**Project 3B: Midterm Exam**

**Name: Juan Villegas**

**Description of the Data**

Gender: M, F

Residency:

Out: Out of State, east: In-state but east CT River, west: In-state but west CT River

VSAT: Verbal SAT

MSAT: Math SAT

WSAT: Writing SAT

Athletics: 1 = team athlete, 0 = not team athlete

Housing: 1 = lives on campus, 0 = commuter

FGEN: 1 = First-Generation college student, 0 = Not First-Generation college student

HS Rank: Rank in high school class

HSGPA: High school GPA (converted to 4-point scale)

Stem: Indicated interest in STEM major (Math, Bio, etc.)

ERG: A ranking of CT high schools based on multiple indicators including socioeconomic level:

1 being highest ranked schools and 9 lowest ranked schools.

GotSchol: 1 = got a scholarship, 0 = did not get a scholarship

**Retained: Y = returned to Eastern for a second year, N = did not return to Eastern for second year**

**1st\_yr\_GPA: GPA after 1 year at Eastern**

------------------------

**The two variables in bold are target variables. The goal is to build models to predict the following:**

* Whether a student will be retained at Eastern for the second year (Success = Not Retained, Failure = Retained)
* Student’s first year GPA.
* We are particularly interested in whether a student will be at risk academically. So, once you have found a model to predict first-year GPA, you can turn to the question of classifying a

student into academically at risk (= Success) or not academically at risk (= Failure).

Using Machine Learning to Develop Predictive Models based on input and output data: High School 🡪 College Classification & Regression

Mata 343: Data Analytics

4/20/2022

Juan Villegas

Table of Contents

[Questions we are trying to answer. (Problem Understanding Phase) 5](#_Toc101444650)

[1st Question 5](#_Toc101444651)

[2nd Question 5](#_Toc101444652)

[3rd Question 6](#_Toc101444653)

[Cleaning up data (Data Preparation Phase) 6](#_Toc101444654)

[Making Sure R reads categorical variables correctly (as.factor) 7](#_Toc101444655)

[Missing Values 8](#_Toc101444656)

[Partitioning & Validating 12](#_Toc101444657)

[Partitioning 12](#_Toc101444658)

[Validating 12](#_Toc101444659)

[Multiple Regression (1st Year GPA) 15](#_Toc101444660)

[Use of MASS Package (AIC Command) 15](#_Toc101444661)

[Explaining my Regression model 20](#_Toc101444662)

[Evaluating my Regression model: 21](#_Toc101444663)

[R^2 21](#_Toc101444664)

[Minmax 21](#_Toc101444665)

[CART MODEL (Retention) 22](#_Toc101444666)

[CART Model: Evaluation 23](#_Toc101444667)

[Academic Risk (Logistic Regression Model) 23](#_Toc101444668)

[Creating my logistic Regression Model (Stepwise) 25](#_Toc101444669)

[Forward Model 25](#_Toc101444670)

[Backward Model 26](#_Toc101444671)

[Both (starting with a full and empty logistic regression model) 27](#_Toc101444672)

[Interpretation of Logistic Regression Model: 29](#_Toc101444673)

[Regression Model Accuracy 30](#_Toc101444674)

# Questions we are trying to answer. (Problem Understanding Phase)

For the first phase of my midterm, I will explain what the question we are trying to answer is and how I am going to tackle each question.

In this midterm, I was given the task to build 3 models to answer 3 questions:

1. Whether a student will be retained at Eastern for the second year (Success = Not Retained, Failure = Retained)
2. Student’s first year GPA.
3. Once, you have created a model to predict 1st year GPA, classify whether a student is at academic risk or not.

\*I use the statistical analysis programming language, R, to create models and answer these questions\*

## 1st Question

To tackle the first question, I will create a CART model that will classify whether a student will be retained for the second year. I’ll make ‘Retained’ be the response/dependent/target variable when creating this model. CART will decide on its own how it chooses to determine if a student is retained for a second year.

## 2nd Question

To answer the second question: I will install the MASS package so that I can use the AIC command. I will create a regression model with `=1st yearGPA’ as the predictor variable and the other variables in my data set as predictor variables. I have the option to run a Forward, Backward or ‘Both’ stepwise. Forward stepwise starts the model with no predictors, just one intercept and searches through all the single-variable models, adding variables, until we find the best one. Backward stepwise starts with all the predictor variables and removes variables with the least statistically significant (the largest p-value) one by one. Both stepwise does both: It starts as forward stepwise but, it also does backwards stepwise. The point of this command is to see what gives the lowest AIC which is almost like adjusted R^2.

After AIC picks my multiple regression model, I will look to see what variables where picked. If I believed important variables are missing, I will add them to my model. If I see variables, I am not fond of, I will remove them.

After this is all done, I will see the accuracy of my model.

## 3rd Question

To answer the 3rd and final question: I am left with two options. K-Nearest Neighbor or Logistic Regression (I could also do another CART model but, I want to display some of the tools I have in my toolbox.). Apart from the classification methods I must pick, I also must do a little bit of research in my domain. I will have to learn: What GPAs correspond with what letter grades, what are Eastern’s GPA requirements for joining certain majors, and what GPA do certain majors require to continue to be in them. I will need to learn this information before I can create a threshold to classify whether a student is at academic risk.

# Cleaning up data (Data Preparation Phase)

The data set I received had a ton of missing values. The continuous variables VSAT, MSAT, WSAT had 323 missing values, the Variable East.West.Out had 62 Missing values, HSRank had 1345 Missing Values, HSGPA had 31 Missing Values, and the variable ERG had 304 missing values.

Another issue this data set had was that R did not read some categorical variables correctly, it read them as continuous.

## Making Sure R reads categorical variables correctly (as.factor)

The first step I took toward cleaning up my data was to make sure that all my variables are correctly understood by R.

I made sure that the variables: ERG, East.West.Out, Athletics, Housing, FGen, Stem, GotSchol are all read as categorical variables.

I gave the variable Athletes the categories: Non-Athletes and Athletes

I gave the variable Housing the categories: Commuter and Lives on Campus

I gave the variable FGen the categories: Not First Gen and First Gen

I gave the variable Stem the categories: Not interested in stem and interested in stem

I gave the variable GotSchol the categories: received no scholarship and received scholarship

I gave the variable ERG the categories: ERG1, ERG2,ERG3,…,ERG9

I gave the variable East.West.Out the categories: East, West, and Out

## Missing Values

Chart, bubble chart

Description automatically generatedFirst, I focused on the variable with the highest number of missing values: HSRank. I created a correlation table of all my variables to see which ones where most correlated to see if I could create a regression model for highly correlated variables. Since there is a good number of variables in my data, I continued by creating a plot that showed me a plot of my variables that are color coated. Text

Description automatically generated with medium confidence

The correlogram to the right displays the most highly correlated variables. The darker the shades of blue or red, the higher the positive or negative correlation between the variables. Immediately, it is noticeable that the variable pair that are highest correlated are: HSRank~HSGPA and VSAT~WSAT.

Text

Description automatically generatedNow That I know that HSRank and HSGPA are closely correlated, I will make a regression model of HSRank with HSGPA as a predicter variable to predict some of the missing values for HSRank. My regression model for HSRank has an R^2 of .60.

According to our model, When GPA is 0, a person predicted HS Rank is -45.11 (our in real life probably 0). For every 1-point increase in GPA, a student’s HS Rank increases by 34.1617.

After applying this linear regression model to all the missing values in our data, we went from having 1345 missing values/records for HSRank to 21.

Text

Description automatically generatedNow I’m going to create a regression model to predict GPA using HS Rank to bring my HSGPA variable to have the same number of missing values.

My regression model states that: when HS Rank is 0, a student predicted HS rank is 0, there HSGPA is 1.84. For every 1 increase in HS Rank, a student’s HSGPA increases by .0205

Now, to confront some of my other missing values like VSAT, WSAT, AND MSAT, I will create histograms of these variables. If these variables have normal distributions, I will replace there missing values with the mean.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Because all 3 histograms have are normally distributed, I will continue by replacing the missing records/values of these variables with the mean.

To finish off the last set of missing values in the variables: East.West.Out and ERG, I will take a different approach. I will distribute my missing values proportionally to each category.

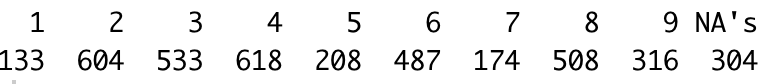
I have 62 missing values in the category East.West.Out. The variables summary looks like this: Text

Description automatically generated

I will divide the count of each variable by 3828 (The total number of records without missing values) and distributed the missing values by proportions. East will receive 24 (41.04% 🡪 1569/3823) of the missing values, Out will receive 4 of the missing values (6.64% 🡪 254/3823) and West will receive 34 (52.31% 🡪 2000/3823) of the missing values.

A picture containing text

Description automatically generated

I have 304 missing values in the category ERG The variables summary looks like this: 

I will divide the count of each variable by 3581 (The total number of records without missing values) and distributed the missing values by proportions. 1 will receive 10 of the missing values, 2 will receive 60 of the missing values and 3 will receive 55 of the missing values, 4 will receive 52 of the missing values, 5 will receive 14 of the missing values, 6 will receive 36 of the missing values, 7 will receive 12 of the missing values, 8 will receive 44 of the missing values and 9 will receive 21 of the missing values.A picture containing text

Description automatically generated

# Partitioning & Validating

Now That I've handled all my missing values, I can begin to partition my data into a training and testing group so that I can have 'foreign' data to see how accurate my model is.

## Partitioning

Before I start partitioning my data set, I will set a seed so that every time I run this partition, my data remains consistent. I set my seed to 1717.

I will my partition my data by a 75-25 split. 75% of my data will go into my training model and 25% will go into my testing model.

## Validating

Now I will begin to validate my variables. First, I will conduct a chi-square test to validate my categorical variables.A picture containing text

Description automatically generatedA picture containing text

Description automatically generated

Since all my p-values are bigger than .1, I know that there is no significance within my partitions, so I can move along with my project. I will now begin to validate my numerical (continuous) variables.

For my Continuous variables, I will create boxplots of all my variables to make sure that the boxplots are similar. Then I will conduct a Kruskal-Wallis and make sure that my P-Values are also bigger than 0.1.

Engineering drawing, box and whisker chart

Description automatically generated

The figure above shows that most of the training boxplots are familiar like there testing boxplots. Now to conduct there Kruskal Wallis Test.

Text

Description automatically generated

The figure above shows that all the Kruskal-Walli’s test for my 6 continuous variables is above 0.1. This means that I successfully completed my partition, and I can continue forward and make some models!

# Multiple Regression (1st Year GPA)

To predict first year GPA, I decided a multiple regression model would be the best type of model I’ve learned so far to use.

I will be using the MASS package so that I can use the AIC command to create my model. I will audit the models that forward/backward/stepwise made and pick the one with the smallest AIC. Then I will see if I should add or exclude certain variables that my model came up with.

I started off by creating a linear model with only the target variable ‘1st year GPA’ and nothing else. After that, I created a multiple regression model called ‘full’ that includes all the predictor variables I have available. The target variable is still 1st year GPA, the predictor variables are everything else from my data set.

## A screenshot of a computer Description automatically generated with medium confidenceUse of MASS Package (AIC Command)

Then I created a forward stepwise model (The model to the right). This model produced an AIC of -163.49. The adjusted r^2 from this model was .1923. The linear regression model it created was:

lm(formula = X1st\_yr\_GPA ~ HSGPA + ERG8 + WSAT + Housing + Stem + East + FGEN + HSRank + ERG6 + ERG3 + ERG4 + MSAT + ERG1, data = MidTermDataTrain)

A screenshot of a computer

Description automatically generated with medium confidenceI continued by creating a backward stepwise model (The model to the right). This model produced an AIC of -167.13. The adjusted r^2 from this model was .1938.

The linear regression model created from this stepwise was:

lm(formula = X1st\_yr\_GPA ~ MSAT + WSAT + HSRank + East + Housing + FGEN + HSGPA + Stem + ERG1 + ERG2 + ERG3 + ERG4 + ERG5 + ERG7 + ERG8, data = MidTermDataTrain)

Lastly, I also created the ‘both’ stepwise but, the stepAIC outputted both the forward and backward linear regression function. I received the Backward linear regression function when I put all the predictor variables into the command, and I received the forward when I put an empty linear function into the command. \*See below\* I decided to pick the backward stepwise selection model because it A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with medium confidencehad the lowest AIC value and highest R^2 value.

After that, I decided to delete all the insignificant variables in my model (the variables whose p-value was bigger than .05).

First, I removed ERG 7 because it was the most insignificant predictor variable. Then, I removed MSAT because it was insignificant in my model. The linear regression model that I was left was with: A screenshot of a computer

Description automatically generated with medium confidence

The multiple linear regression model that I was left with was:

lm(formula = X1st\_yr\_GPA ~ WSAT + HSRank + East + Housing + FGEN + HSGPA + Stem + ERG3 + ERG4 + ERG8, data = MidTermDataTrain)

I was not satisfied with the predictor variables that R decided where significant in this dataset. I was left in awe that going to a high school who was ranked #1 (ERG 1) stopped being significant in my model. In my opinion, it would’ve made sense to have ERG 1 & 8 in the same model because these are opposite ends of the scale from each other. It did not make sense to me how ERG 3 & 4 where significant. I decided to remove ERG 3 & 4. I also decided to remove HS rank from the selection because of how highly correlated it is with HSGPA. Both of those predictor variables are in the model, so I decided to stick with HSGPA. This is the model I decided to stick with:

A screenshot of a computer

Description automatically generated with low confidence

lm(formula = X1st\_yr\_GPA ~ WSAT + East + Housing + FGEN + HSGPA + Stem + ERG8, data = MidTermDataTrain)

## Explaining my Regression model

My regression model states:

A student’s 1st year GPA is predicted to be -.8717925 (0) when all our predictor variables in our model are 0. For every 1 increase in a student’s WSAT score, there 1st GPA will increase by .0014681. When a student lives in the East side of CT, there estimated 1st year GPA is predicted to increase by .1580384. If a student lives on campus, there predicted 1st year GPA increases by .3715278. When a student is classified as “First generation” they’re GPA is predicted to decrease by .1125036. For every 1 increase in a student HSGPA, there predicted 1st year GPA is predicted to increase by .8320443. If a student declares to be interested in stem, there predicted 1st year GPA is estimated to decrease by .2302108. Lastly, when a student’s high school rank is 8, there estimated GPA decreases by .3228736.

This regression makes sense to me because: since this is a liberal arts school, it would make sense that students with higher WSAT grade are predicted to have a higher GPA because the majors that they can choose from can involve writing since it is a liberal arts school. It makes sense that a student that lives on the east side of Connecticut (closer to Eastern) also has an increase in their GPA. They’re closer to home so they have a lot mor resources than out of state or students living further away. It also makes sense that students living is significant when a 1st GPA is higher. A student living on campus does not have to worry about traveling and to be frank, first year students are required to have the most expensive hurley pass so they don’t have to worry about food, and they also have all the resources they need at school. Commuter students might find it difficult to achieve a good GPA having to drive everywhere and having a reduction in time and convince to do homework and meet with professors during office hours. It also makes sense that first generation students might see a decrease in their GPA. First generation students must learn everything regarding college by themselves. They don’t have a guidance at home, and they could honestly not be receiving the support that non-first gen students receive. First gen students probably must navigate how to afford college by themselves and focus on other things like scholarships as well. It also makes sense to me that students interested in Stem have less predicted 1st year GPAs… These students are going to liberal arts school to receive Stem education. This could mean that they were possibly denied from stem schools because of their past grades and they’re trying to give it a shot at whatever school they were accepted into. This could also mean that the stem education provided in the liberal arts school was not up to par and student couldn’t do well in these classes.

## Evaluating my Regression model:

### R^2

The R^2 in my model was not as good as I would have wanted it to be. R^2 measures how much variability in the dependent variables can be explained the model. Adjusted R^2 also does this but, it considers how many predictor variables are in the regression model. I will give my adjusted R^2 because I made a multiple regression model (there was a numerous amount of predictor variables). My adjusted R^2 was 0.1857. This means that my about 18.6% of dependent variability can be explained by my regression model.

### Minmax

Using Min-Max as a form to see the accuracy of my model, I learned that my model has an accuracy of 76.97%.

# CART MODEL (Retention)

A picture containing timeline

Description automatically generated

My CART Model of retention shows a baseline model that states: 75% of the students in this model are retained.

Leaf node 2 states that if a student’s first year GPA is less than 1.3, 88% of those students are not retained. This accounts for 12% of our actual dataset.

Leaf node 24 states that 70% of students whose GPAs are between 1.3 and 1.7 and received more or had a WSAT of 435 where not retained. This accounted for 1% of our dataset.,

Leaf node 25 states that 71% of students whose GPAs are between 1.3 and 1.7 and received a WSAT score less than 435 where retained. This accounted for 0% of our dataset.

Leaf node 13 states that 68% of the students who has a GPA less than or equal to 1.6 where retained. This accounted for 1% of our data.

Leaf node 7 states that if a student’s GPA is more than 1.7 then 84% of those students were attained for a second year. This accounts for 86% of our dataset.

## CART Model: Evaluation

My Cart Model correctly classified retained students 84.19% of the time and incorrectly classified unretained students as retained 15.81% of the time

My model found 97% of all retained students.

My model found 42% of all students not retained.

My model incorrectly classifies students as retained 57.02% of the time

My model incorrectly classified unretained students as retained 2.14% of the time

# Academic Risk (Logistic Regression Model)

Even Though our CART Model showed us that 75% of people in our data set stay for another year when their GPA is bigger than 1.3, This doesn’t show us whether we should consider these students at academic risks or not. To determine a threshold for a student to be at academic risk I decided to research a little bit on ECSU’s 2021-22 Catalog. I learned this:

* Students need a 2.0 GPA to become a communications major
* BIO, EES, and Political Science majors need a 2.0 to graduate
* Students are dismissed from the majors: communications, BIS, Finance, if their GPA falls below a 2.5 for 2 consecutive semesters. (2.7 for Business Admin majors)
* Students need a 2.3 GPA to graduate with a Comp sci. or Psychology degree
* To enter the early childhood program or Elementary Teach program, students need a 3.0 GPA and need to maintain it to graduate.
* Individualized majors need a 2.7 to graduate
* P.E. majors need a 2.7 GPA to graduate.

I continued to research my domain and looked on bigfuture.collegeboard.org to see what letter grades correspond to GPA.

I was provided with this:

A+ 97-100 4.0

A 93-96 4.0

A- 90-92 3.7

B+ 87-89 3.3

B 83-86 3.0

B- 80-82 2.7

C+ 77-79 2.3

C 73-76 2.0

C- 70-72 1.7

D+ 67-69 1.3

D 65-66 1.0

F Below 65 0.0

After learning a bit about my domain, I decided to make my ‘At academic risk’ threshold at 1.7. I believe if a student has a C- in all his classes and his GPA is a 1.7, he should be considered at risk because there are some programs/majors that he will not be able to get into. Although there is a saying that says ‘C’s get degrees’ I don’t believe that a C- is acceptable. A 1.7 GPA prevents students from graduating from most of the majors listed above.

## Creating my logistic Regression Model (Stepwise)

To create my logistic regression model I decided to start off with using the MASS package so that I can use the AIC command… Again.

I did the same thing as I did in my multiple linear regression model: created an empty logistic regression model with only my target variable of AtRisk.

I also created my ‘Full’ logistic regression model with my target variable of ‘AtRisk’ with everything else included in my dataset as predictor variables.

I let R create a Forward, Backwards and 2 Both models using AIC.

These are my results:

### Forward Model

My Forward stepwise model looked like this:

This model has an AIC of 2149.9 and an R^2 of 0.08421308 (8.42%).Table

Description automatically generated

### Backward Model

My Forward stepwise model looked like this: Table

Description automatically generated

This model has an AIC of 2150.9 and an R^2 of 0.08723831 (8.72%).

### Both (starting with a full and empty logistic regression model)

My 1st both stepwise models looked like this: This model started off with a full logistic regression model. Table

Description automatically generated with medium confidence

This model has the same AIC as the backward model.

Table

Description automatically generatedMy 2nd both stepwise models looked like this: This model started off with a an empty logistic regression model.

This model has the same AIC as the forward model.

I decided to go with my forward regression model because it had the lowest AIC.

#### Nit Picking The model I chose

Now that I chose the forward stepwise model, I am going to get rid of all the insignificant coefficients in my model. The model I was left was is this:

glm(formula = AtRisk ~ HSGPA + Housing + ERG8 + Stem + ERG5 + East + ERG1 + FGEN, family = binomial, data = MidTermDataTrain)

After running a summary of my model, I noticed that ERG1 was the most insignificant to my model. I removed this from my model and ran another summary of my model with ERG1

This model now had all significant coefficients but, I didn’t think that it was correct to have only ERG 5 and 8 in my model. I decided to get rid of ERG 5 because it’s a split down the middle for how good a school was rated. I believe that ERG5 being the only significant ‘mediocre’ score for a school is off.

The final model that I was left with after nit picking was this: glm(formula = AtRisk ~ HSGPA + Housing + ERG8 + Stem + East, family = binomial, data = MidTermDataTrain).

This model has an ‘R^2’ of 0.07572459 (7.57%).

## Interpretation of Logistic Regression Model:

My logistic regression model and there coefficients look like this:

Estimate Std.

(Intercept) -2.9567

HSGPA 1.3172

HousingLives On Campus 1.1054

ERG81 -0.6387

StemInterested in Stem -0.5294

East1 0.3902

For HSGPA, The odds of a student being at Academic Risk Decrease by a factor of 3.732954 (exp(1.3172)) for every 1 increase in HSGPA.

For students that live on campus, The odds of a student being at Academic Risk Decrease by a factor of 3.020432 (exp(1.1054)) compared to when they don’t live on campus.

For students whose high schools where ranked 8, The odds of a student being at Academic Risk increases by a factor of 0.5279783 (exp(-0.6387)) compared to students whose high school was not ranked 8.

For students who are interested in stem, The odds of a student being at Academic Risk increases by a factor of 0.5889582 (exp(-0.5294)) compared to students who aren’t interested in stem.

For students who live on the East side of Connecticut, The odds of a student being at Academic Risk decreases by a factor of 1.477276 (exp(0.3902)) compared to students that live anywhere else.

I logistic regression model makes sense because, I believe that students who had lower GPAs in high school have a higher odd of being at academic risk in college than students with higher GPA’s. It also makes sense to me that if a student is living on campus, there risk of being at academic risk is a lot lower than a student who isn’t living on campus. This student has a lot more resources and presumably a lot less responsibilities and stressors affecting them. I also believe that if a student goes to a poorly ranked high school (the poorest one available), then the odds that they will be at academic risk are high than students who go to better rated high school’s because the better rated high schools probably prepared them better. It also makes sense that the students interested in stem also have higher odds of being at academic risk than the other students because stem majors tend to be more extensive and harder degrees than other degrees. I can see how living closer to Eastern can cause less odds of resulting at being at academic risk because these students are closer to home than other students. These students could have more guidance and family support compared to students not as close to home.

## Regression Model Accuracy

Using Min-Max my model has an accuracy of 0.7466062 (74.66%).

I was going to use the ROCR package to make a table to make a ROCR curve plot so I could get my best True positive and False Positive rate to create a cut off to create a table to see the accuracy but, I got lost and confused and this report needs to be handed in.