Stat 472 Final Project

Samantha Bothwell, Wulf Novak, and Crystal Wu

May 10, 2019

Group Name: ÷ and Conquer

Abstract

During the 2010 fall semester, students and faculty were given surveys to fill out to describe themselves and the environment of Calculus I classes. We aim to answer the following questions relating to the data:

1. What Instructor Qualities Have the Largest Impact on Student Grade?
2. How Does Student Performance Differ Between States?
3. How Likely Are Students to Overestimate/Underestimate Their Grades?

Question 1: What Instructor Qualities Have the Largest Impact on Student Grade?

I Introduction

At the end of their Calculus I course, students were asked to evaluate their teacher quality based on a variety of characteristics. Additionally, instructors were asked questions related to their teaching style. To evaluate a "good" vs a "bad" teacher we used variables that pertained only to instructor quality. For this project we will evaluate the instructor qualities that separates an "F" student vs an "A" student. The response variable will be the student reported end of semester Calculus I grades.

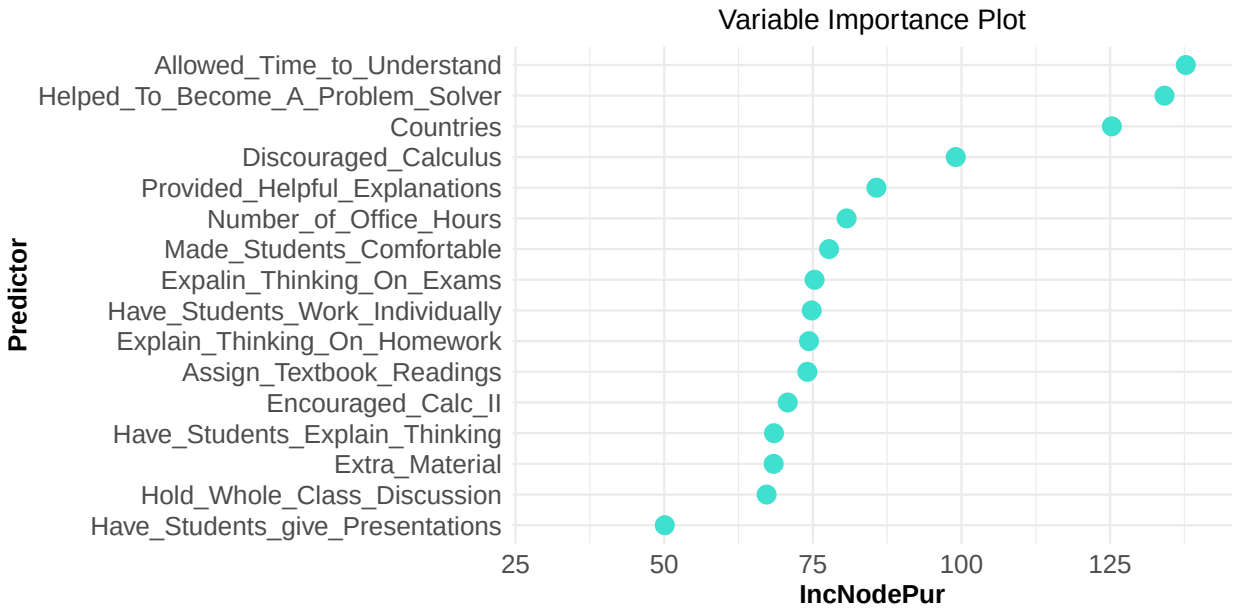
II Methods

To perform analysis we turned to the random forest algorithm. The benefits to this algorithm is its predictive power and that it accounts for overfitting. Where this algorithm falls short lies in its interpretability. We felt it an appropriate mode for analysis due to its ability to rank the important variables in predicting the response variable. Using only complete data, we resulted in 3261 observations.

III Analysis

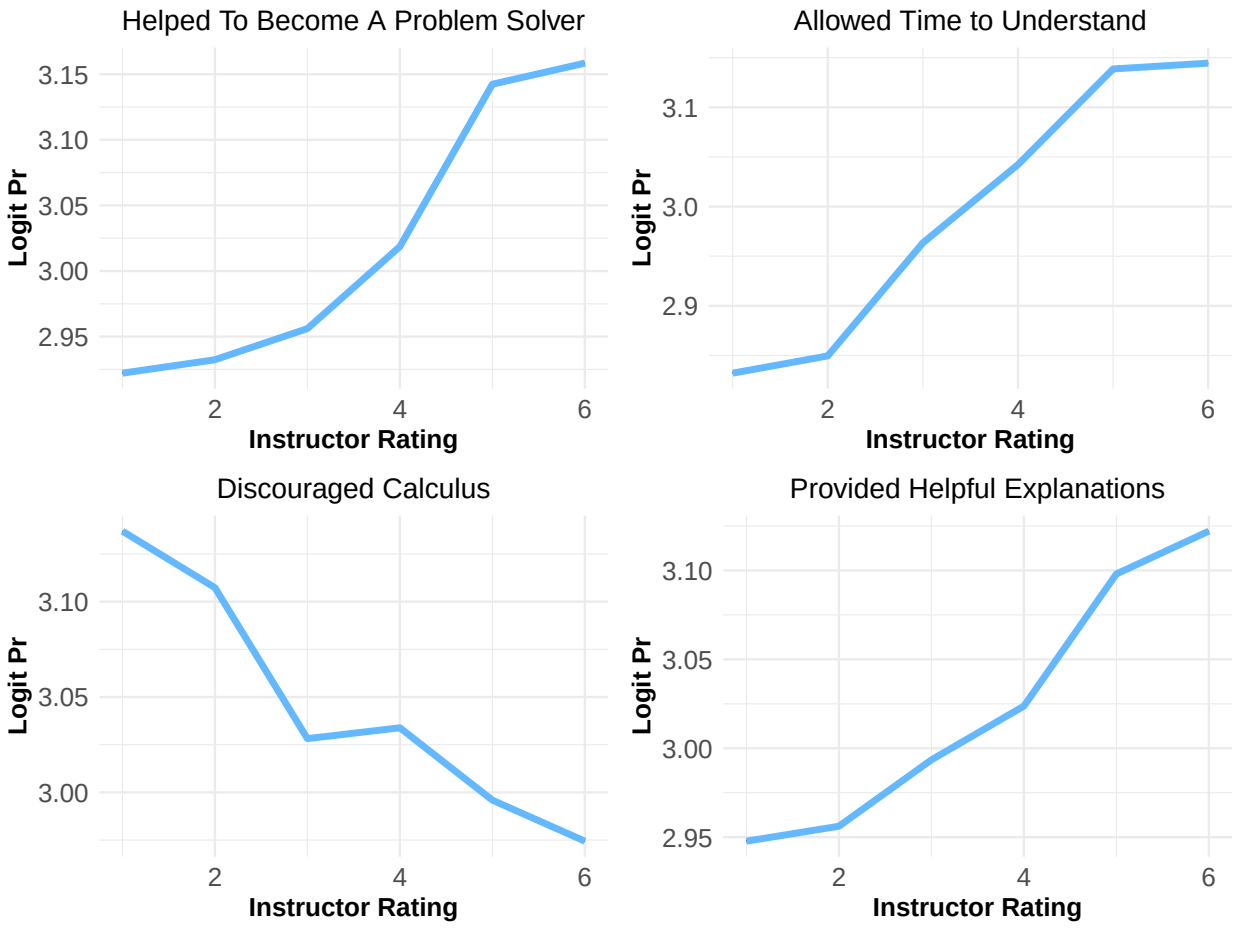
Using the random forest algorithm, we can visualize the variables that are most important in predicting the response with a variable importance plot, as seen in Figure 1.

Figure 1. Instructor Quality Variable Importance Plot



Looking at Figure 1 we can identify some of the top variables. The top variables that have high importance are "My calculus instructor helped me become a better problem solver", "My caluclus instructor allowed time for me to understand difficult ideas", "My calculus instructor provided explanations that were understandable", and "My calculus instructor discouraged me from wanting to continue taking calculus". We can further view the relationships between the student grades and these variables with partial plots under the `R` random forest package.

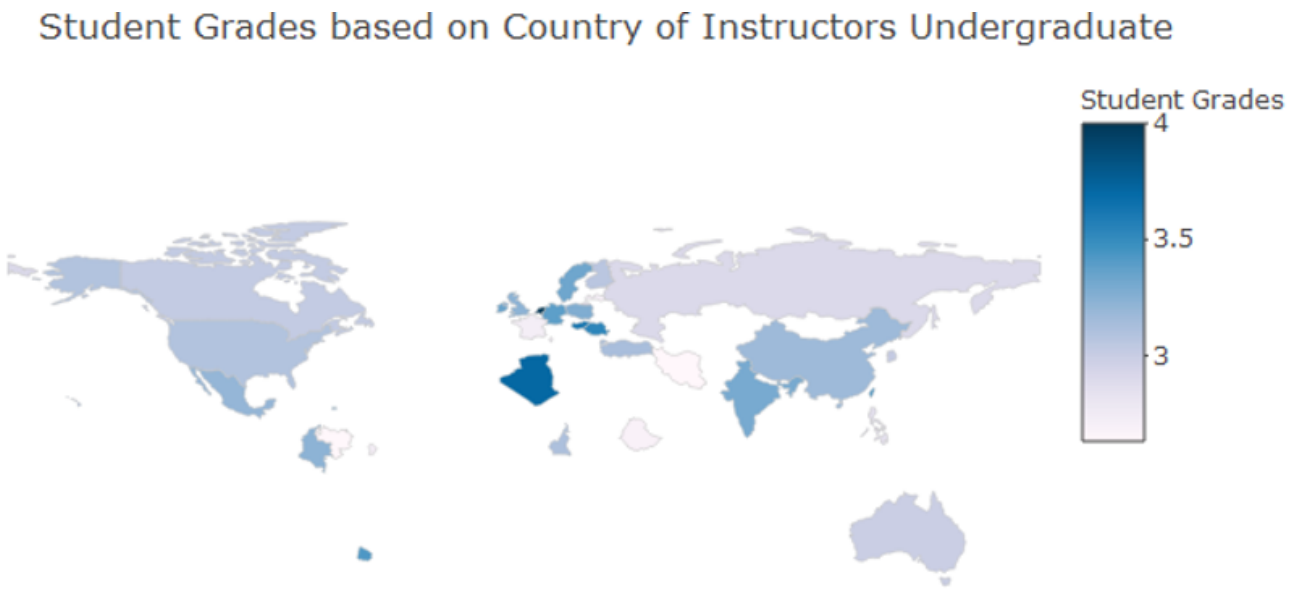
Figure 2. Top Variable Partial Plots



Each of the instrucor ratings is on a Likert scale where 1 is strongly disagree and 6 is strongly agree. Looking at the partial plots, each variable except "Discouraged Calculus" has a positive association with higher grades when the instructor rating is higher. For "Discouraged Calculus", the more an instructor discouraged students from taking Calculus II, the lower the students' grade.

The other quality of the instructor that was ranked highly was the country in which that instructor received their undergraduate degree. Figure 3 shows the relationship between final calculus grades for students by the instructor's home country (For an interactive version, please see the supporting figure.

Figure 3. Student Grades by Instructor's Country



IV Conclusion and Discussion

The random forest algorithm was used due to it's strong predictive power. It was able to provide us information about which variables were best in explaining our response variable.

We concluded that the most important instructor qualities to better student grades are:

1. "My calculus instructor helped me become a better problem solver"
2. "My caluclus 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"

Some of these variables could exhibit some bias. For example, a student with an A would probably be more likely to rate their instructor higher. Since we don't know the level of bias in these questions, we still believe the variables could help explain optimal qualities of the instructor.

Question 2: How Does Student Performance Differ Between States?

I Introduction

The Calculus dataset offers numerous variables to evaluate students. These include variables about the students scores in high school math classes, perception of mathematics, belief in their own ability, and even how many hours they would study weekly. We thought we would take advantage of this available data to gain an understanding of the qualities that lead to students achieving higher grades in their calculus class. The resulting analysis focuses on which variables make the most impact on a student and their end of calculus grade.

II Methods

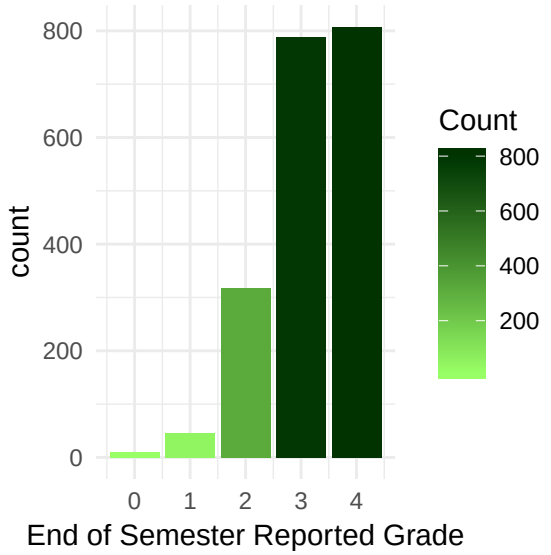
The initial chosen method to analyze the dataset was neural networks. It was intended to use the powerful predictive abilities of neural networks to predict the grade a student would recieve given selected variables. After thorough research, and a large amount of time spent attempting to optimize the neural network, neural networks proved to be an unideal method. The large amount of input variables (30) in addition to the multiple layers and nodes/neurons needed to produce an accurate, proved to have numerous downsides. Namely, the computation time would exceed 40 minutes - I decided to stop the network, feeling that given the many parameters I would need to tweak and optimize, it was not something I would be able to accomplish in the timeframe I had. I opted to use only 8 variables and fewer layers. This network would take 3 minutes if the model converged, but produce large errors and poor predictions. I attempted to find the optimal cost function, activation functions, and number of layers and nodes - but ultimately failed in creating a useful network.

In order to produce results for the project, I opted to use the random forest algorithm.

III Analysis

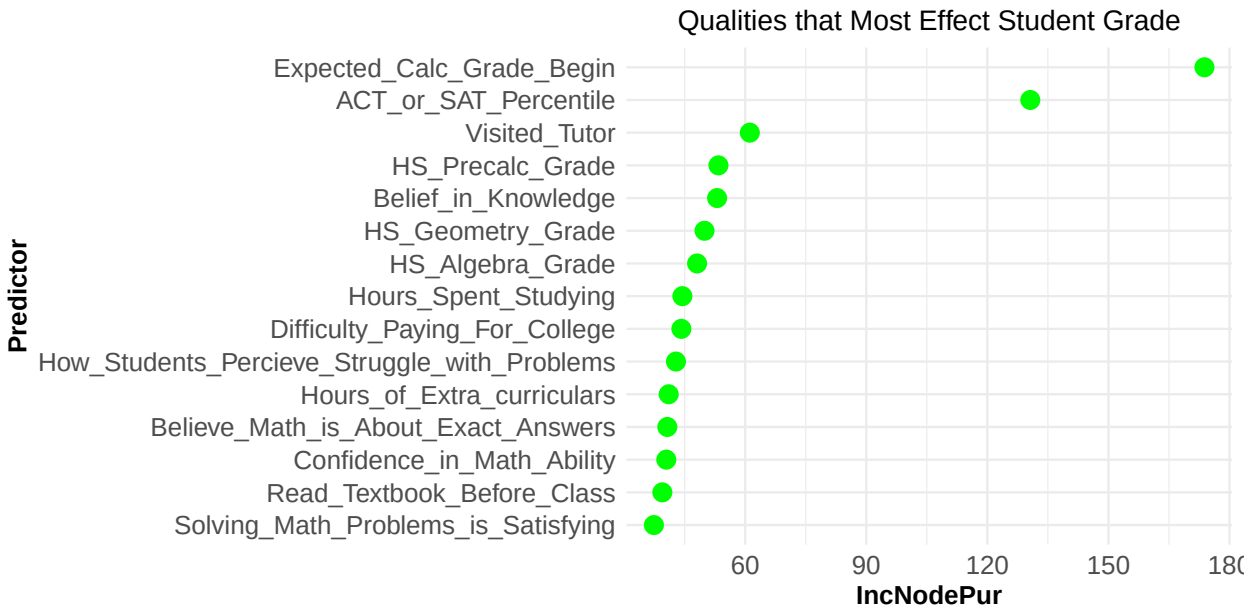
```
##           used (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  891033 47.6    1442291  77.1  1442291  77.1
## Vcells 7803153 59.6   24335780 185.7 30387325 231.9
```

Figure 4: End of Semester Grades



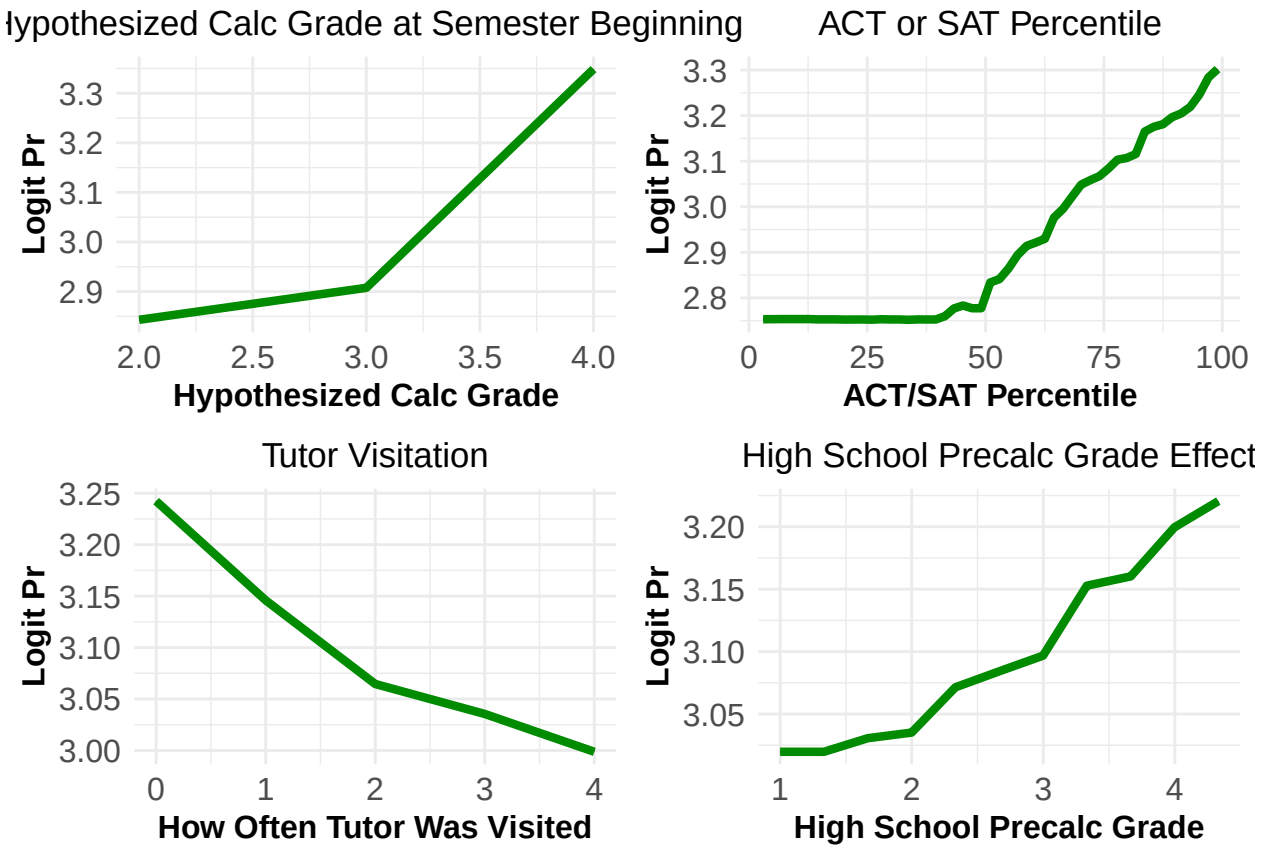
This plot features the number of students for each of the grades reported. Notably, the results are left skewed, with the majority of students receiving A's or B's. Since the dataset was limited to those who took the pre-semester survey AND the post-semester survey, it is possible that students who received a higher final grade were more likely to respond. Additionally, the end of semester grades were student reported, and were either the final grade or the expected final grade. This may have also lead to the bias in grades.

Figure 5: Variable Importance Plot



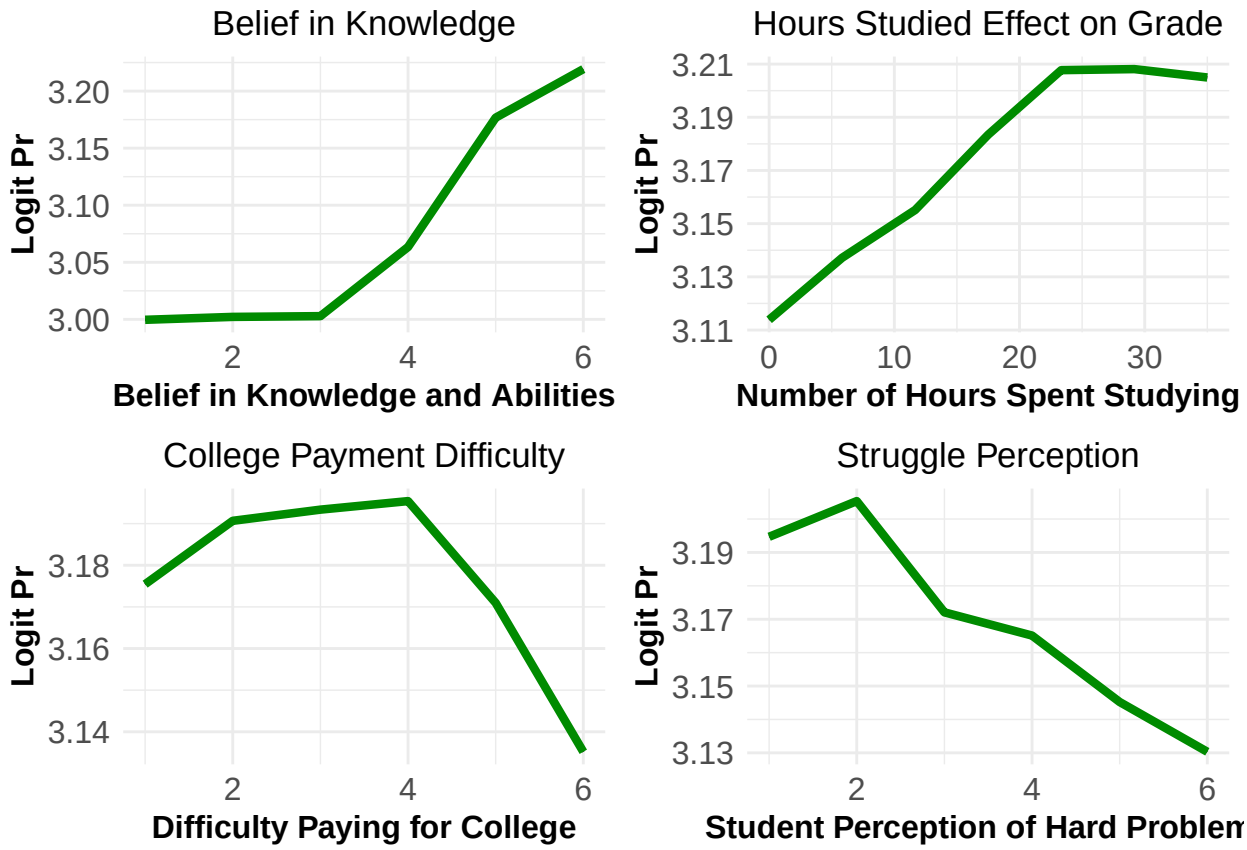
The above plot shows the top 15 variables that effected the students reported final calculus grade. The expected calc grade in the beginning of the semester most effected their final calc grade, which possibly accents the importance of student expectation when heading into the course. Numerous variables that would be determined before the student entered college calculus, such as High School Pre-calc grade, High School Geometry Grade, and ACT or SAT Percentile, also played a large role in the student's final calculus grade. Other variables that judged the confidence of the student and their perception of math problems or struggle were evident.

Figure 6: Partial Plots (Top Variables)



These plots illustrate 4 of some of the top variables. The plot for hypothesized calc grade shows how students who predicted they would recieve an A played a large role in producing students who did in fact receive an A. ACT/SAT Percentile is left skewed, likely due to the selectivity of colleges. How often a tutor was visited was based on a likert scale where 0 = never visited a tutor, and 4 = daily visits. The majority of students never visited a tutor, and the majority recieved A's and B's, so it's possible that tutor visitation's importance on student performance was skewed by these facts. Lastly, High School Pre-calc, in addition to the other High School grade vairables, had a similar effect on final calculus grades, where higher high school grades corresponded to a higher calculus grade.

Figure 7: Partial Plots (Other Interesting Variables)



The above chart shows many interesting qualities. The Belief in Knowledge plot is based on a likert scale from the following question. "I believe I have the knowledge and abilities to succeed in this course" Where 1 = Strongly disagree, and 6 = Strongly Agree. Again, the student's perception of their abilities played an important role in their final calc grade. Next, it is unsurprisingly shown that students who study more perform better - interestingly, the gain after studying 20 hours a week is marginal. The next plot is particularly insightful, as it features that effect of college payment difficulty plays on final calculus grade. 1 = strongly disgree with a difficulty for paying for college, and 6 = strongly agree. Students who responded as agree or strongly agree experienced lower final calc grades, which is unfortunate, considering that college is generally seen as a means for individuals to rise in socio-economic status, grades being effected by difficulty to pay for college is unideal. The last plot looks at the student responses to the following question: "When Experiencing a Difficulty in my math class I" where 1 = "Try to figure it out on my Own", and 6 = "Quickly seek help or give up trying". This plot indicates that students that indepently solve problems achieve higher grades than students who immediately seek help without putting in much effort.

IV Conclusion and Discussion

In conclusion, the qualities that most effect student's grade are high school grades in math classes and ACT/SAT percentile, their beliefs and confidence in achieving a higher grade, and also the amount of effort they put into studying - a difficulty in paying for college also played a role.

Question 3: How Likely Are Students to Overestimate/Underestimate Their Grades?

I Introduction

A lot of studies have shown that students' expectation of their grades would have an effect on their actual grades.Students who overestimating themselves may face grade disputes in their college life (Rooseveltf, 2009). They may complain about their grades and not work hard in the future. Finally, they would get failed in their college life. We do not want such a situation to happen anymore.

Therefore, our study is to figure out what would be the cause to make students underestimate themselves or overestimate themselves. We also want to use the result of this study to find some possible ways to activate the student's confidence level about taking a calculus course.

II Methods

The response variable in this question is the difference between student's expected calculus I grade and their expected grade after the calculus courses end. One thing we want to figure out is how likely college students would overestimate/underestimate their grade.

We used the logistic regression model for this question. The classification method is for the data set which is already separated into several groups, which is to make a prediction or take decisions based on the data set we have. Because our response variable is one categorical variable, we decided to use the binary logistic regression model.

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 * Percentile + \beta_2 * previous\ calculus\ experience + \beta_3 * choice + \beta_4 * home\ suppport$$

$$+\beta_5 * worktime + \beta_6 * extra\ activity\ time + \beta_7 * math\ preparaion\ time + \beta_8 * difficulty\ of\ paying\ college$$

$$Expectation = \begin{cases} 1, & \text{if students increase their expectation on grade} \\ 0, & \text{if students increase their expectation on grade} \end{cases}$$

The predicted variable we consider in this model is that the previous SAT/ACT grade; previous high school mathematics experience; home supporting level; expected time for working; expected time for doing

extra activities; expected time for studying for Calculus I.

We selected the data that related to our predict variables. Then we removed all the missing value and decided to not include the data that students have the same expectation before and after the class. We got 781 observations in our data set. The expectation increasing group includes 123 people and for the decreasing group, we have 658 people.

III Analysis

We used the binary logistic regression model to analyze the relationship between the students' expected grade and SAT/ACT grade, previous high school experience, home environment supportive level, expect spending time on work, extra activities, and study. From the output of the logistic regression, we found that two of the variables are significant when predicting the expectation difference on students' grade: previous high school experience and home environment supportive level. As shown in Fig 8, more high school calculus experience and higher supportive level of home environment would increase the likelihood that students increasing expectation on their final grade.

Figure 8: Odds of Switching

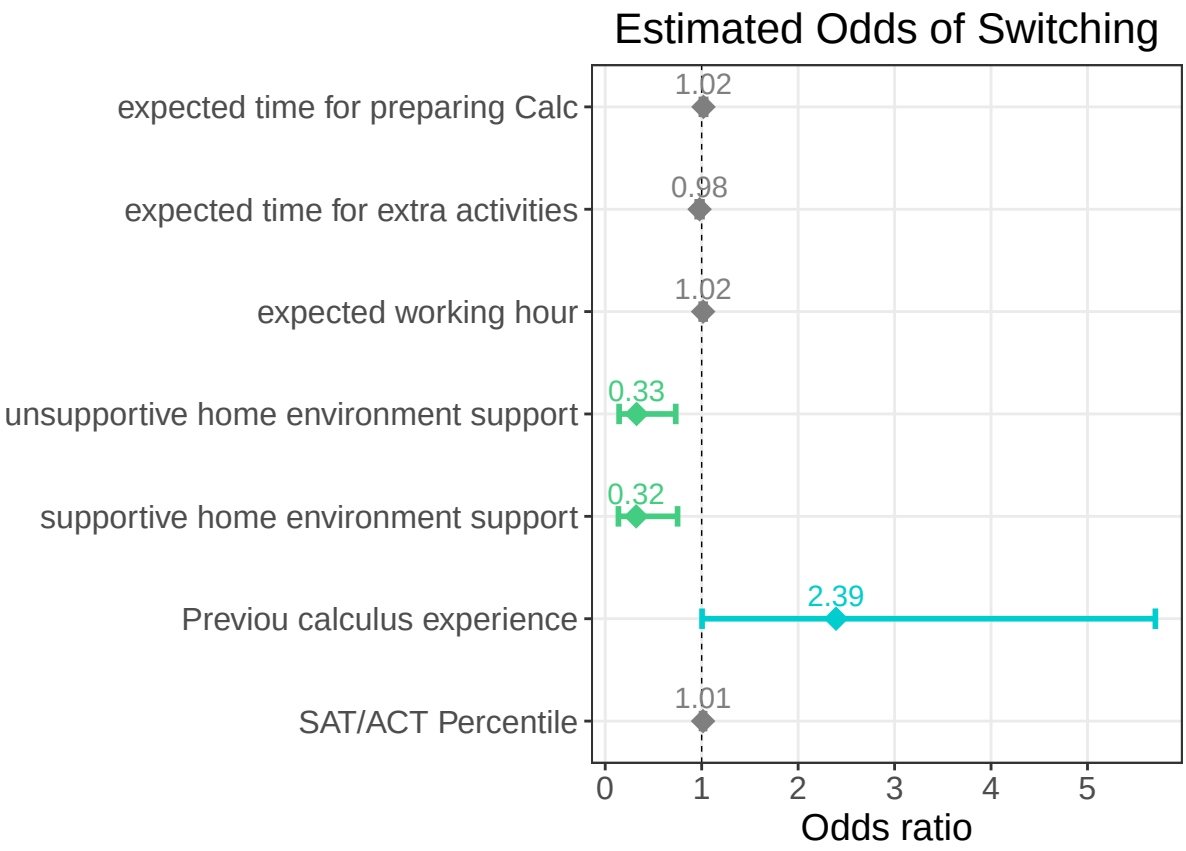
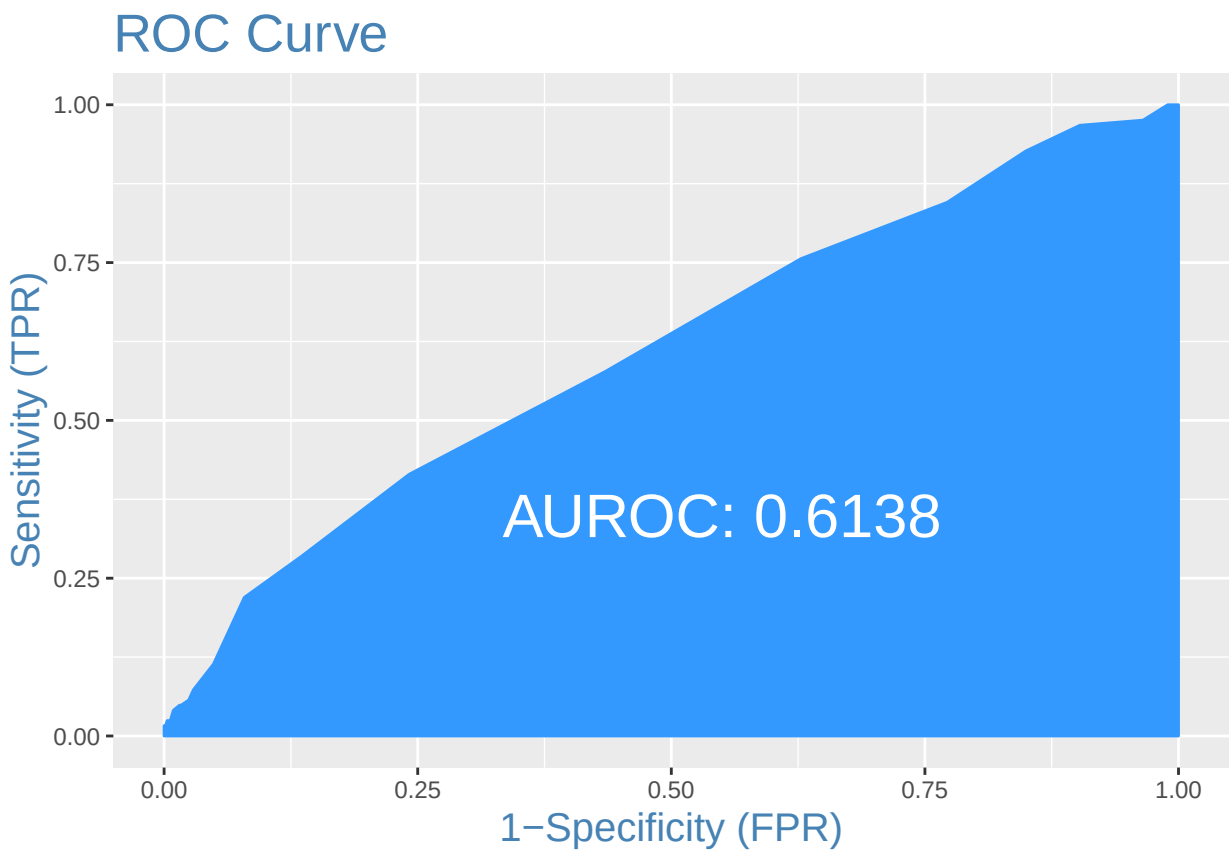


Table 1: Confusion Matrix

	Predict: False	Predict: True
Actual: False	1	0
Actual: True	657	123

However, from the confusion matrix (Table 1), we can see that the true positive rate is low (TN=.158) and the false negative rate is high. Based on the ROC curve graph (Fig. 9), we can see the area under the curve is .6183. The rate of successful classification by the logistic model is 61.38%.

Figure 9: ROC Curve



IV Conclusion and Discussion

We chose the binomial logistic regression model because we believe the regression is easy to compute and interpret. The output from the regression model can be used as a baseline to measure the performance of some other complex Algorithms. However, from the logistic regression model checking result, we found that the rate of successful classification is not as high as we expected, and the accuracy of the model is low. We

consider that the reason for a poor logistic regression model is that we did not identify all the important independent variables in the mode.

Another thing we should consider in this data is that we do not have two nice scaled groups. The sample size of overestimate group is much larger than that of the underestimate group. Because our data are collected from the student survey and it is a volunteer job, most of the complete case we find in our data set are students who got an A/B grade in the calculus course. Students may have bias on their expected grades. One more thing we can do in the future is to do a randomly split on our data set and separate the data into a training set and testing set to avoid the sample.

The model can accurately identify patients receiving low quality care with test set accuracy being equal to 15.9% which is not good our baseline model, and the successful classification rate for this model is 61.38%.

We may not want to use the probabilities returned by the logistic regression model to prioritize student's expectation for their calculus course grades.

Acknowledgements

We would like to thank Dr. Aaron Nielson, Youngseok Song, and Nahali Mhatre for their recommendations and assistance throughout the project. We would also like to thank professors, grad studnets, and instructors in the CSU Statistics department who have been pivotal to our understanding of statistical analysis. Among these are Dr. Daniel Cooley, Dr. Bailey Fosdick, Dr. Ander Wilson, Dr. Jay Breidt, Dr. Zachary Weller, Ben Prytherch, Charlie Vollmer, Joshua Hewitt, Michael Creutzinger MS. and David Clancy MS.. This project was adapted from the code provided by Dr. Bailey Fosdick and Dr. Jessica Ellis-Hagman and based on their original paper "Women 1.5 Times More Likely to Leave STEM Pipeline after Calculus Compared to Men: Lack of Mathematical Confidence a Potential Culprit"

References

[1] Santra, A. K., Christy, C. J. "Genetic algorithm and confusion matrix for document clustering" International Journal of Computer Science Issues (IJCSI), 9(1), 322, 2012

[2] Hosmer Jr, D. W., Lemeshow, S., Sturdivant, R. X. "Applied logistic regression" John Wiley and Sons., 2013

[3] Pandey, P. "A Guide to Machine Learning in R for Beginners: Logistic Regression" A Medium Corporation

[4] Grogan, Michael "Decision Trees and Random Forests in R" DataScience+

Code Appendix

Question 1 code

```
dat <- read.csv("D:/School/Spring 2019/Stat 472/Calculus Retention/maalongdatafile_ANON.csv")



# New data frame
dat1 <- dat[,c(1,10,16,52,61,113,157,184:217,229:232,268,277,278,280,301,302,304,310,
               312:318,325,329,330,331,349,351,353,422:427,327)]


# Clean data
cleandat <- dat1[!(is.na(dat1$sqr25grade)) & !(is.na(dat1$ip33countryunder)) & !(dat1$sq18ask==""),]
```

```
countries <- rep(NA,nrow(cleandat))
countries[cleandat$ip33countryunder == 1] <- "USA"
countries[cleandat$ip33countryunder == 15] <- "AUS"
countries[cleandat$ip33countryunder == 38] <- "CAN"
countries[cleandat$ip33countryunder == 46] <- "CMR"
countries[cleandat$ip33countryunder == 47] <- "CHN"
countries[cleandat$ip33countryunder == 48] <- "COL"
countries[cleandat$ip33countryunder%in%c(56,57)] <- "DEU"
countries[cleandat$ip33countryunder == 62] <- "DZA"
countries[cleandat$ip33countryunder == 69] <- "ETH"
countries[cleandat$ip33countryunder == 70] <- "FIN"
countries[cleandat$ip33countryunder == 75] <- "FRA"
countries[cleandat$ip33countryunder == 78] <- "GBR"
countries[cleandat$ip33countryunder == 100] <- "HUN"
countries[cleandat$ip33countryunder == 102] <- "IRL"
countries[cleandat$ip33countryunder == 104] <- "IND"
countries[cleandat$ip33countryunder == 107] <- "IRN"
countries[cleandat$ip33countryunder == 109] <- "JOR"
countries[cleandat$ip33countryunder == 120] <- "KOR"
```

```
countries[cleandat$ip33countryunder == 133] <- "LVA"
countries[cleandat$ip33countryunder == 152] <- "MEX"
countries[cleandat$ip33countryunder == 161] <- "NLD"
countries[cleandat$ip33countryunder == 173] <- "PHL"
countries[cleandat$ip33countryunder == 175] <- "POL"
countries[cleandat$ip33countryunder == 178] <- "PRI"
countries[cleandat$ip33countryunder == 184] <- "ROU"
countries[cleandat$ip33countryunder == 185] <- "RUS"
countries[cleandat$ip33countryunder == 191] <- "SWE"
countries[cleandat$ip33countryunder == 218] <- "TUR"
countries[cleandat$ip33countryunder == 221] <- "TWN"
countries[cleandat$ip33countryunder == 226] <- "URY"
countries[cleandat$ip33countryunder == 230] <- "VEN"

cleandat <- cbind(cleandat, countries)
cleandat$countries <- as.factor(cleandat$countries)
```

```
# Change SAT/ACT scores to percentile
SATscore <- rev(seq(200,800,by=10)) #61 numbers
SATper <- c(99,99,99,98,97,97,96,95,95,94,93,91,90,88,87,85,83,82,79,77,75,
            73,70,67,64,62,59,55,52,49,45,42,40,36,33,30,27,24,21,19,16,14,
            12,10,9,7,6,5,4,3,3,2,2,1,1,1,1,1,0,0,0)

SATPerc = SATper[match(cleandat$sp3satmathscore,SATscore)]

ACTscore <- 36:1
ACTper <- c(99,99,99,99,98,96,95,92,90,87,83,79,74,68,62,56,50,43,36,30,24,
            18,12,7,4,1,1,1,1,1,1,1,1,1,1,1)

ACTPerc = ACTper[match(cleandat$sp7actmathscore, ACTscore)]

Perc <- rowMeans(cbind(SATPerc,ACTPerc), na.rm = T )
cleandat <- cbind(cleandat, Perc)
```

```r
cleandat <- cleandat[,-c(2,3,40:42,58)]
```

```r
i <- sapply(cleandat, is.factor)
cleandat[i] <- lapply(cleandat[i], as.character)


# Change variables into a number for pca
for(i in 1:dim(cleandat)[1]){
  for(j in 1:dim(cleandat)[2]){
    cleandat[i,j] <- ifelse(cleandat[i,j]=="Strongly agree", 6,
                      ifelse(cleandat[i,j]=="Agree", 5,
                      ifelse(cleandat[i,j]=="Slightly agree", 4,
                      ifelse(cleandat[i,j]=="Slightly disagree", 3,
                      ifelse(cleandat[i,j]=="Disagree", 2,
                      ifelse(cleandat[i,j]=="Strongly disagree", 1,
                      ifelse(cleandat[i,j]=="Not at all", 0,
                      ifelse(cleandat[i,j]=="Very often", 5,
                      ifelse(cleandat[i,j]=="help students learn to reason through
                             problems on their own", 0,
                      ifelse(cleandat[i,j]=="work problems so students know how to do them",3,
                      ifelse(cleandat[i,j]=="answered to question if no one responded quickly", 3,
                      ifelse(cleandat[i,j]=="waited for a student to answer", 0,
                      ifelse(cleandat[i,j]=="helped me figure out how to solve the problem", 3,
                      ifelse(cleandat[i,j]=="solved the problem for me", 0,
                      ifelse(cleandat[i,j]=="never", 0,
                      ifelse(cleandat[i,j]=="some class sessions", 1,
                      ifelse(cleandat[i,j]=="about half the class sessions", 2,
                      ifelse(cleandat[i,j]=="most class sessions", 3,
                      ifelse(cleandat[i,j]=="every class session", 4,
                      ifelse(cleandat[i,j]=="Not selected", 0,
                      ifelse(cleandat[i,j]=="Yes", 1,
                             cleandat[i,j])))))))))))))))))))))
  }
}
```

```r
cleandat[2:68] <- sapply(cleandat[2:68],as.integer)
```

```r
cleandat <- cleandat[-c(1:4,57,41,43,70,62:67)]
```

```r
cleandat <- cleandat[complete.cases(cleandat),]
```

```r
# Rename column
colnames(cleandat) <- c("Final Calc I Grade", "Waits_For_Answers","Solved_Problems_For_Students",
                        "Instr_Asked_Questions", "Listened_To_Questions",
                        "Discussed_Calculus_Applications",
                        "Allowed_Time_to_Understand", "Helped_To_Become_A_Problem_Solver",
                        "Provided_Helpful_Explanations", "Available_For_Appointments",
                        "Discouraged_Calculus", "Showed_How_To_Work_Problems",
                        "Have_Students_Work_Together",
                        "Hold_Whole_Class_Discussion", "Have_Students_give_Presentations",
                        "Have_Students_Work_Individually", "Lecture", "I_Ask_Questions",
                        "Have_Students_Explain_Thinking", "Extra_Material",
                        "Explain_Thinking_On_Homework",
                        "Expalin_Thinking_On_Exams", "Assign_Textbook_Readings",
                        "Made_Students_Nervous",
                        "Encouraged_Calc_II", "Treated_Students_as_Capable",
                        "Made_Students_Comfortable",
                        "Encourage_Office_Hours", "Presented_Multiple_Methods",
                        "Didnt_Speak_English_Well",
                        "Made_Class_Interesting", "Assigned_Homework", "Collected_Homework",
                        "Use_Technology_to_Find_Answers", "Use_Technology_to_Check_Answers",
                        "Use_Technology_For_Illustrating_Ideas","Instructor_Posistion",
                        "Experience_Teaching_Calculus", "Interested_in_Teaching_Calc",
                        "Interested_in_Teaching_Adv_Math", "Interested_in_Improving_Teaching",
                        "Primary_Role_as_an_Instructor", "Break_down_into_Subskills",
                        "Application_Problems",
                        "Discuss_Difficult_Problems_with_Colleagues", "Use_PreAssessments",
```

```
                          "Follow_Textbook",
                          "Use_Alternate_Sources", "Highest_Degree", "Highest_Field_Degree",
                          "Number_of_Calc_Sections_Teaching", "Number_of_Office_Hours",
                          "Help_Students_Outside_of_Office_Hours", "Number_of_Exams_Given",
                          "Country_of_Undergraduate", "Countries")
```

```
library(randomForest)
set.seed(472)

i <- sapply(cleandat, is.character)
cleandat[i] <- lapply(cleandat[i], as.factor)
i <- sapply(cleandat, is.integer)
cleandat[i] <- lapply(cleandat[i], as.numeric)

# dat.imputed <- rfImpute(sqr25grade ~ ., cleandat)
fit=randomForest(`Final Calc I Grade`~., data=cleandat)
```

Figure 1. Instructor Quality Variable Importance Plot

```
#Shows which variables are most important for predicting grade
IncNodePur <- fit$importance[1:55]
names <- colnames(cleandat)[2:56]

Imp <- as.data.frame(cbind(names, IncNodePur))
Imp$IncNodePur <- as.numeric(as.character(Imp$IncNodePur))
Imp$names <- factor(Imp$names, levels = Imp$names[order(Imp$IncNodePur)])
Imp <- Imp[order(-IncNodePur),]
Imp <- Imp[c(1:15,30),] # Choose top 15 variables



library(ggplot2)
p <- ggplot(Imp, aes(IncNodePur, names)) +
  geom_point(size=4, shape=16, col = "turquoise") +
  xlim(30,140) + ylab("Predictor") + labs(title="Variable Importance Plot") +
```

```
  theme_minimal() + theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold"))
```

Figure 2. Top Variable Partial Plots

```
library(pdp)
library(vip)

pa <- partial(fit, pred.var = "Helped_To_Become_A_Problem_Solver", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("Instructor Rating") + labs(title="Helped To Become A Problem Solver") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "steelblue1", lwd = 1.5)

pb <- partial(fit, pred.var = "Allowed_Time_to_Understand", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("Instructor Rating") + labs(title="Allowed Time to Understand") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "steelblue1", lwd = 1.5)

pc <- partial(fit, pred.var = "Discouraged_Calculus", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("Instructor Rating") + labs(title="Discouraged Calculus") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "steelblue1", lwd = 1.5)
```

```
pd <- partial(fit, pred.var = "Provided_Helpful_Explanations", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("Instructor Rating") + labs(title="Provided Helpful Explanations") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "steelblue1", lwd = 1.5)
```

Figure 3. Student Grades by Instructor's Country

```
library(webshot)
library(plotly)

newdat <- aggregate(cleandat[,1], list(cleandat$Countries), mean)

# light grey boundaries
l <- list(color = toRGB("grey"), width = 0.5)

# specify map projection/options
g <- list(
  showframe = FALSE,
  showcoastlines = FALSE,
  projection = list(type = 'Orthographic')
)

p <- plot_geo(newdat) %>%
  add_trace(
    z = newdat$x, colors = 'PuBu',
    locations = ~newdat$Group.1, marker = list(line = l)
  ) %>%
  colorbar(title = 'Student Grades') %>%
  layout(
    title = 'Student Grades based on Country of Instructors Undergraduate',
```

```
    geo = g
  )
p
```

Question 2 code

```
rm(list=ls())
gc()
```

```
##           used (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells 1031695 55.1    1770749  94.6  1770749  94.6
## Vcells 7985241 61.0   35185510 268.5 43889881 334.9
```

```
data <- read.csv("D:/School/Spring 2019/Stat 472/Calculus Retention/maalongdatafile_ANON.csv")
zipcode1 <- read.csv("D:/School/Spring 2019/Stat 472/Calculus Retention/maalongdatafile_ZIP.csv")

### Newly chosen variables to predict good vs bad students
p2dt <- data[,c(1,10,16, 32:40,seq(43,64,3),95,96,98:101,105:107,111,115:117,121,122,137,151:155,252:255

### Project 2 Data Frame
p2dt <- cbind(p2dt,data$sqr25grade)

### Add Zip codes based on student ID
library(zipcode)
data(zipcode)
zipc <- zipcode1$sp45zip
zipc <- clean.zipcodes(zipc)
zipcode2 <- cbind(zipcode1$studid, zipc); colnames(zipcode2) <- c('studid','zip')
p2dt <- merge(p2dt, zipcode2, by ="studid")

### Here we will omit unnecessary variables, and variables that don't have adequate values

# Only contains complete zip codes
```

```
statedt <- subset(p2dt, nchar(as.character(p2dt$zip))==5)

# Merge State names with zip code
statedt <- merge(statedt, zipcode, by='zip' )
testdt <- statedt
# If yes in any of the sp14 questions, then the value is 1, otherwise it equals 0
statedt$sp14 <- ifelse(statedt$sp14calc == 'Yes' | statedt$sp14calcab  == 'Yes'| statedt$sp14calcbc  ==

# Recoding spr26calc 2 so that yes = 1, otherwise 0
statedt$spr26calc2 <- ifelse(statedt$spr26calc2 == 'Yes', 1, 0)
statedt$spr28c2req <- ifelse(statedt$spr28c2req == 'Yes', 1, 0)

# Recode spr31intend so that 'not at all certain' = 0, and 'very certain' = 3
statedt$spr31intend <- as.character(statedt$spr31intend)
statedt$spr31intend <- ifelse(statedt$spr31intend == 'Not at all certain',0,statedt$spr31intend)
statedt$spr31intend <- ifelse(statedt$spr31intend == 'Very certain',3,statedt$spr31intend)

# Recode spr32diffic,spr33unsuccess, spr37study, sp43fatheduc, sp44motheduc 0 and 3 entries (similar to
statedt$spr32diffic <- as.character(statedt$spr32diffic)
statedt$spr32diffic <- ifelse(statedt$spr32diffic == 'I try hard to figure it out on my own', 0, stated
statedt$spr32diffic <- ifelse(statedt$spr32diffic == 'I quickly seek help or give up trying', 3, stated

# Recoding spr33unsuccess
statedt$spr33unsuccess <- as.character(statedt$spr33unsuccess)
statedt$spr33unsuccess <- ifelse(statedt$spr33unsuccess == 'a natural part of solving the problem', 0, s
statedt$spr33unsuccess <- ifelse(statedt$spr33unsuccess == 'an indication of my weakness in mathematics

# Recoding spr37study
statedt$spr37study <-as.character(statedt$spr37study)
statedt$spr37study <- ifelse(statedt$spr37study == 'memorize it the way it is presented', 0, statedt$sp
statedt$spr37study <- ifelse(statedt$spr37study == 'Make sense of the material, so that I understand it
```

```r
# Recoding sp43fatheduc
statedt$sp43fatheduc <- as.character(statedt$sp43fatheduc)
statedt$sp43fatheduc <- ifelse(statedt$sp43fatheduc == 'Did not finish high school', 1, statedt$sp43fatl
statedt$sp43fatheduc <- ifelse(statedt$sp43fatheduc == 'High school', 2, statedt$sp43fatheduc)
statedt$sp43fatheduc <- ifelse(statedt$sp43fatheduc == 'Some college', 3, statedt$sp43fatheduc)
statedt$sp43fatheduc <- ifelse(statedt$sp43fatheduc == 'Four years of college', 4, statedt$sp43fatheduc)
statedt$sp43fatheduc <- ifelse(statedt$sp43fatheduc == 'Graduate school', 5, statedt$sp43fatheduc)

# Recoding sp44motheduc
statedt$sp44motheduc <- as.character(statedt$sp44motheduc)
statedt$sp44motheduc <- ifelse(statedt$sp44motheduc == 'Did not finish high school', 1, statedt$sp44motl
statedt$sp44motheduc <- ifelse(statedt$sp44motheduc == 'High school', 2, statedt$sp44motheduc)
statedt$sp44motheduc <- ifelse(statedt$sp44motheduc == 'Some college', 3, statedt$sp44motheduc)
statedt$sp44motheduc <- ifelse(statedt$sp44motheduc == 'Four years of college', 4, statedt$sp44motheduc)
statedt$sp44motheduc <- ifelse(statedt$sp44motheduc == 'Graduate school', 5, statedt$sp44motheduc)

# Recoding sp52yearcoll
statedt$sp52yearcoll <- as.character(statedt$sp52yearcoll)
statedt$sp52yearcoll <- ifelse(statedt$sp52yearcoll == 'Freshman', 1, statedt$sp52yearcoll)
statedt$sp52yearcoll <- ifelse(statedt$sp52yearcoll == 'Sophomore', 2, statedt$sp52yearcoll)
statedt$sp52yearcoll <- ifelse(statedt$sp52yearcoll == 'Junior', 3, statedt$sp52yearcoll)
statedt$sp52yearcoll <- ifelse(statedt$sp52yearcoll == 'Senior', 4, statedt$sp52yearcoll)
statedt$sp52yearcoll <- ifelse(statedt$sp52yearcoll == 'Graduate Student', 5, statedt$sp52yearcoll)
statedt$sp52yearcoll <- ifelse(statedt$sp52yearcoll == 'Other', 6, statedt$sp52yearcoll)

# Recoding sp57work, sp58extra, sp59prep
statedt$sp57work <- as.character(statedt$sp57work)
statedt$sp57work <- ifelse(statedt$sp57work == "0 hours",0, statedt$sp57work)
statedt$sp57work <- ifelse(statedt$sp57work == "1-5 hours",3, statedt$sp57work)
statedt$sp57work <- ifelse(statedt$sp57work == "6-10 hours",8, statedt$sp57work)
statedt$sp57work <- ifelse(statedt$sp57work == "11-15 hours",13, statedt$sp57work)
statedt$sp57work <- ifelse(statedt$sp57work == "16-20 hours",18, statedt$sp57work)
```

```
statedt$sp57work <- ifelse(statedt$sp57work == "21-30 hours",25.5, statedt$sp57work)
statedt$sp57work <- ifelse(statedt$sp57work == "More than 30",35, statedt$sp57work)

statedt$sp58extra <- as.character(statedt$sp58extra)
statedt$sp58extra <- ifelse(statedt$sp58extra == "0 hours",0, statedt$sp58extra)
statedt$sp58extra <- ifelse(statedt$sp58extra == "1-5 hours",3, statedt$sp58extra)
statedt$sp58extra <- ifelse(statedt$sp58extra == "6-10 hours",8, statedt$sp58extra)
statedt$sp58extra <- ifelse(statedt$sp58extra == "11-15 hours",13, statedt$sp58extra)
statedt$sp58extra <- ifelse(statedt$sp58extra == "16-20 hours",18, statedt$sp58extra)
statedt$sp58extra <- ifelse(statedt$sp58extra == "21-30 hours",25.5, statedt$sp58extra)
statedt$sp58extra <- ifelse(statedt$sp58extra == "More than 30",35, statedt$sp58extra)

statedt$sp59prep <- as.character(statedt$sp59prep)
statedt$sp59prep <- ifelse(statedt$sp59prep == "0 hours",0, statedt$sp59prep)
statedt$sp59prep <- ifelse(statedt$sp59prep == "1-5 hours",3, statedt$sp59prep)
statedt$sp59prep <- ifelse(statedt$sp59prep == "6-10 hours",8, statedt$sp59prep)
statedt$sp59prep <- ifelse(statedt$sp59prep == "11-15 hours",13, statedt$sp59prep)
statedt$sp59prep <- ifelse(statedt$sp59prep == "16-20 hours",18, statedt$sp59prep)
statedt$sp59prep <- ifelse(statedt$sp59prep == "21-30 hours",25.5, statedt$sp59prep)
statedt$sp59prep <- ifelse(statedt$sp59prep == "More than 30",35, statedt$sp59prep)

# Recoding sq43text, sq34office, sq34online, sq34tutor, sq36outside
statedt$sq34text <- as.character(statedt$sq34text)
statedt$sq34office <- as.character(statedt$sq34office)
statedt$sq34online <- as.character(statedt$sq34online)
statedt$sq34tutor <- as.character(statedt$sq34tutor)
statedt$sq36outside <- as.character(statedt$sq36outside)

statedt$sq34text <- ifelse(statedt$sq34text == 'never', 0, statedt$sq34text)
statedt$sq34text <- ifelse(statedt$sq34text == 'some class sessions', 1, statedt$sq34text)
statedt$sq34text <- ifelse(statedt$sq34text == 'about half the class sessions', 2, statedt$sq34text)
statedt$sq34text <- ifelse(statedt$sq34text == 'most class sessions', 3, statedt$sq34text)
```

```
statedt$sq34text <- ifelse(statedt$sq34text == 'every class session', 4, statedt$sq34text)

statedt$sq34office <- ifelse(statedt$sq34office == 'never', 0, statedt$sq34office)
statedt$sq34office <- ifelse(statedt$sq34office == 'some class sessions', 1, statedt$sq34office)
statedt$sq34office <- ifelse(statedt$sq34office == 'about half the class sessions', 2, statedt$sq34offi
statedt$sq34office <- ifelse(statedt$sq34office == 'most class sessions', 3, statedt$sq34office)
statedt$sq34office <- ifelse(statedt$sq34office == 'every class session', 4, statedt$sq34office)

statedt$sq34online <- ifelse(statedt$sq34online == 'never', 0, statedt$sq34online)
statedt$sq34online <- ifelse(statedt$sq34online == 'some class sessions', 1, statedt$sq34online)
statedt$sq34online <- ifelse(statedt$sq34online == 'about half the class sessions', 2, statedt$sq34onli
statedt$sq34online <- ifelse(statedt$sq34online == 'most class sessions', 3, statedt$sq34online)
statedt$sq34online <- ifelse(statedt$sq34online == 'every class session', 4, statedt$sq34online)

statedt$sq34tutor <- ifelse(statedt$sq34tutor == 'never', 0, statedt$sq34tutor)
statedt$sq34tutor <- ifelse(statedt$sq34tutor == 'some class sessions', 1, statedt$sq34tutor)
statedt$sq34tutor <- ifelse(statedt$sq34tutor == 'about half the class sessions', 2, statedt$sq34tutor)
statedt$sq34tutor <- ifelse(statedt$sq34tutor == 'most class sessions', 3, statedt$sq34tutor)
statedt$sq34tutor <- ifelse(statedt$sq34tutor == 'every class session', 4, statedt$sq34tutor)

statedt$sq36outside <- ifelse(statedt$sq36outside == 'Yes', 1, statedt$sq36outside)
statedt$sq36outside <- ifelse(statedt$sq36outside == 'No', 0, statedt$sq36outside)

# SAT/ACT combining
SATscore <- rev(seq(200,800,by=10)) #61 numbers
SATper <- c(99,99,99,98,97,97,96,95,95,94,93,91,90,88,87,85,83,82,79,77,75,73,70,67,64,62,59,55,52,49,45
SAT = SATper[match(statedt$sp3satmathscore,SATscore)]
ACTscore <- 36:1
ACTper <- c(99,99,99,99,98,96,95,92,90,87,83,79,74,68,62,56,50,43,36,30,24,18,12,7,4,1,1,1,1,1,1,1,1,1,
ACT = ACTper[match(statedt$sp7actmathscore, ACTscore)]

# creating average standardized test percentile
```

```
X_Percentile<- apply(cbind(SAT,ACT), 1, mean, na.rm = T )


# recombining with the data frame statedt
statedt <- cbind(statedt, X_Percentile)


# Recode some survey stuff in our data
i <- sapply(statedt, is.factor)
statedt[i] <- lapply(statedt[i], as.character)
for(i in 1:dim(statedt)[1]){ for(j in 1:dim(statedt)[2]){
statedt[i,j] <- ifelse(statedt[i,j]=="Strongly agree", 6,
ifelse(statedt[i,j]=="Agree", 5,
ifelse(statedt[i,j]=="Slightly agree", 4,
ifelse(statedt[i,j]=="Slightly disagree", 3,
ifelse(statedt[i,j]=="Disagree", 2,
ifelse(statedt[i,j]=="Strongly disagree", 1,
ifelse(statedt[i,j]=="Not at all", 0,
ifelse(statedt[i,j]=="Very often", 5,
statedt[i,j])))))))))
}}
```

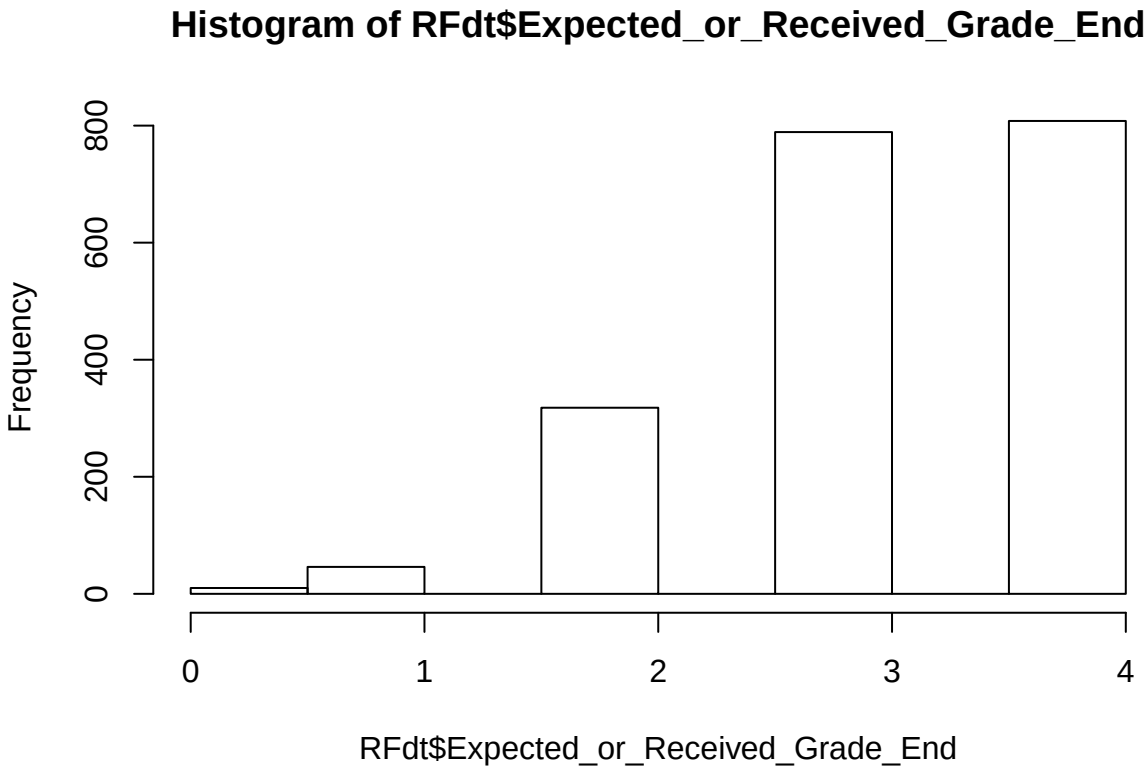
```
# Ommitting uneccesary columns
NNdt <- statedt[,-c(1:13,16,18:21,30, 41,49:53)]


# Changing everything to numeric and omitting NAs/empty cells
NNdt = as.matrix(as.data.frame(lapply(NNdt, as.numeric)))
NNdt <- NNdt[complete.cases(NNdt),]


# experiment with scaling
NNdt <- data.frame(NNdt)


# Creating training and testing dataset
# n <- round(.75*(nrow(NNdttest)))
# training_sample = sample(nrow(NNdttest), size = n, replace = FALSE)
```

```
# train <- NNdttest[training_sample,]
# test <- NNdttest[-training_sample,]

# Creating training and testing dataset
# n <- round(.75*(nrow(NNdt)))
# training_sample = sample(nrow(NNdt), size = n, replace = FALSE)
# train <- NNdt[training_sample,]
# test <- NNdt[-training_sample,]

# train <- data.frame(train)
# test <- data.frame(test)
# library(neuralnet)

# All data
# nn <- neuralnet(
# as.factor(data.sqr25grade)~.,
# data=train, hidden=c(15), err.fct="ce",
# linear.output=FALSE, lifesign = 'full', learningrate = .01)
# trainS <- scale(train)

# Some Data
# nn <- neuralnet(
# as.factor(sqr25grade)~sp57work + sp58extra + sp59prep + sp14 + X_Percentile + sp15geograde + sp15alg2
# data=trainS, hidden=c(7), err.fct="ce",
# linear.output=FALSE, lifesign = 'full', learningrate = .01)

# plot(nn)

## Confusion Matrix
# p = predict(nn,test)
# prediction = apply(p,1,which.max)
# table(test$spr25grade,prediction)
```

Figure 4: End of Semester Grades

```
# Ommitting uneccesary columns
NNdt <- statedt[,-c(1:13,16,18:21,30, 41,49:53)]


# Changing everything to numeric and omitting NAs/empty cells
NNdt = as.matrix(as.data.frame(lapply(NNdt, as.numeric)))
NNdt <- NNdt[complete.cases(NNdt),]


# experiment with scaling
NNdt <- data.frame(NNdt)


# Random Forest
RFdt <- NNdt


# Placing Expected_or_Received_Grade_End at the end of the dataframe
RFdt <- RFdt[,-c(28)]
RFdt <- cbind(RFdt, NNdt$sqr25grade)


# Renaming columns
colnames(RFdt) <- c('HS_Geometry_Grade','HS_Algebra_Grade', 'HS_Precalc_Grade',
                    'Expected_Calc_Grade_Begin', 'Intend_to_take_Calc_2',
                    'Calc_2_Required_for_Major', 'Belief_in_Knowledge',
                    'Confidence_in_Math_Ability', 'Understanding_of_Studied_Math',
                    'Certainty_in_Intentions_After_College', 'Perception_of_Difficulty',
                    'Memorize_or_Understand_While_Studying', 'How_Students_Percieve_Struggle_with_Proble
                    'Solving_Math_Problems_is_Satisfying','Father_Education','Mother_education', 'Year_
                    'Difficulty_Paying_For_College','Read_Textbook_Before_Class',
                    'Visit_Instructor_Office_Hours','Used_Online_Tutoring','Visited_Tutor',
                    'Met_With_Students_To_Study','Taken_HS_Calc','ACT_or_SAT_Percentile',
                    'Expected_or_Received_Grade_End')


# High Proportion of students in data set received an A or B. Perhaps students who chose
```

```r
# to respond to the survey were more likely to have higher grade.


hist(RFdt$Expected_or_Received_Grade_End)
```




```r
### Random Forest
set.seed(80085)
library(randomForest)
fit = randomForest(`Expected_or_Received_Grade_End` ~ ., data = RFdt)


library(ggplot2)
ggplot(data=RFdt, aes(RFdt$Expected_or_Received_Grade_End)) +
  geom_bar(aes(fill=..count..)) +
  scale_fill_gradient("Count", low = "#99FF66", high = "#003300") +
  labs(x='End of Semester Reported Grade') + theme_minimal()
```

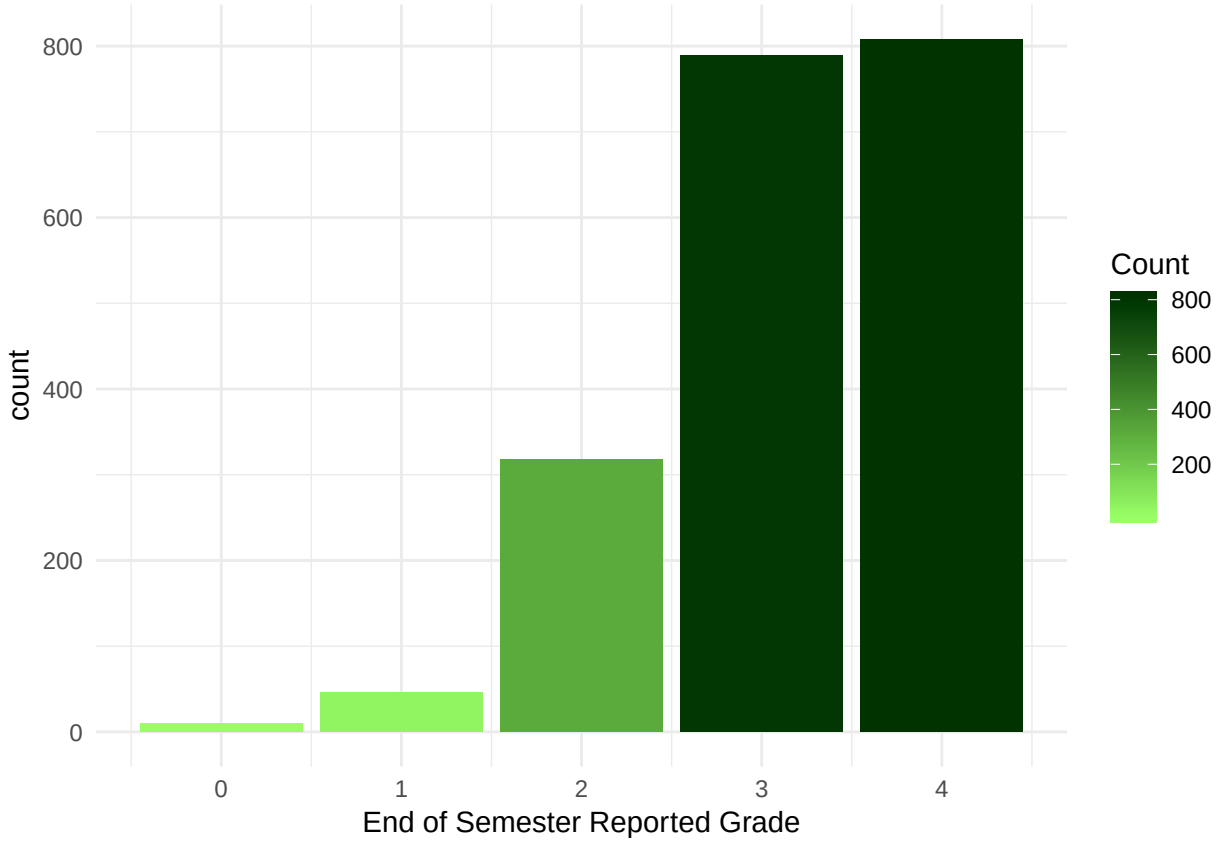


Figure 5: Variable Importance Plot

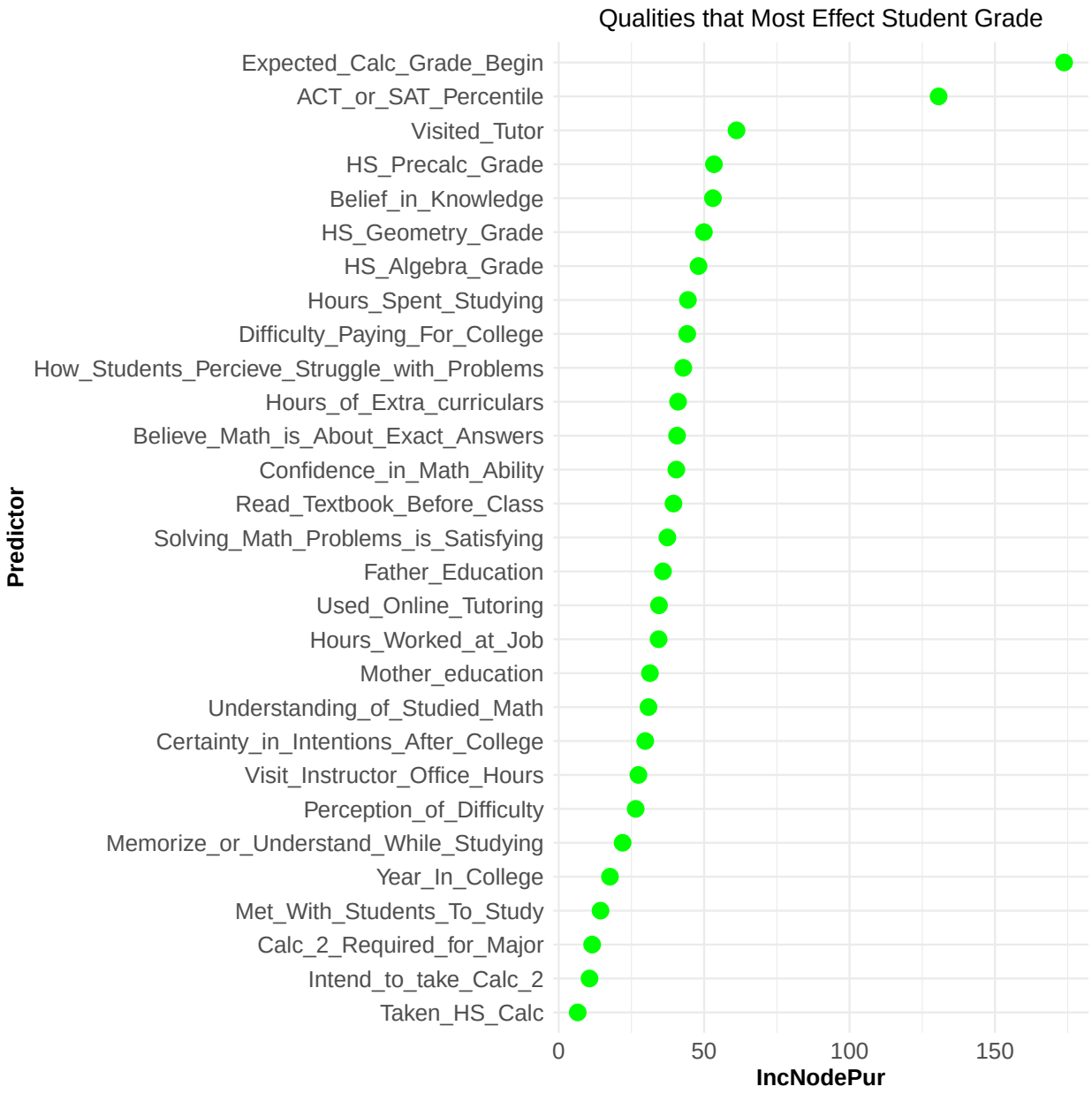


Figure 6: Partial Plots (Top Variables)

```
### Time for Partial plots!

library(pdp)

library(vip)


pa <- partial(fit, pred.var = "Expected_Calc_Grade_Begin", plot = TRUE,

        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +

  xlab("Hypothesized Calc Grade") + labs(title="Hypothesized Calc Grade at Semester Beginning") +
```

```
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "green4", lwd = 1.5)

pb <- partial(fit, pred.var = "ACT_or_SAT_Percentile", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("ACT/SAT Percentile") + labs(title="ACT or SAT Percentile") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "green4", lwd = 1.5)

pc <- partial(fit, pred.var = "Visited_Tutor", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("How Often Tutor Was Visited") + labs(title="Tutor Visitation") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "green4", lwd = 1.5)

pd <- partial(fit, pred.var = "HS_Precalc_Grade", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("High School Precalc Grade") + labs(title="High School Precalc Grade Effect") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "green4", lwd = 1.5)

grid.arrange(pa, pb, pc, pd)
```

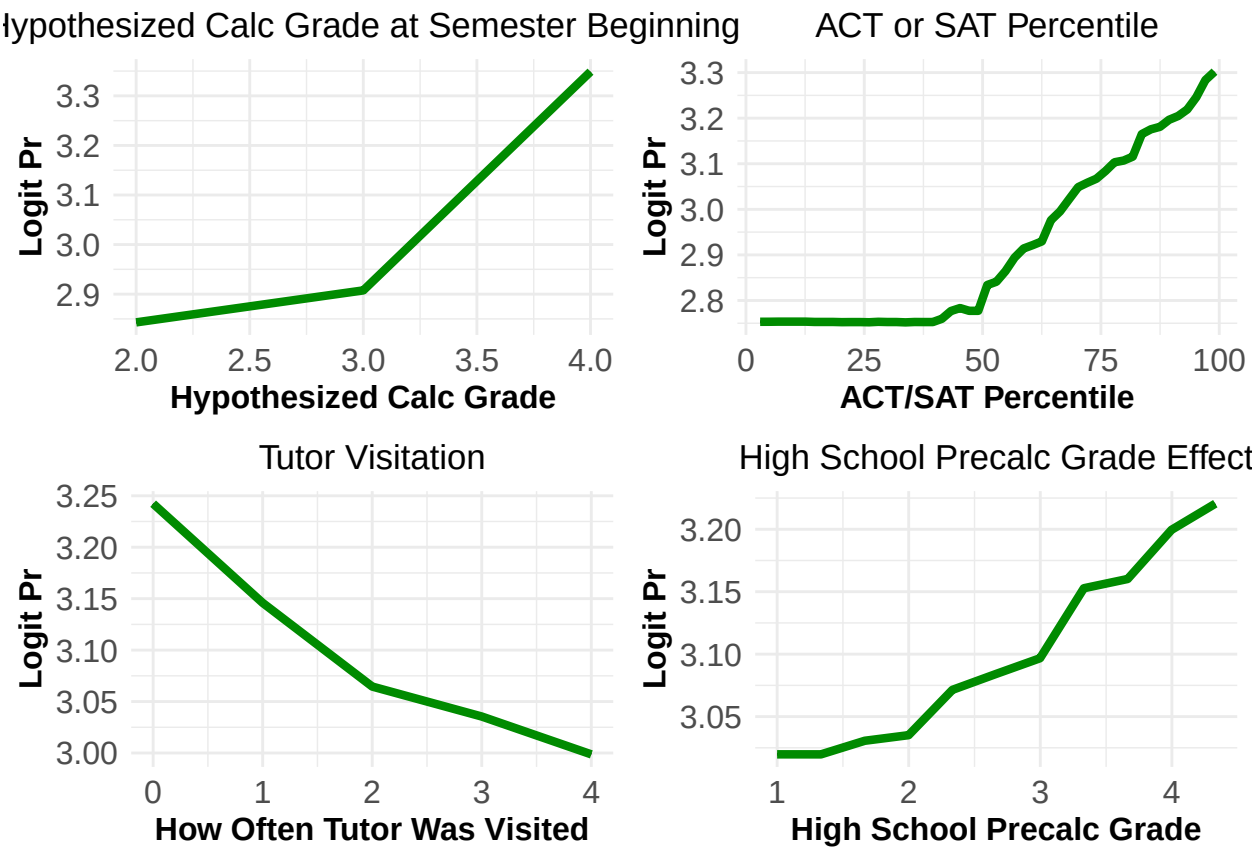
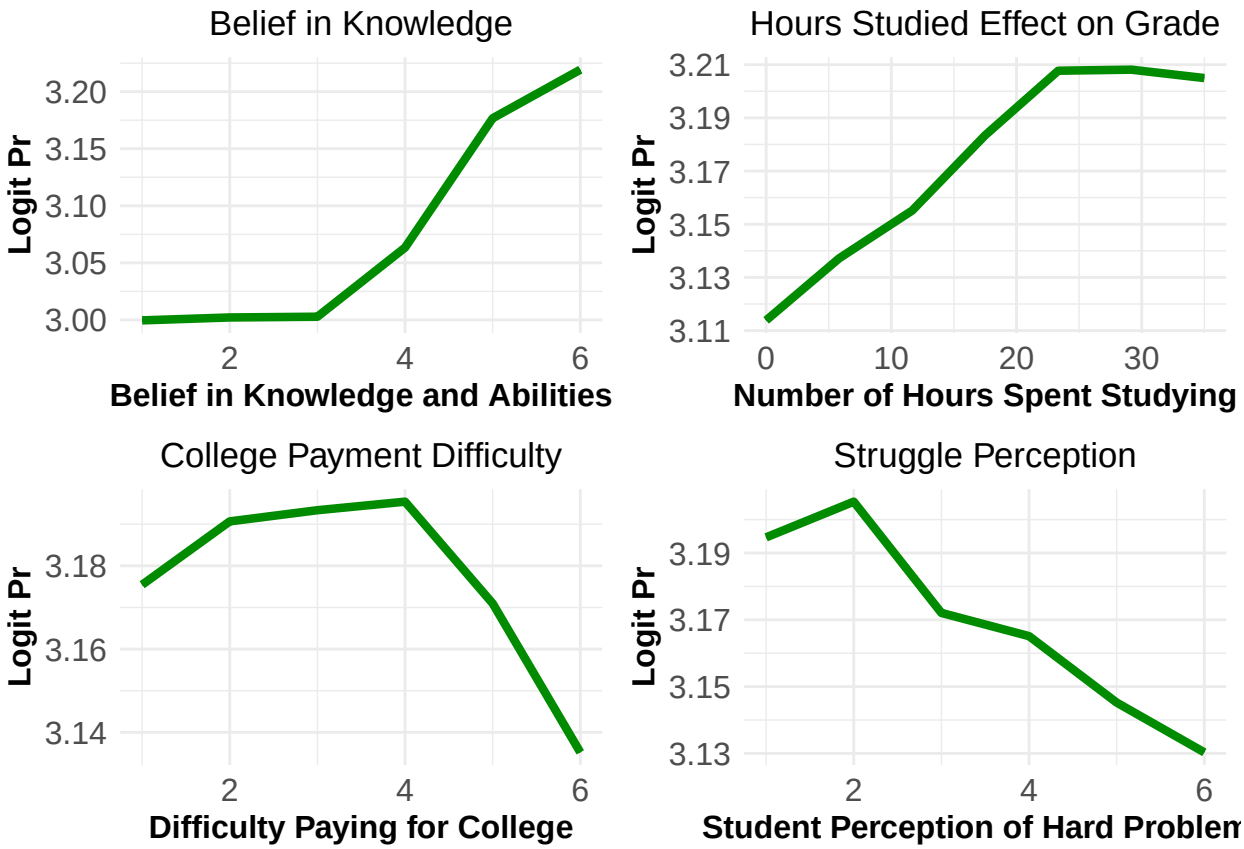


Figure 7: Partial Plots (Other Interesting Variables)

```
## More Plots...
pe <- partial(fit, pred.var = "Belief_in_Knowledge", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("Belief in Knowledge and Abilities") + labs(title="Belief in Knowledge") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "green4", lwd = 1.5)


pf <- partial(fit, pred.var = "Hours_Spent_Studying", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("Number of Hours Spent Studying") + labs(title="Hours Studied Effect on Grade") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
```

```
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "green4", lwd = 1.5)

pg <- partial(fit, pred.var = "Difficulty_Paying_For_College", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("Difficulty Paying for College") + labs(title="College Payment Difficulty") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "green4", lwd = 1.5)

ph <- partial(fit, pred.var = "How_Students_Percieve_Struggle_with_Problems", plot = TRUE,
        plot.engine = "ggplot2") + theme_minimal() + ylab("Logit Pr") +
  xlab("Student Perception of Hard Problems") + labs(title="Struggle Perception") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text=element_text(size=12),
        axis.title=element_text(size=12,face="bold")) +
  geom_line(color = "green4", lwd = 1.5)

grid.arrange(pe, pf, pg, ph)
```



Question 3 code

```
library(tidyr)
library(modelr)
library(broom)
library(InformationValue)
library(car)


raw <- read.csv("D:/School/Spring 2019/Stat 472/Calculus Retention/maalongdatafile_ANON.csv")
### Variables chosen based off of Crystal's google doc section


#previous SAT/ACT grade
SATscore <- rev(seq(200,800,by=10)) #61 numbers
SATper <- c(99,99,99,98,97,97,96,95,95,94,93,91,90,88,87,85,83,82,79,77,75,73,70
```

```
            ,67,64,62,59,55,52,49,45,42,40,36,33,30,27,24,21,19,16,14,12,10,9,7,
            6,5,4,3,3,2,2,1,1,1,1,1,0,0,0)
SAT = SATper[match(raw$sp3satmathscore,SATscore)]
ACTscore <- 36:1
ACTper <- c(99,99,99,99,98,96,95,92,90,87,83,79,74,68,62,56,50,43,36,30,24,18,
            12,7,4,1,1,1,1,1,1,1,1,1,1,1)
ACT = ACTper[match(raw$sp7actmathscore, ACTscore)]
X_Percentile<- apply(cbind(SAT,ACT), 1, mean, na.rm = T )
X_Percentile <-as.numeric(X_Percentile)
raw <- cbind(raw,X_Percentile)


X_prev_calc <- rep(NA,nrow(raw))
X_prev_calc[raw$sp15geograde>0|raw$sp15alg2grade>0|raw$sp15imgrade>0|
              raw$sp15precgrade>0|raw$sp15triggrade>0|raw$sp15statgrade>0|
              raw$sp15calcgrade>0|raw$sp15othgrade>0]<-"1"
X_prev_calc[raw$sp17abgrade>0]<-"1"
X_prev_calc[raw$sp17bcgrade>0]<-"1"
X_prev_calc[raw$sp18calccol == "Yes"]<-"0"
X_prev_calc[is.na(X_prev_calc) & raw$sp18calccol == "No"]<-""
raw <- cbind(raw,X_prev_calc)



raw$spr36choice <- as.character(raw$spr36choice)
raw$choice <- ifelse(raw$spr36choice == "I would never take another mathematics course",1,
              ifelse(raw$spr36choice == "I would continue to take mathematics",3,
              ifelse(raw$spr36choice ==2,2,"")))


raw$sp54homesup <- as.character(raw$sp54homesup)
raw$supp <- ifelse(raw$sp54homesup == "Not at all",0,
            ifelse(raw$sp54homesup =="Somewhat",1,
            ifelse(raw$sp54homesup =="Strongly",2,
            ifelse(raw$sp54homesup == "Very Strongly", 3,""))))
```

```
raw$supp_family <- ifelse(raw$sp55encno == 'Yes' | raw$sp55encmoth  == 'Yes'|
                          raw$sp55encfath  == 'Yes'| raw$sp55encsib  == 'Yes'|
                          raw$sp55encrel == 'Yes'| raw$sp55enccouns  == 'Yes'|
                          raw$sp55encmatht  == 'Yes'| raw$sp55encotht  == 'Yes'|
                          raw$sp55enccoach== 'Yes', 1, 0)


raw$sp61pay <- as.character(raw$sp61pay)
raw$sp61pay_code <- ifelse(raw$sp61pay =="Strongly Agree", 5,
                    ifelse(raw$sp61pay =="Agree", 4,
                    ifelse(raw$sp61pay =="Slightly agree",3,
                    ifelse(raw$sp61pay =="Slightly disagree", 2,
                    ifelse(raw$sp61pay =="Disagree", 1,
                    ifelse(raw$sp61pay =="Strongly disagree", 0,"")))))
```

```
myvar <- c('X_Percentile','X_prev_calc','supp','supp_family','sp57workl','sp58extral','sp59prepl','sqr2
clean <- raw[myvar]
clean$diff <- clean$sqr25grade-clean$spr25grade
clean <- clean[which(clean$supp != "" & clean$X_prev_calc !="" &clean$diff != ""),]
clean <- clean[complete.cases(clean),]
clean$supp <- as.factor(clean$supp)
clean$supp_family <- as.factor(clean$supp_family)
clean1 <- clean[which(clean$diff!=0),]
clean1$expect <- ifelse(clean1$diff> 0,1,0)
```

Figure 8: Odds of Switching

```
model1 <- glm(formula = expect ~ X_Percentile+X_prev_calc+supp+sp57workl+sp58extral+sp59prepl, family =
predicted1 <- plogis(predict(model1, clean1))
optCutOff1 <- optimalCutoff(clean1$diff, predicted1)[1]
sum <- summary(model1)
or <- exp(sum$coefficients)
sum_est <- sum$coefficients[c(2,3,4,5,6,7,8)]
sum_se <- sum$coefficients[c(2,3,4,5,6,7,8),2]
```

```r
lower <- sum_est - 1.96*sum_se
upper <- sum_est + 1.96*sum_se
or_est <- exp(sum_est)
or_lower <- round(exp(lower),3); or_upper <- round(exp(upper),3)
CI <- rbind(or_lower, or_upper)
ORc1 <- round(exp(sum$coefficients[2:8,1]),3)
ORc2 <- paste("(",or_lower,",",or_upper,")")
ORdt <- cbind(ORc1, ORc2)
colnames(ORdt) <- c("Odds Ratio", "Odds Ratio CI")
rownames(ORdt) <- c("SAT/ACT Percentile","Previou calculus experience","supportive home environment sup

library(ggplot2)
# Create labels
boxLabels = c("SAT/ACT Percentile","Previou calculus experience","supportive home environment support",
# Enter summary data. boxOdds are the odds ratios, boxCILow is the lower bound of the CI,
# boxCIHigh is the upper bound.
yAxis = c(1,2,3,4,5,6,7)
df <- data.frame(
boxOdds = or_est,
boxCILow = or_lower,
boxCIHigh = or_upper
)
# Plot
p <- ggplot(df, aes(x = boxOdds, y = yAxis))
p + geom_vline(aes(xintercept = 1), size = .25, linetype = "dashed") +
geom_errorbarh(aes(xmax = boxCIHigh, xmin = boxCILow), size = 1, height = .2,
color = ifelse(or_upper < 1, "seagreen3", ifelse(or_lower > 1, "cyan3", "grey50"))) +
geom_point(size = 4, pch = 18,
color = ifelse(or_upper < 1, "seagreen3", ifelse(or_lower > 1, "cyan3", "grey50"))) +
theme_bw() +
theme(panel.grid.minor = element_blank()) +
scale_y_continuous(breaks = yAxis, labels = boxLabels) +
```

```
scale_x_continuous(breaks = seq(0,7,1) ) +
ylab("") +
xlab("Odds ratio") +
ggtitle("Estimated Odds of Switching") +
theme(plot.title = element_text(hjust = 0.5, size = 16),
axis.text=element_text(size=12),
axis.title=element_text(size=14)) +
geom_text(aes(label = round(boxOdds,2)),vjust = -0.6,
color = ifelse(or_upper < 1, "seagreen3", ifelse(or_lower > 1, "cyan3", "grey50")))
```

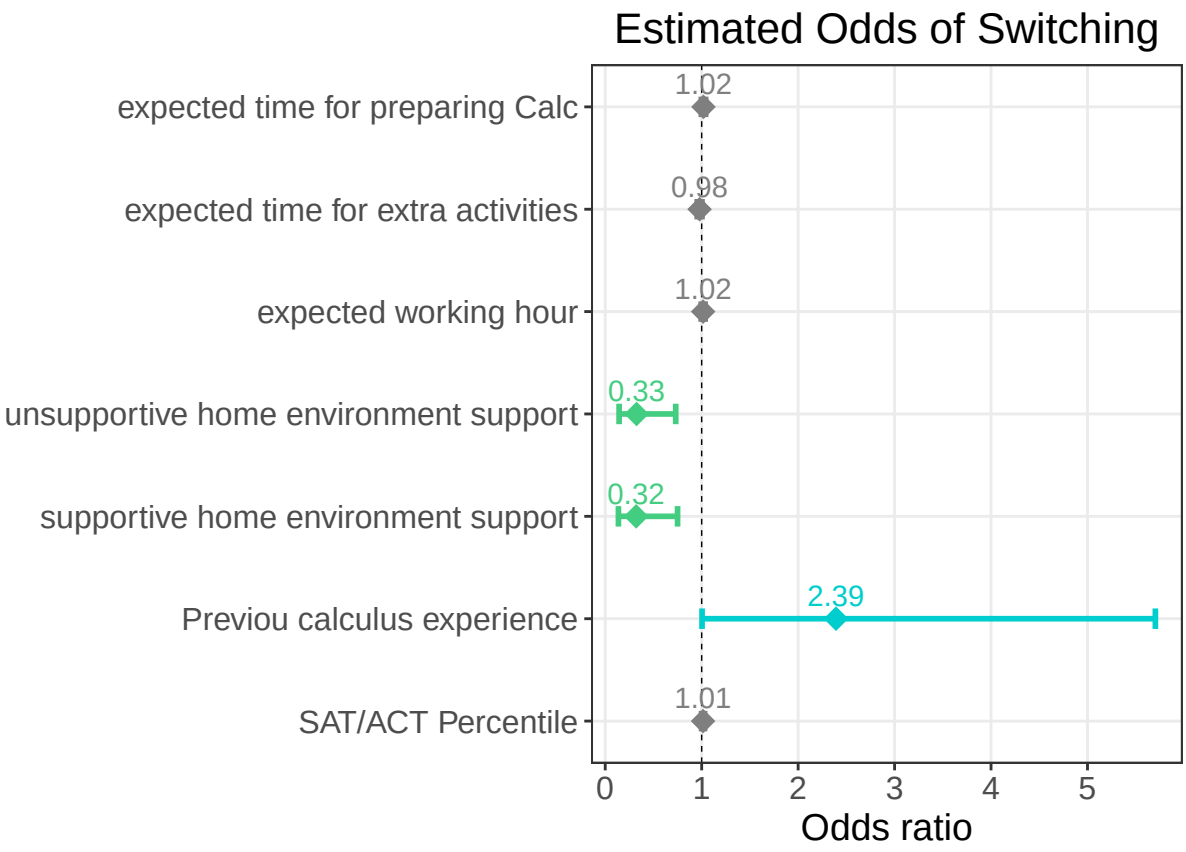


Table 1: Confusion Matrix

```
table <- confusionMatrix(clean1$expect, predicted1, threshold = optCutOff1)
colnames(table) <- c("Predict: False","Predict: True")
row.names(table) <- c("Actual: False","Actual: True")
table
```

Figure 9: ROC Curve

```
a <- sensitivity(clean1$expect,predicted1,threshold = optCutOff1)
b <- specificity(clean1$expect,predicted1,threshold = optCutOff1)
plotROC(clean1$expect, predicted1)
```