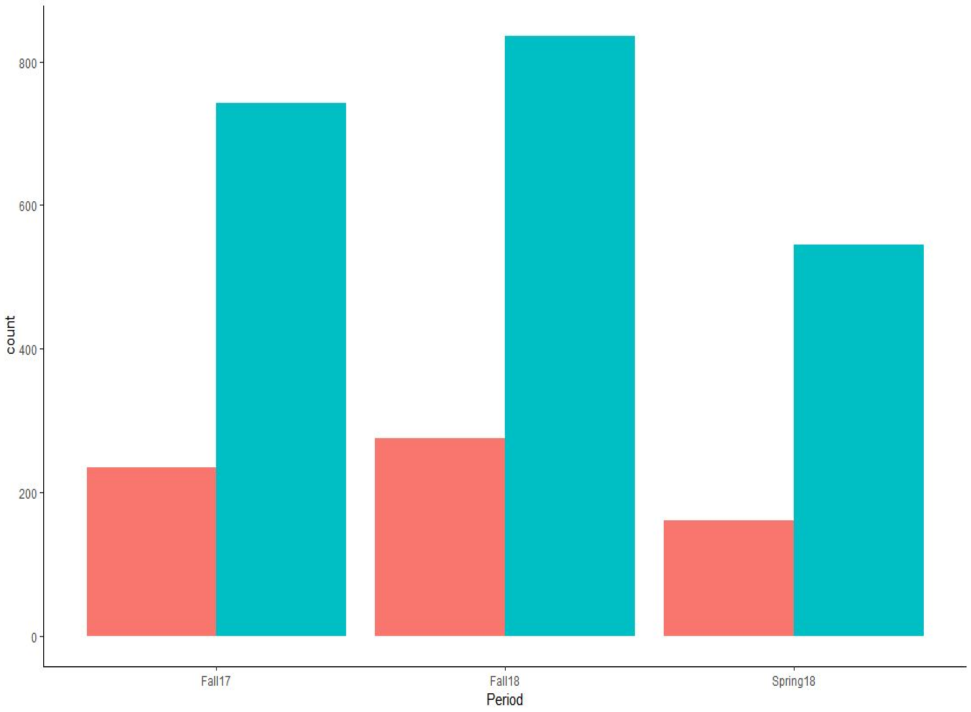
- Demographic and class data
    - Demographic: Prior academic statistics (HS GPA, SAT), ethnicity, college, year, etc.
    - Class: Grades for various assignments
    - Response: Final grade in the class transformed as Pass/Fail
    - Over 3000 instances collected over 3 semesters (Fall 2017-Fall 2018), 131 total variables
- Problem:
    - 24% failure rate-high failure rates lead to bottlenecking
- Proposed Solution:
    - Develop an accurate model that can detect early in the semester if a student is likely to fail the class
    - “Early Alert” system intervention

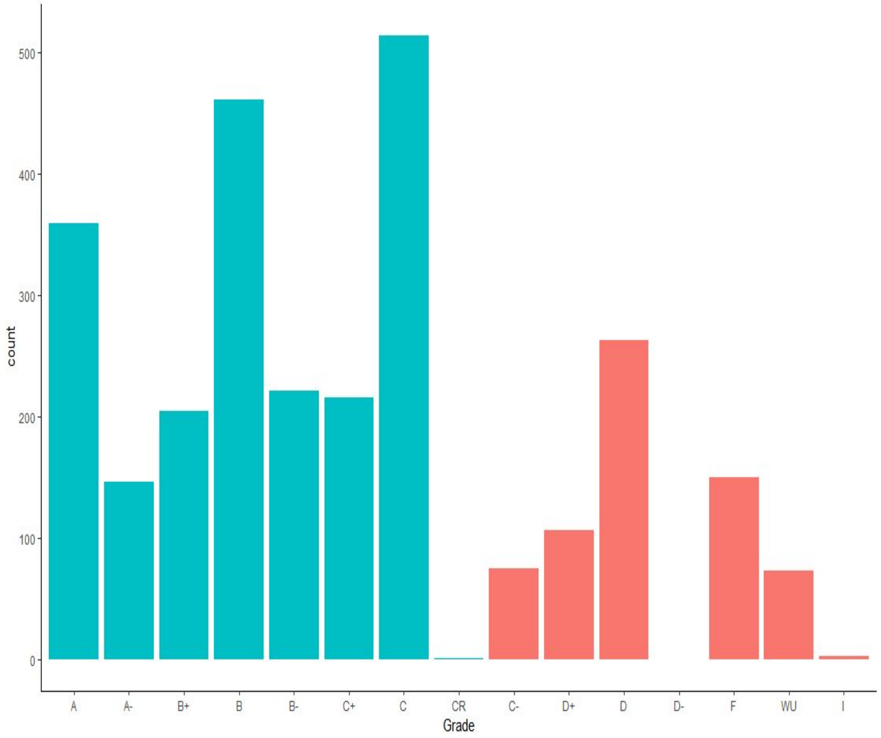
Data Preprocessing

1. Drop irrelevant and repetitive variables
   - Lots of missing values
   - Second major, ethnicity subgroups, etc
2. Throw out variables highly correlated to response
   - Term GPA, Total GPA (after class), academic status, term units earned, etc.
3. Imputing:
   - Conversion from ACT to SAT using official concordance tables
   - Use MICE (Multivariate Imputation by Chained Equations) w/ RF method to impute HSGPA and SAT
4. Throw out observations where we can’t impute missing features
   - Fathers/mothers education, ethnicity, grade, etc.

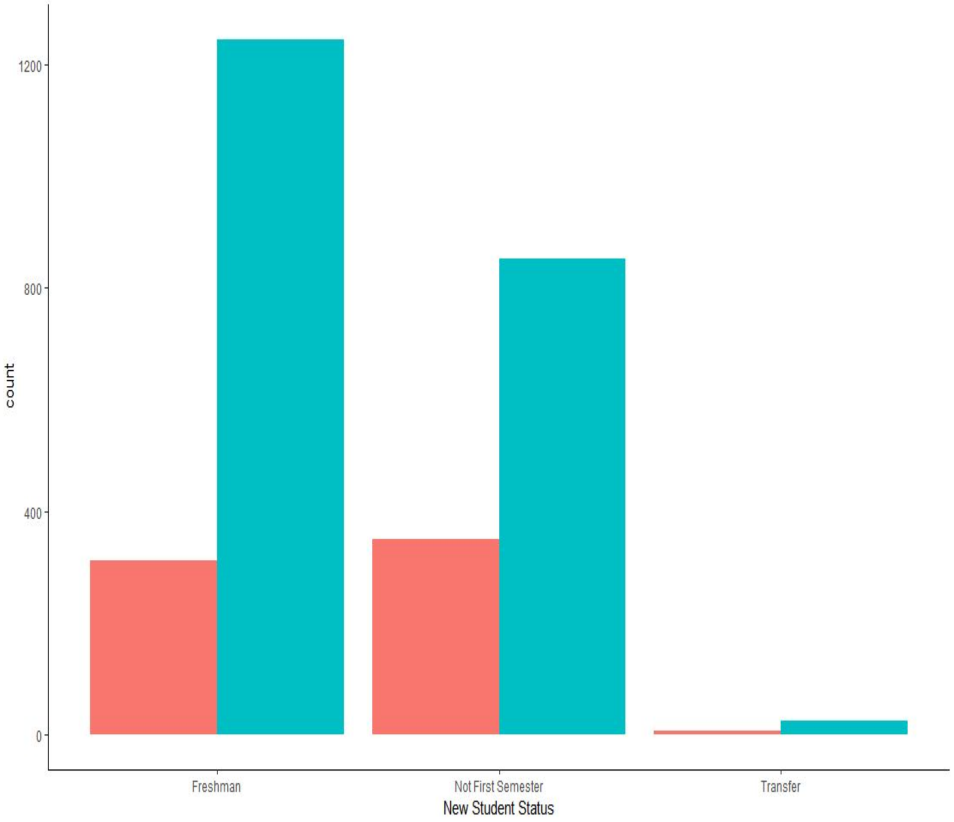

Plot of Missingness
Count
2000
1000
0
Var
class
si.attendance
uuid
birth.year
enrollment.sts.desc
ethnicity.desc
gender
major.name
matric.period
term.gpa
termunits.earned
total.gpa
total.units.earned
trans.units.acc
college
grade
term.unit.enrolled
hsgpa
ed.father.parent.guardian1
ed.mother.parent.guardian2
satrd.composite
satrd.math
satrd.verbal
ap.credits
act.composite
act.eng
act.math
sat.composite
satmath
sat.verbal
acad.sts.desc

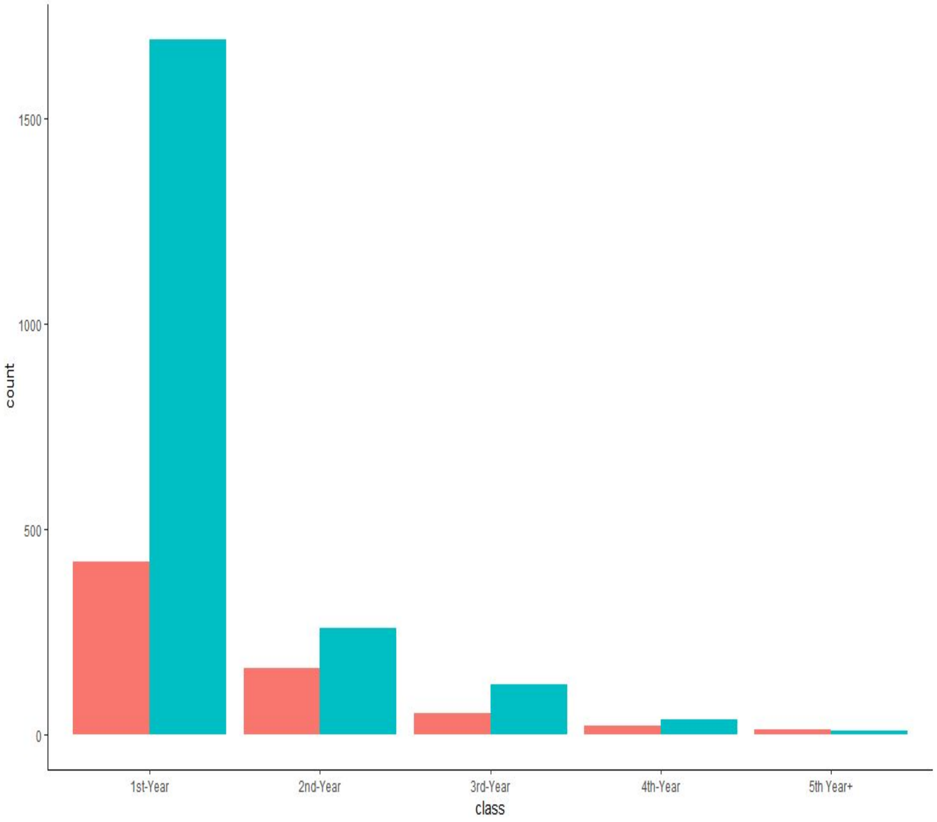
Exploratory Data Analysis



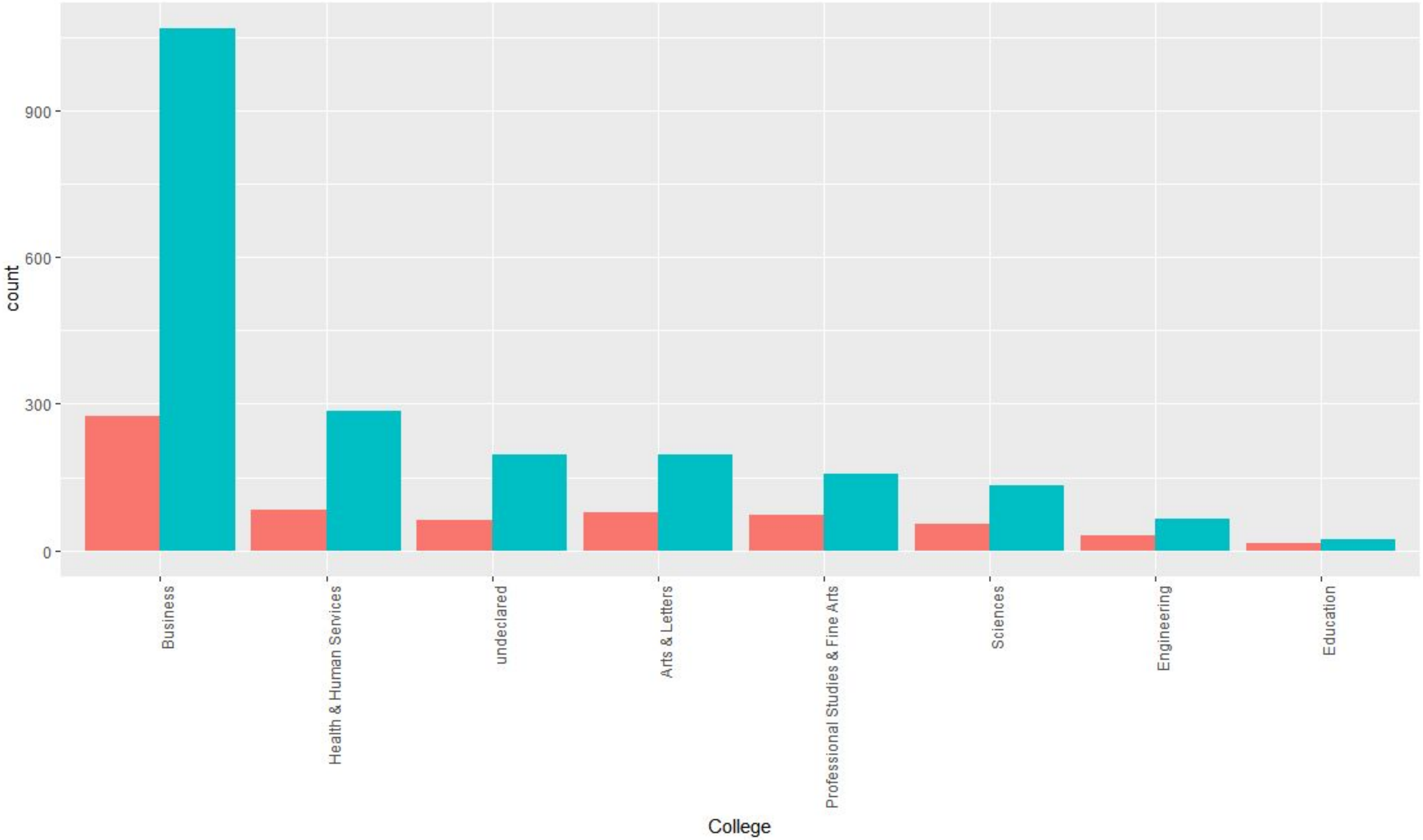


EDA: Year and Matriculation

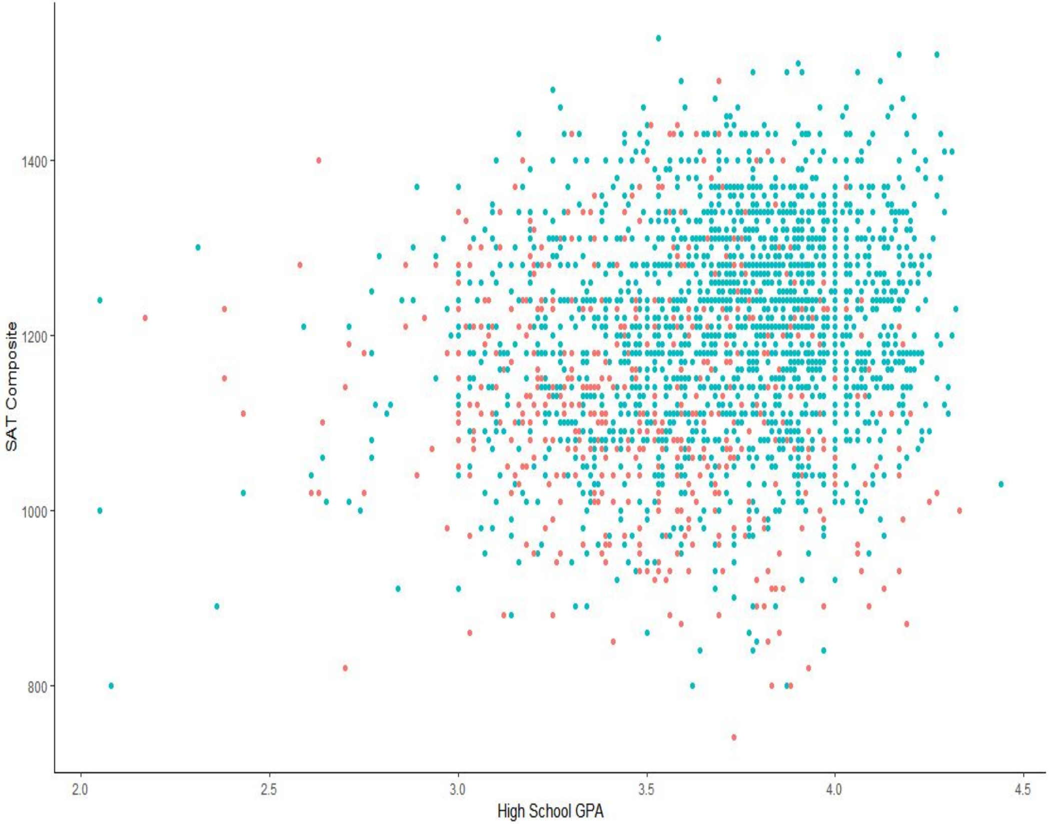




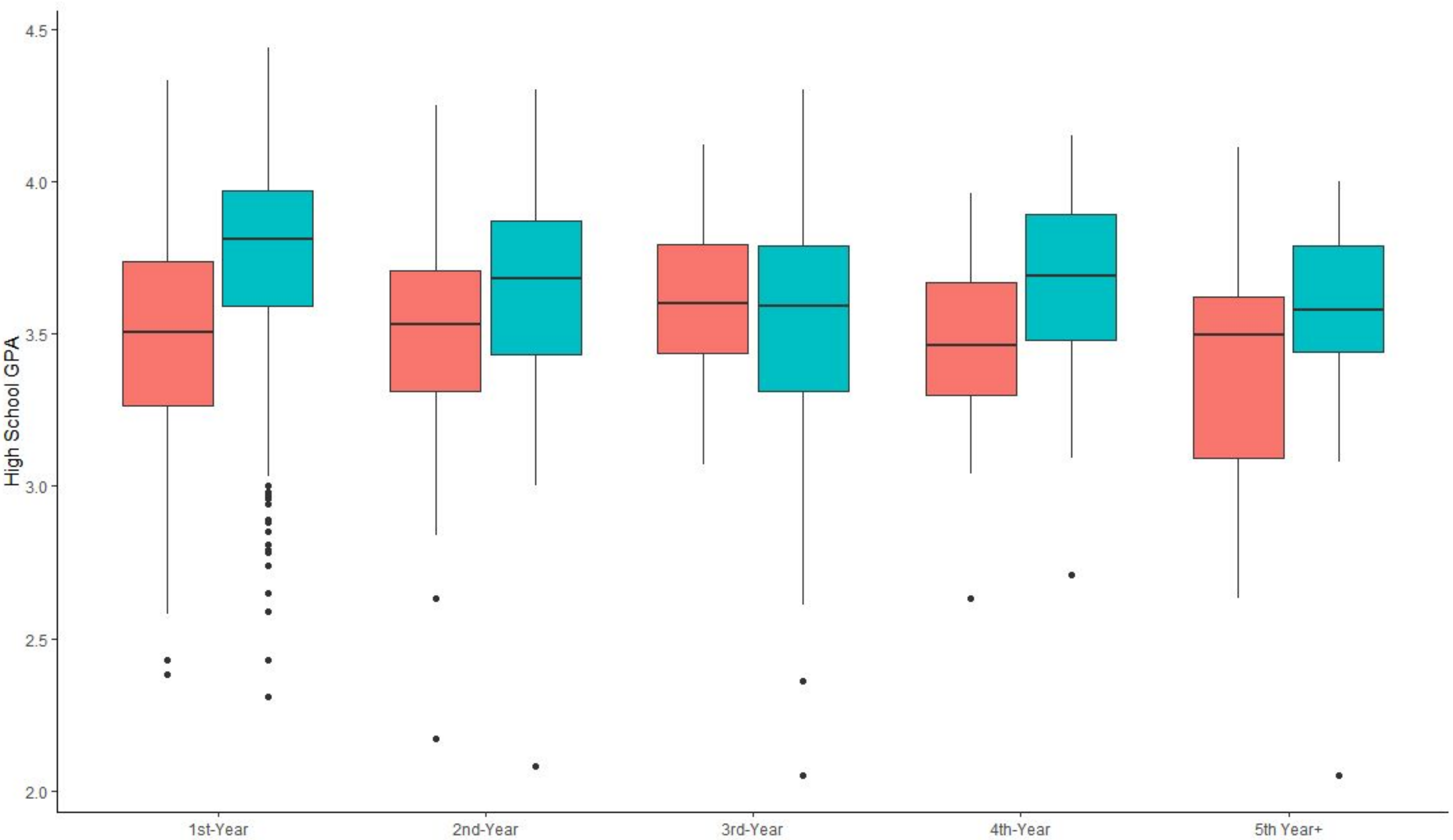
EDA: Distribution of Colleges



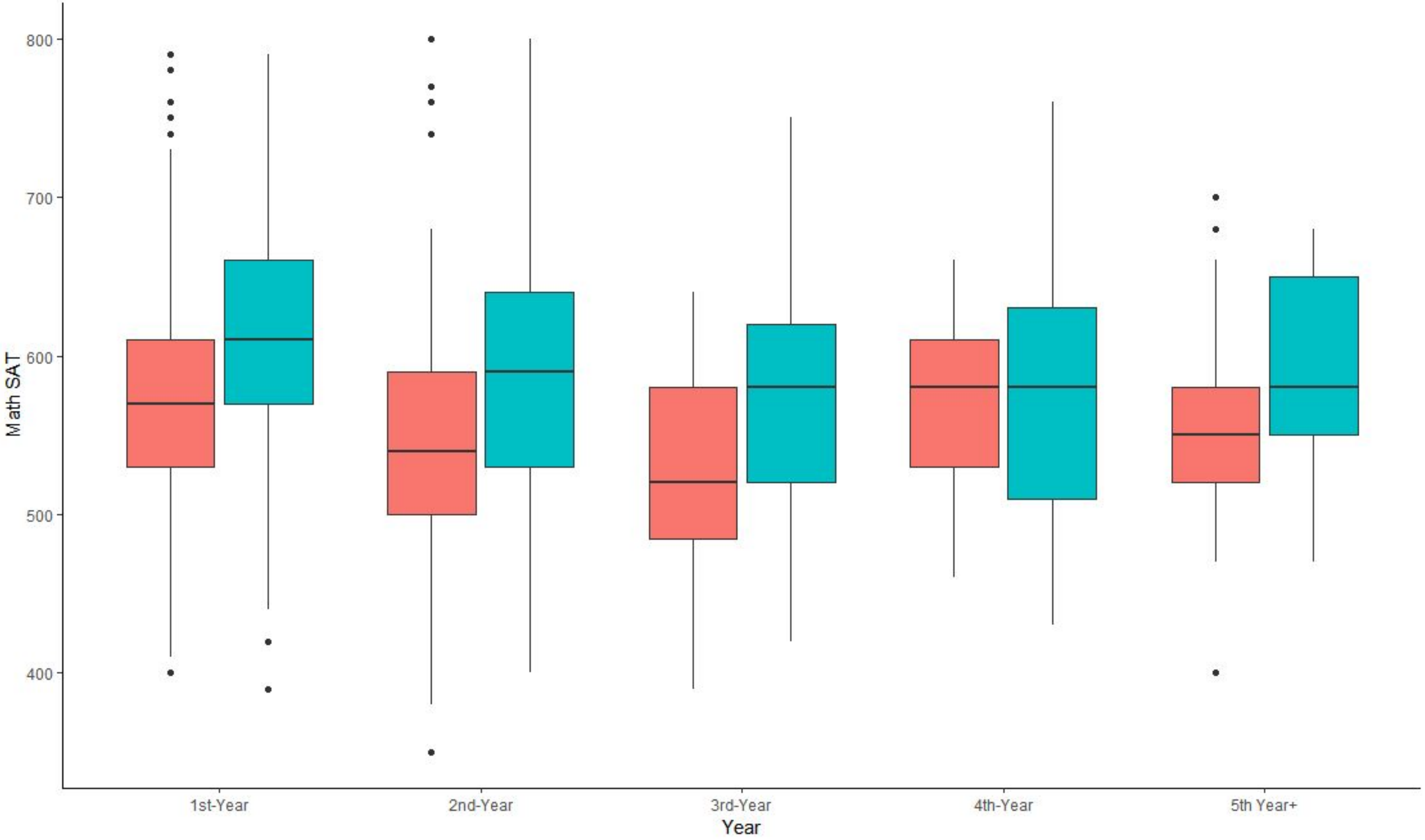
EDA: SAT Composite



EDA: High School GPA

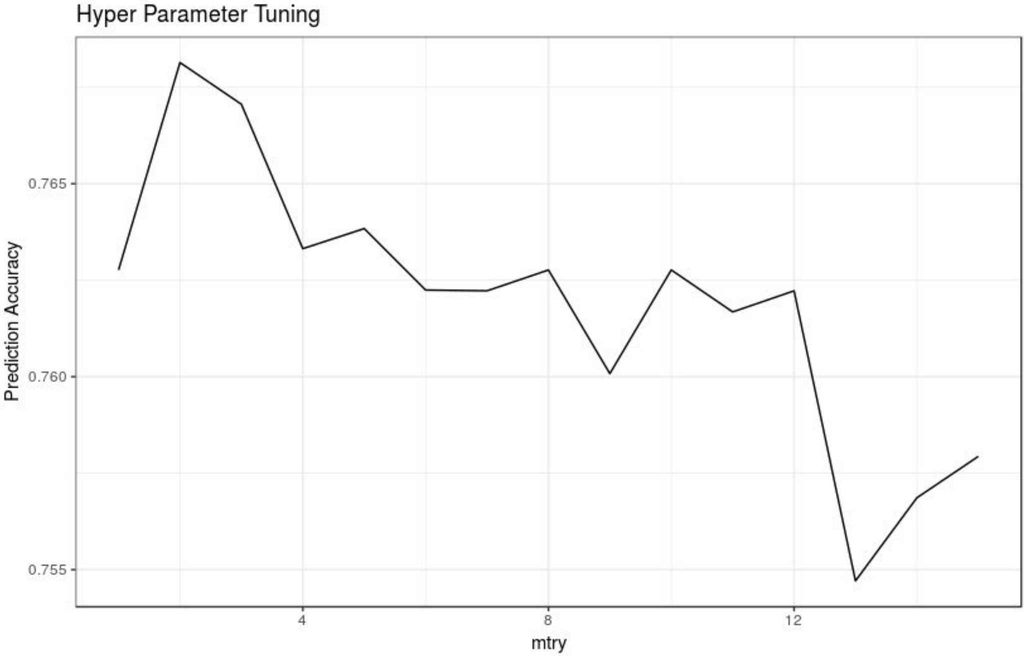


EDA: Math SAT

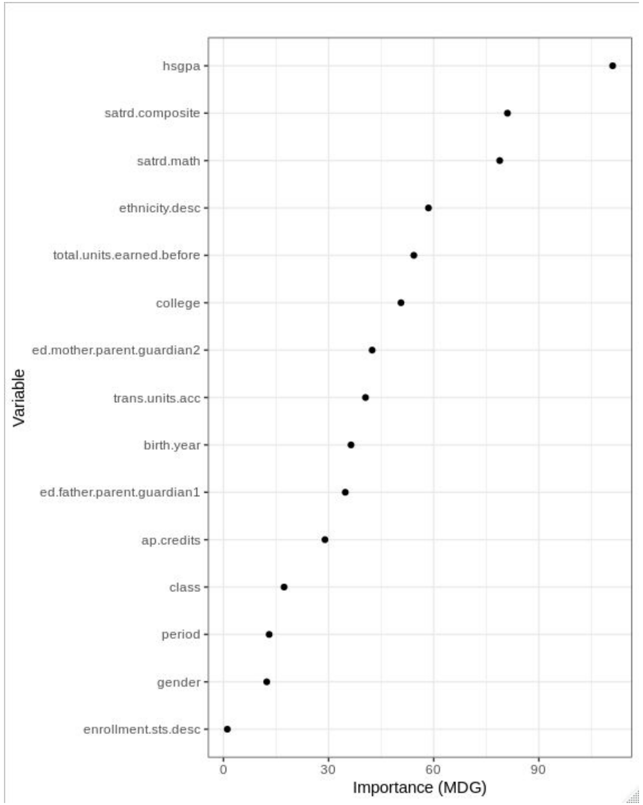

800
700
600
500
400
Math SAT
1st-Year
2nd-Year
3rd-Year
4th-Year
5th Year+
Year

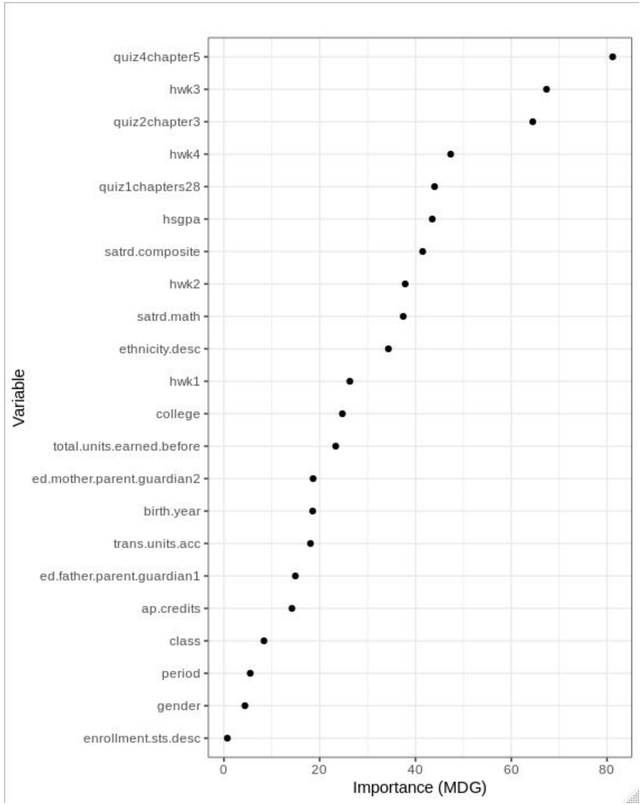
Random Forest

- Randomly split data into ⅔ training and ⅓ testing
- Performed tuning on training
- Predicted based on testing



Random Forest Importance

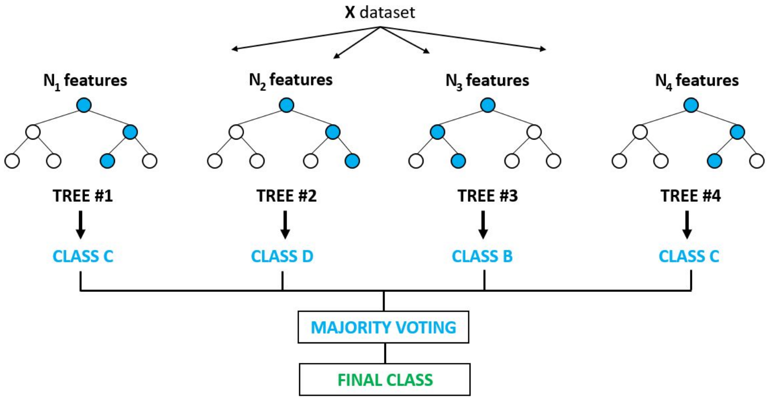


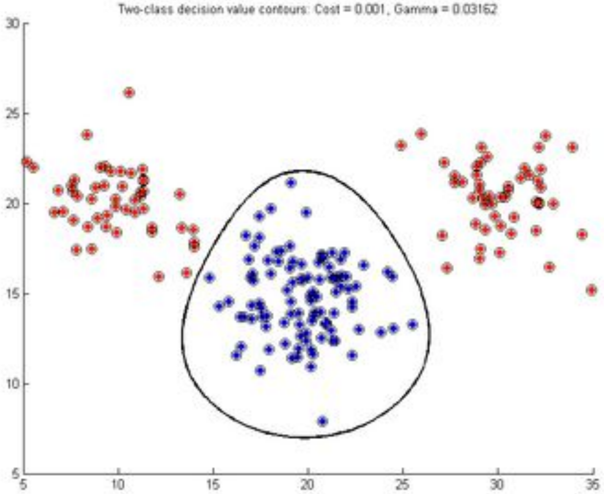


SVM

- Classification technique where hyperplanes are created to separate and classify the data in some feature space into different regions.
- Parameters:
  - Gamma: values of 0.1 and 0.01 are used for the before_data model and after_data model, respectively
  - Cost: value of 1 is used for both models
  - Kernel: Radial
- Number of support vectors for before_data model: 1107
- Number of support vectors for after_data model : 779

Results





Time Frame	Test Misclass. Rate
Before	22.5% (2 mtry)
~4 weeks	16.2% (4 mtry)

Time Frame	Test Misclass. Rate
Before	23.7% (γ=0.1, cost=1)
~4 weeks	16.9% (γ=0.01, cost=1)

Conclusion

- Technique choice inconclusive
  - RF more intuitive
  - Lots of support vectors; SVM could be overfitting
- Not a complete early alert system
  - Week-by-week approach would be better
- There’s a point in the semester where students can’t recover
- Incorporating variables summarizing student behavior could be stronger
- Want to get the highest accuracy as early as possible to establish an intervention in time