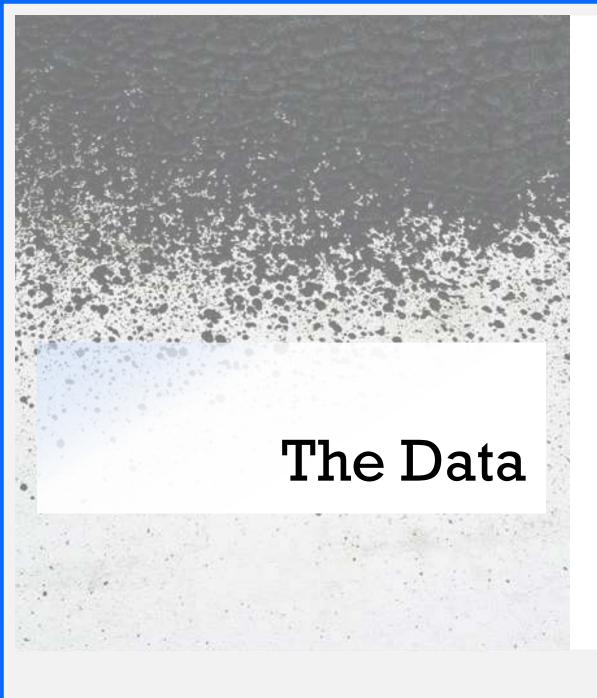
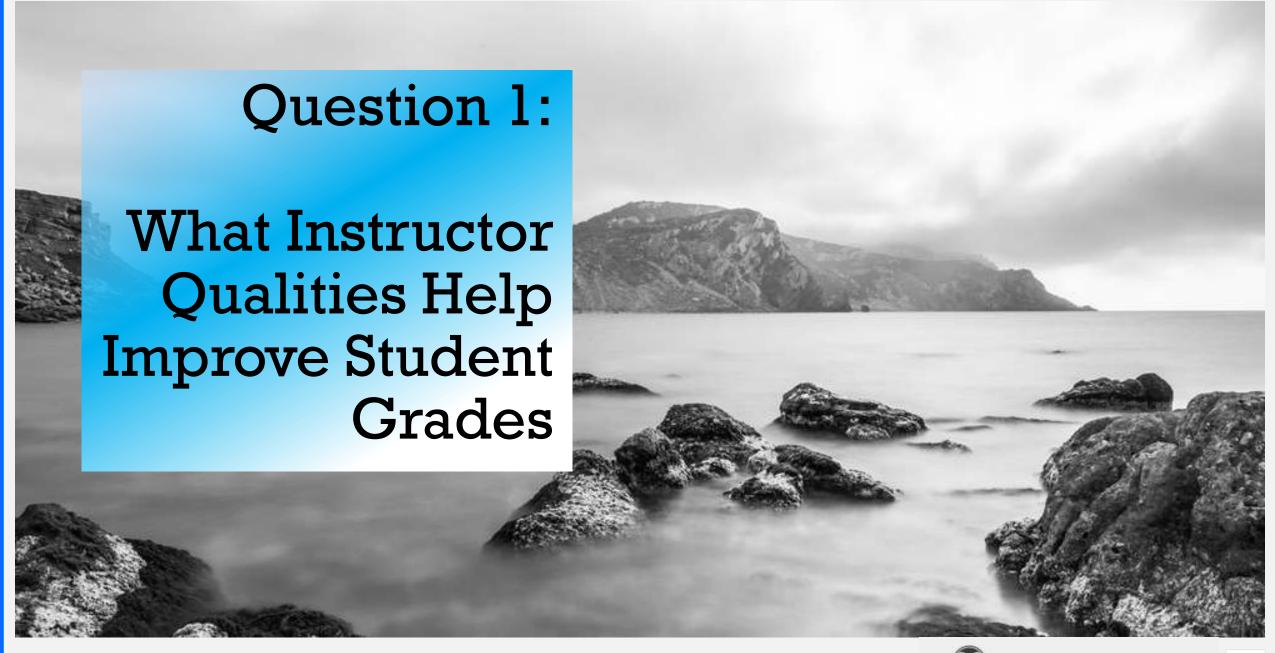
Stat 472 Final Presentation

Samantha Bothwell, Wulf Novak, and Crystal Wu



Calculus I data

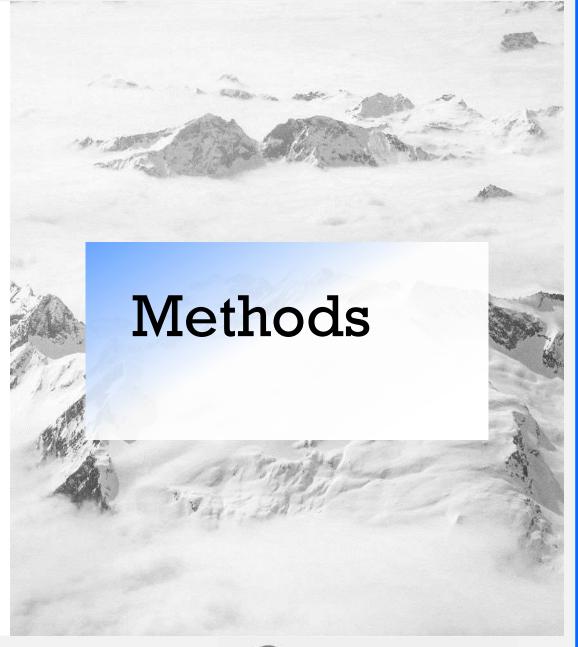
- Data for this project was collected at the beginning and end of the 2010 fall semester.
- Instructors, Students, and Departments were given surveys to complete
- The data could exhibit some bias since it was a volunteer survey – students with higher grades may have been more likely to complete the survey.



- Response Variable: Student reported Calc I grades on the post-term survey.
- Algorithm: The random forest algorithm will be implemented to determine the important variables in predicting grade.
 - Random Forest and AUCRF package
 - Builds multiple decision trees and merges them to create a prediction
 - Evaluated relationships between the student final Calc I grade and the various predictor variables

Data:

- Only variables relating to instructor quality were included.
- 55 predictor variables were used
- Any students with incomplete observations were taken out
- The resulting sample size was 4,271

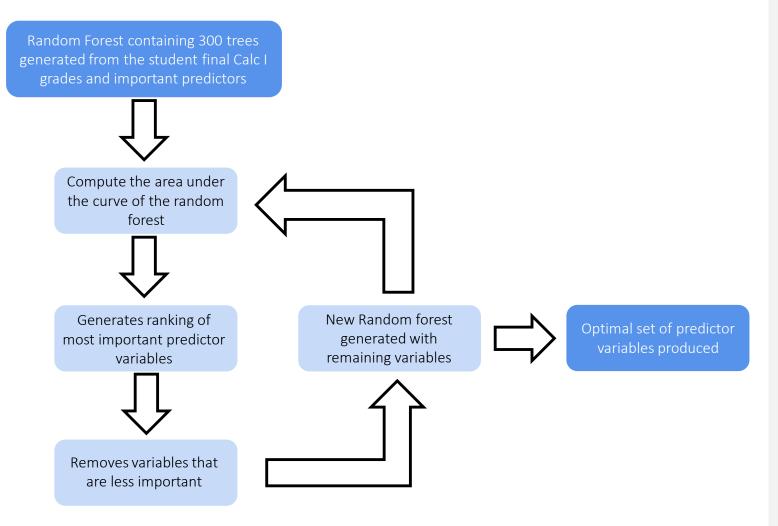




The Random Forest Algorithm

Advantages of Random Forest

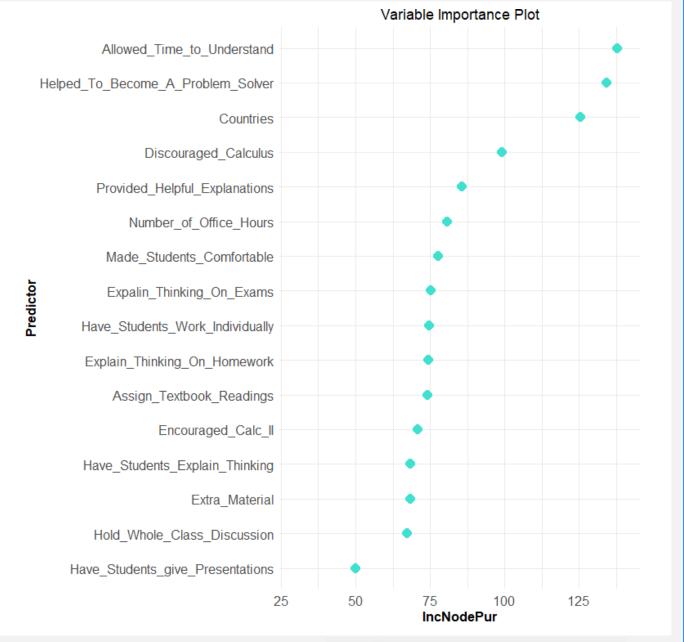
- High predictive performance
- Reliable for estimating important predictors
- It offers efficient estimates of the test error without the additional cost of repeated model training



Analysis of Variables

Top Variables

- 1. "My Calculus Instructor Helped Me Become A Better Problem Solver"
- 2. "My Calculus Instructor Allowed Time For Me To Understand Difficult Ideas"
- 3. Country of the Instructor's Undergraduate Degree
- 4. "My Calculus Instructor Discouraged Me From Wanting To Continue Taking Calculus"
- 5. "My Calculus Instructor Provided Explanations That Were Understandable"





Analysis of Variables (continued)

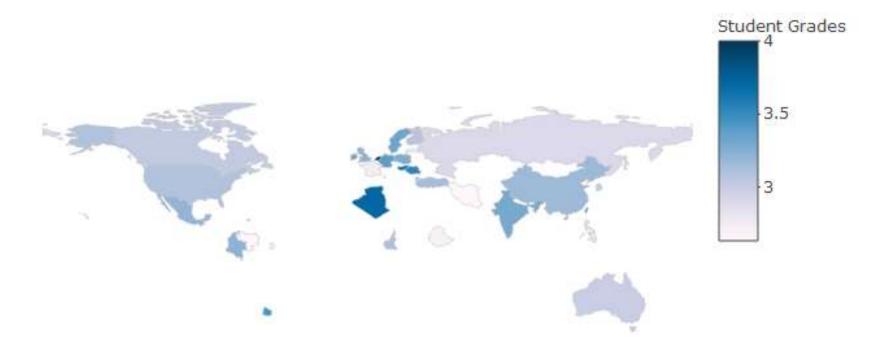


Instructors were rated for each question on a Likert scale from 1-6

- 1: Strongly Disagree
- **6**: Strongly Agree

Analysis of Variables (continued)

Student Grades based on Country of Instructors Undergraduate



file:///D:/School/Spring%202019/Stat%20472/Calculus%20Retention/FinalProject Pt1 Visual.html

Question 2:

What Qualities
Are Most
Important For
Predicting Good
Student Grades?



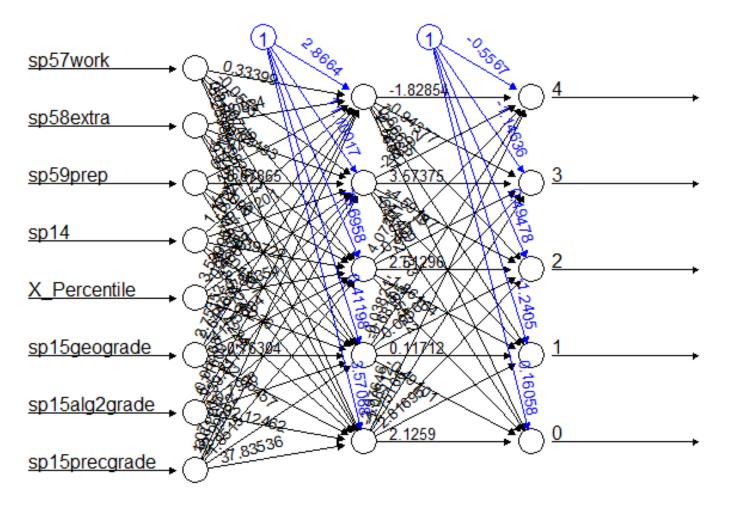
- Response Variable: End of semester student reported calculus 1 grade
- Algorithm: Neural Networks
 - Neural Networks are used in facial and image recognition, self-driving cars, and language translation
 - They are exceptional for prediction, often used to win prediction competitions
 - Complicated

Data:

- Pre-College Information Including: ACT/SAT Score, Algebra 2,
 Pre-calc, Geometry, High school Calc
- Perception or Belief in Ability: Confidence in math ability, confidence in knowledge, perception of difficulty
- Studying styles: Visited tutor, Hours spent studying, Met with students to study
- 30 Variables total
- The resulting sample size was 1,971 observations



The Neural Network Output



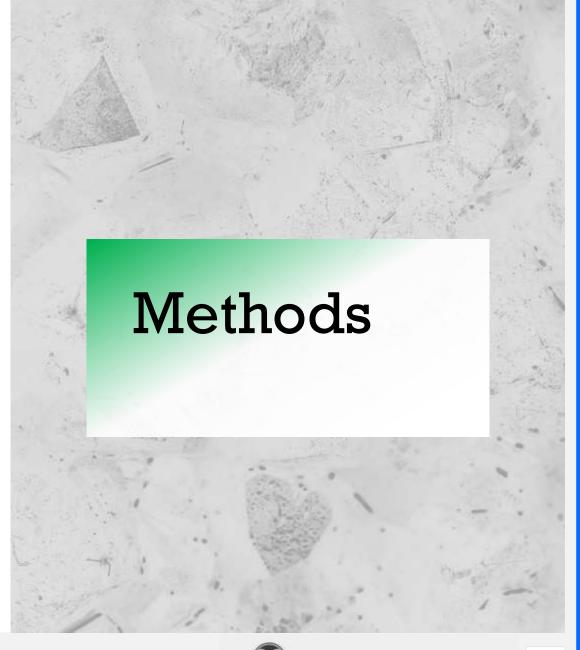
- Reduced number of variables in model to lower computation time
- Tried many combinations of layers, and number of neurons per layer
- Low interpretability, which isn't useful for determining the qualities of students who achieve high grades.

Error: 3025.147098 Steps: 65770

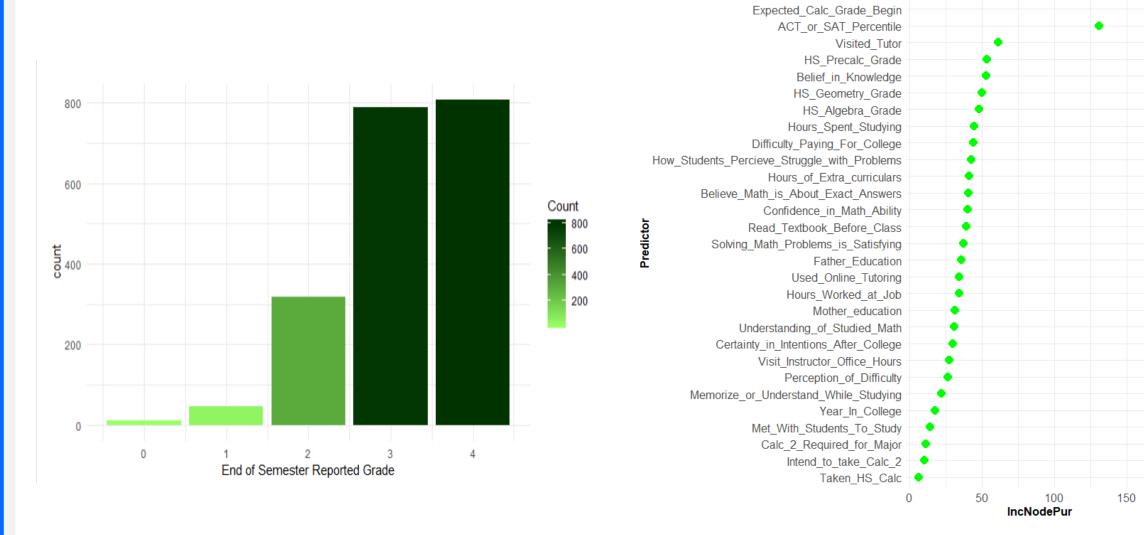
- Response Variable: End of semester student reported calculus 1 grade
- Algorithm: Neural Networks Random Forest ©
 - Neural Metworks are used in facial and image recognition/ self-driving cars, and language translation
 - They are exceptional for prediction, often used to wing prediction competitions
 - Complicated

Data:

- Pre-College Information Including: ACT/SAT Score, Algebra 2,
 Pre-calc, Geometry, High school Calc
- Perception or Belief in Ability: Confidence in math ability, confidence in knowledge, perception of difficulty
- Studying styles: Visited tutor, Hours spent studying, Met with students to study
- 30 Variables total
- The resulting sample size was 1,971 observations



Results





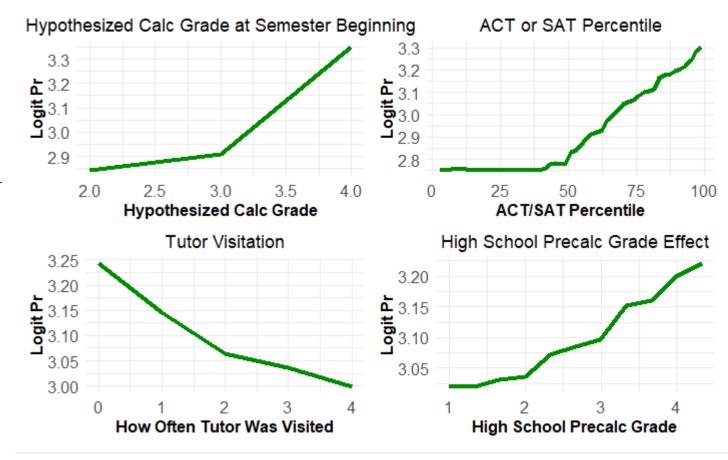
Qualities that Most Effect Student Grade

Analysis of Variables

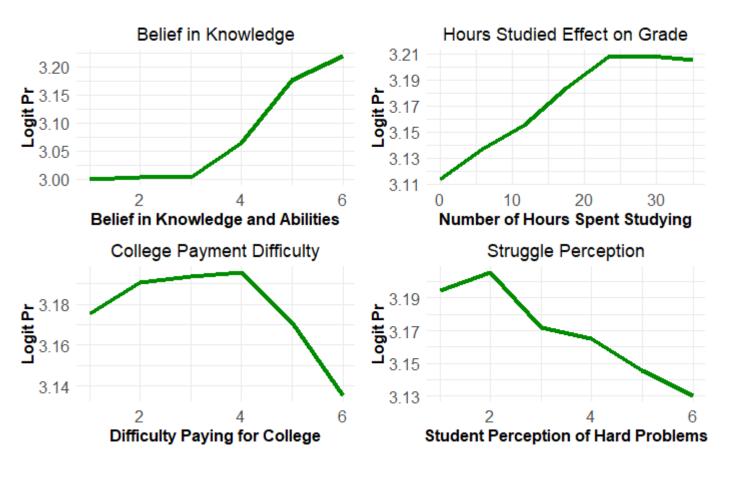
Top Variables

- 1. Those who predicted a 'B' or above for themselves were likely to perform better
- 2. Notably, doing better than average on the ACT/SAT improved grades. Also, the ACT/SAT percentile was left skewed, likely due to colleges selectively accepting students partly based on ACT/SAT scores.
- 3. 0 = Never, 4 = Ever class session.

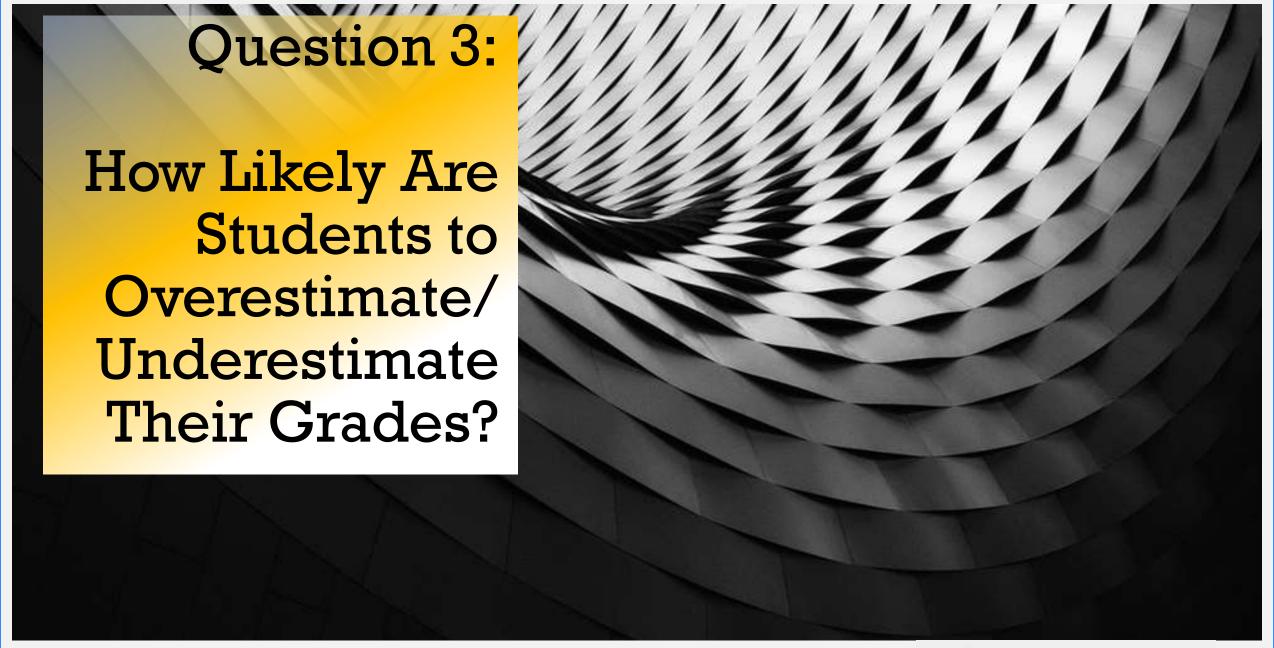
 Majority of students never visited, but most students obtained A's or B's.
- 4. High School precalc, algebra 2, and geometry all unsurprisingly were strong predictors for Calc 1 final grade.



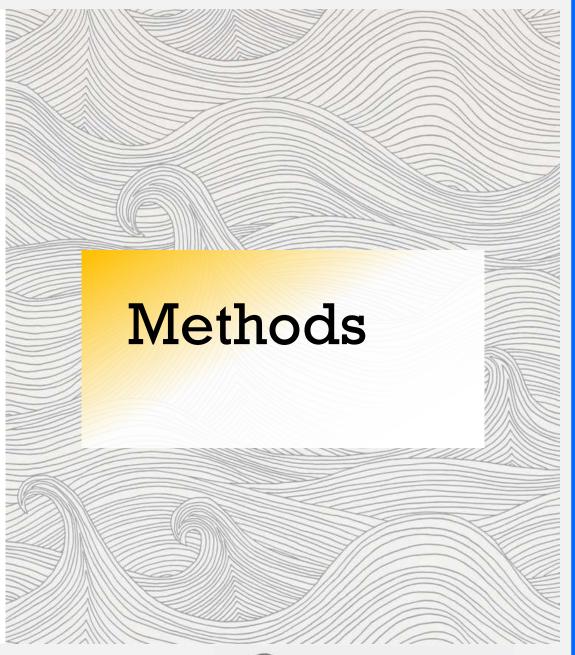
Analysis of Variables (continued)



- Belief in Knowledge and Ability: "I believe I have the knowledge and abilities to succeed in this course"
 - 1 = Strongly Disagree,
 - 6 = Strongly Agree
- Hours Studied: Weekly hours studied.
 Studying more than 20 hours for calculus didn't seem effective.
- Difficulty Paying for College: "I am anticipating difficulty paying for college"
 - 1 = Strongly Disagree,
 - 6 = Strongly Agree
- Struggle Perception: "When Experiencing a Difficulty in my math class I"
 - 1 = "Try to figure it out on my Own",
 - 6 = "Quickly seek help or give up trying"



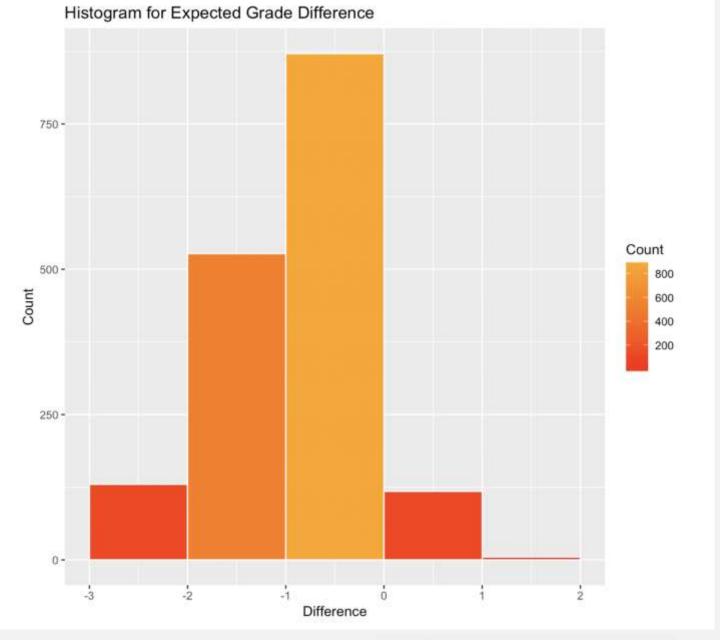
- Algorithm: Classification will be approximate because our response variable is categorical.
 - Binomial logistic regression model
 - Our output has two type of values (positive/negative)
 - Calculate the likelihood of students that could increase their expectation with those variables we select.
- Response Variable: The difference between the students' expected grade before the semester began and after the course end.
- Predict variables:
 - SAT/ACT Percentile
 - Previous high school mathematics experience
 - Home environment supporting level
 - Expected time for working
 - Expected time for doing extra activities
 - Expected time for studying for calculus I





Data Preparation

- Selected the data related to our variables in this question
- Removed all the missing values.
- The number of observations for the clean data was 781.



Logistic Regression Model

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * Percentile + \beta_2 * previous calculus experience + \beta_3 * home support + \beta_4 * expected work time + \beta_5 * expected extra activity tim + \beta_6 * expected study time$$

Expectation difference
$$(y) = \begin{cases} 1 & if student increase their expectation; \\ 0 & if students decrease their expectation; \end{cases}$$

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6)}}$$

Output of Logistic Regression Model

- Log (Odds) = 0.057 + 1.014*Percentile + 2.393*Previous calculus experience + 0.321*home support_1 + 0.325*home support_2 + 1.015*expected work time + 0.979*expected extra activity time + 1.019*expected study time
- Previous calculus experience (p = .049)
 & home environment support (p = .008)
 has significant effect on students'
 expectation of their grade
- From the confusion matrix

• Accuracy: .159

Sensitivity: .158

• Specificity: 1

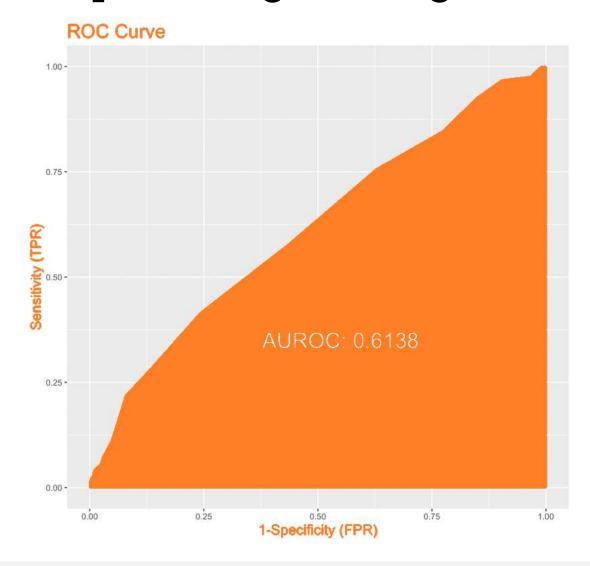
• Precision: 1

Confusion matrix

whether students will increase their expected grade or not

N = 781	Predicted: Decrease	Predicted: Increase
Actual: Decrease	1	0
Actual: Increase	657	123

Output of Logistic Regression Model



 From the ROC curve, the Area Under the Curve (AUC) is 61.38%; The rate of successful classification by the logistic model is 61.38%

Discussion

- Some ideas about logistic regression model:
 - Advantages:
 - Be easy to compute and interpret
 - Can be a baseline to measure the performance of other complex Algorithms
 - Drawbacks:
 - Not useful when solving the non-linear problems;
 - Logistic regression is not a useful tool unless you have already identified all the important independent variables
- Some issue with the data:
 - Did not have a nice distribution for the difference
 - Did not figure out all the important independent variables
 - Did not randomly split our data set into a training set and testing set
- We may not use the probabilities returned by the logistic regression model to prioritize student's expectation for their calculus course grades.

