STEM Retention:

What factors influence students' decisions
to stay in or leave STEM majors?

Jessica Linford
Texas A&M University
April 23, 2022

Cleaning and Filtering the Data

Our first task upon receiving the data was to understand the variables, deal with missing values, and clean the data set for analysis. The original data set consisted of 57,635 observations for 352 variables, including several duplicated variables and breakdowns of GPA and major by term from 2003 to 2016.

We were not sure whether all the students in the data set had been involved in STEM majors at some point, so we first needed to create a list of which majors were STEM majors. The data set had a column which categorized the final major for each student as either STEM or Not STEM, so we used that as a starting point. We then scanned every column in the data set that included majors and created a list of major codes that were not in the final major column. To decide whether these majors should be categorized as STEM, we used the following sources: Texas A&M course catalogs dating back to 2003[1] and lists of majors classified as STEM by the National Science Foundation[2] and the Department of Homeland Security[3]. For each major on our unknown major list, we started by searching for the major code in old course catalogs. In many cases, we found that the code corresponded to a major that was already classified, but the major code had changed over time. In the other cases, we consulted the NSF and DHS lists to make a categorization. Once all majors had been classified, we compiled a complete list of STEM majors. We then scanned through all the majors listed for each student. Because we were interested in determining the main factors involved in STEM retention, we dropped students from the data set if none of the majors for the student were classified as STEM. After removing these students and students who did not graduate, we were left with 33,497 students remaining in the dataset.

Our response variable for the models is the classification of the final major as STEM or Not STEM. We decided to use the following explanatory variables: ethnic group, gender, cumulative GPA, total transfer hours, hours registered during 1st Term, SAT critical reading score, SAT math score, Pell grant eligibility, first generation student (Yes/No), admission type (freshman or transfer), total semesters, first semester GPA, and initial major category (STEM/Not STEM). First semester GPA and initial major category were new variables created from the breakdowns of GPA and major by term.

Because there were some unusual categories of ethnic groups, we combined the 565 students originally coded as “Other (Collapsed)” and the 261 students originally coded as “International” under the label “Other”. Similarly, we addressed unusual admission categories by recoding the 136 students

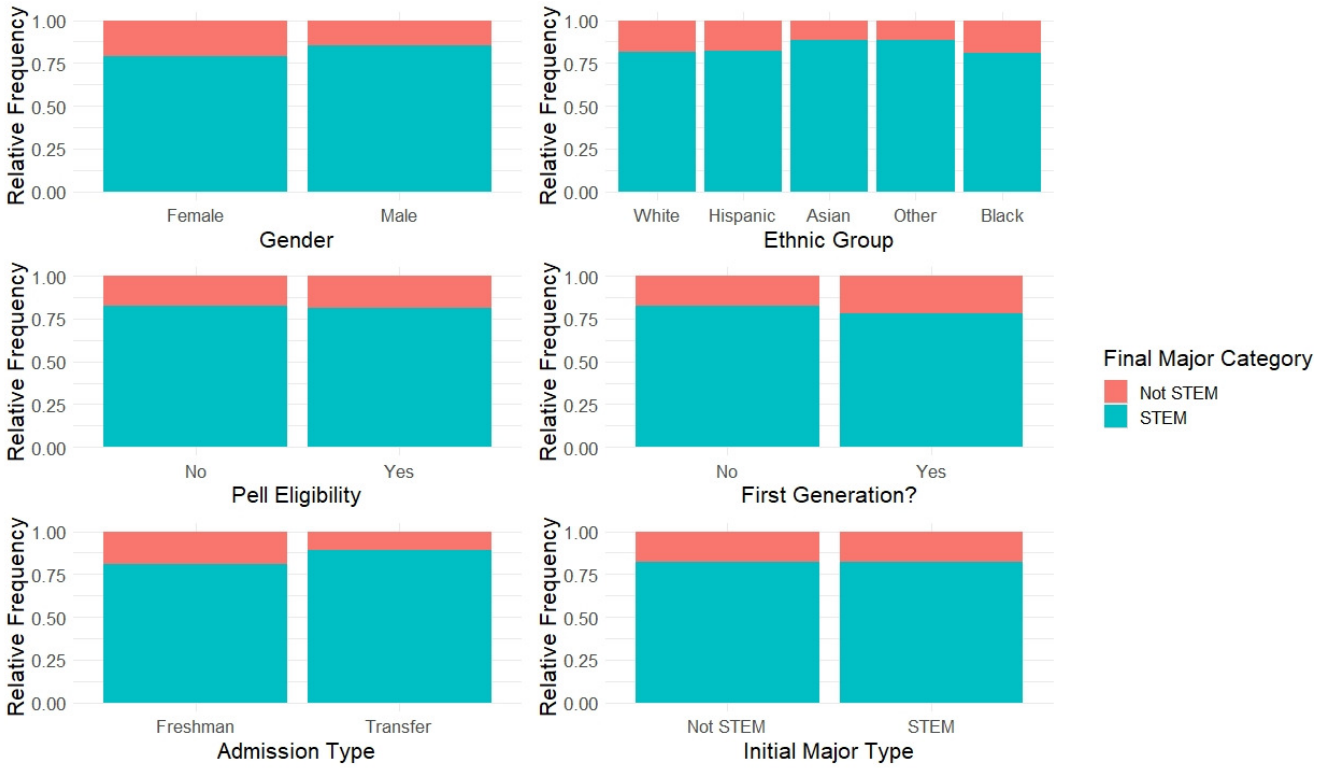
[1] https://regcatalogarchives.tamu.edu/ and https://catalog.tamu.edu/archives/
[2] https://www.ccm.edu/wp-content/uploads/2019/03/NSF-STEM-Classification-of-Instructional-Programs.pdf
[3] https://www.ice.gov/sites/default/files/documents/stem-list.pdf

originally coded as “International Transfer” as “Transfer” and the 170 students originally coded as “International Freshman” as “Freshman”.

Finally, we removed students with missing values of any of our chosen variables from the data set. There were 1,335 students with missing admission types and 4,857 students with missing SAT scores. We also removed 46 students with the unusual admission types “Katrina”, “Summer Transient”, and “Postbacc Undergraduate” and two students with impossible SAT reading and math scores (579 and 960, respectively). After removing these students, we were left with a final dataset with 27,410 observations of 14 variables.

Exploratory Data Analysis – Categorical Variables

The table and graphs on the following page show a comparison of the proportions of students who graduated as STEM majors for the different levels of each categorical explanatory variable. The differences in proportions were not large for any of the variables. We found that males are slightly more likely to stay in STEM majors than females (85% for males compared to 79% for females). For the different ethnic groups, those categorized as “Asian” or “Other” are most likely to stay in STEM (88% for both groups), while about 82% of Whites, 82% of Hispanics, and 81% of Blacks stay in STEM. Pell eligibility appears to be mostly unrelated to STEM retention, as there was little difference in the percentages of non-Pell-eligible and Pell-eligible students who stuck with STEM majors (82% vs. 81%). Those who were not 1st generation students were more likely to finish a STEM major than 1st generation students (83% vs. 78%). Transfer students are also more likely to stay in STEM than those who enter as freshmen (89% vs. 81%). It seems to make little difference whether students begin in STEM or switch to a STEM major later in their college journeys. Both groups ended up with 82% of students graduating with STEM majors.

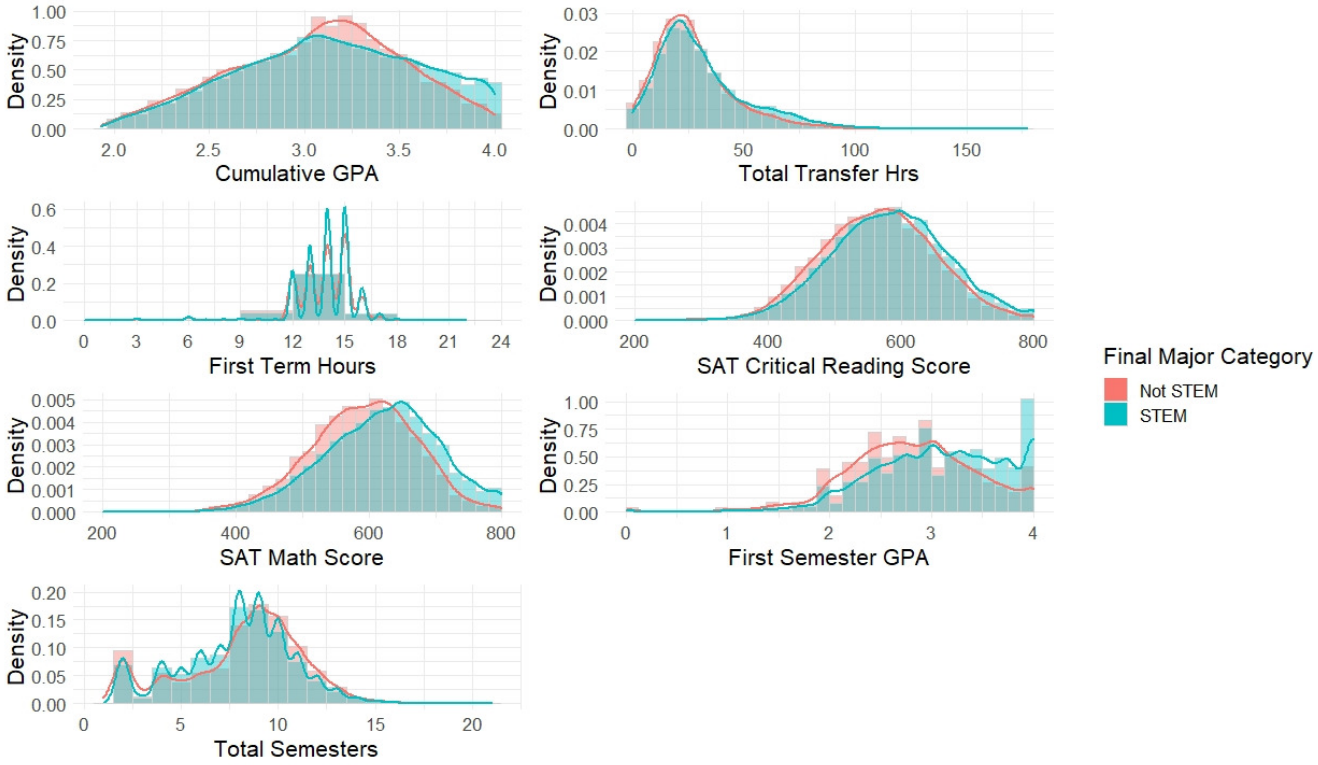

Relative Frequency
1.00
0.75
0.50
0.25
0.00
Female
Male
Gender
White
Hispanic
Asian
Other
Black
Ethnic Group
No
Yes
Pell Eligibility
First Generation?
Freshman
Transfer
Admission Type
Not STEM
STEM
Initial Major Type
Final Major Category
Not STEM
STEM

	Final Major Category		
Factor	**STEM**	**Not STEM**	**Total**
Gender			
Female	10,061 (79%)	2,717 (21%)	n = 12,778
Male	12,477 (85%)	2,155 (15%)	n = 14,632
Ethnic Group			
White	16,406 (82%)	3,690 (18%)	n = 20,096
Hispanic	3,394 (82%)	749 (18%)	n = 4,143
Asian	1,366 (88%)	184 (12%)	n = 1,550
Other	730 (88%)	96 (12%)	n = 826
Black	642 (81%)	153 (19%)	n = 795
Pell Eligible?			
No	20,898 (82%)	4,499 (18%)	n = 25,397
Yes	1,640 (81%)	373 (19%)	n = 2,013
First Generation Student?			
No	20,988 (83%)	4,443 (17%)	n = 25,431
Yes	1,550 (78%)	429 (22%)	n = 1,979
Admission Type			
Freshman	18,628 (81%)	4,391 (19%)	n = 23,019
Transfer	3,910 (89%)	481 (11%)	n = 4,391
Initial Major Category			
STEM	21,270 (82%)	4,600 (18%)	n = 25,870
Not STEM	1,268 (82%)	272 (18%)	n = 1,540

Exploratory Analysis – Quantitative Variables

We also compared the distributions of the quantitative variables for students who graduated in STEM vs. those who did not. There were small differences between the two groups; most notably, mean SAT reading and math scores were both higher for the group who graduated in STEM (reading = 583.4, math = 621.6) than for the non-STEM graduates (reading = 566.3, math = 591.9). The mean first semester GPA was also higher for the STEM group (3.1 vs. 2.8 for non-STEM). Those who graduated in STEM had a slightly higher median number of transfer hours (26) than those who did not (24). Finally, the non-STEM group took slightly longer, on average, to finish than the STEM group (mean 8.2 semesters for the non-STEM group vs. 7.9 semesters for the non-STEM group).

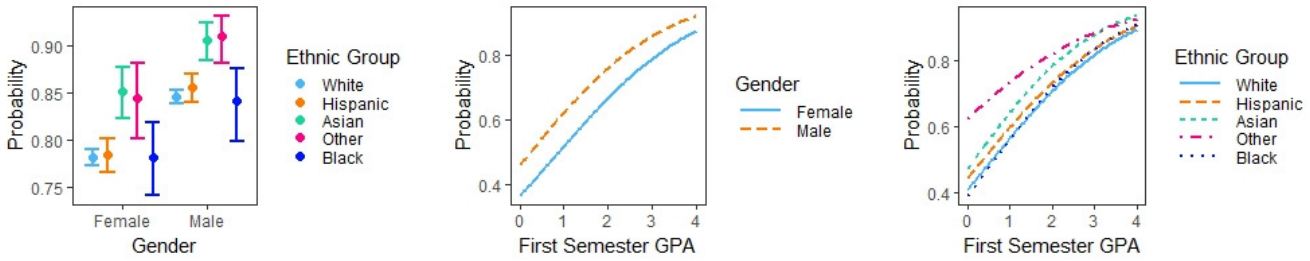
The histograms and table below summarize the differences in the quantitative variables between the STEM graduates and the non-STEM graduates.



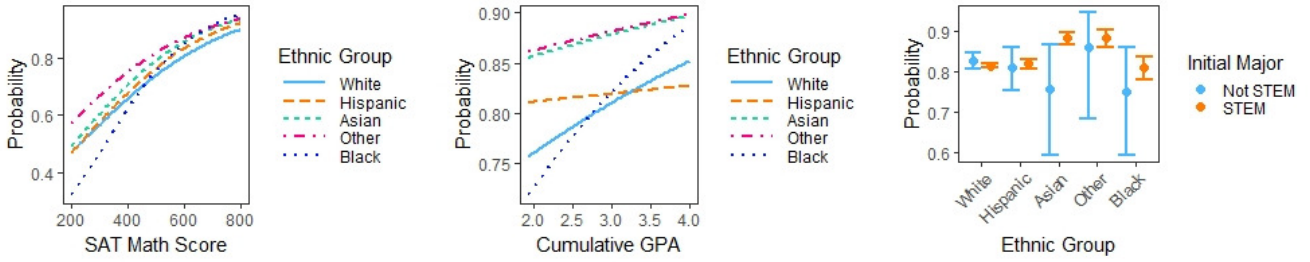
	Final Major Category	
Factor	**STEM**, N = 22,538	**Not STEM**, N = 4,872
Cumulative GPA		
Median (IQR)	3.1 (2.8, 3.5)	3.1 (2.8, 3.4)
Mean (SD)	3.1 (0.5)	3.1 (0.4)
Transfer Hours		
Median (IQR)	26.0 (17.0, 40.0)	24.0 (15.0, 35.0)
Mean (SD)	30.9 (20.1)	27.1 (16.9)
First Semester Hours		
Median (IQR)	14.0 (13.0, 15.0)	14.0 (13.0, 15.0)
Mean (SD)	13.9 (1.8)	13.8 (1.8)
SAT Reading Score		
Median (IQR)	580.0 (520.0, 640.0)	570.0 (510.0, 620.0)
Mean (SD)	583.4 (86.5)	566.3 (84.3)
SAT Math Score		
Median (IQR)	630.0 (560.0, 680.0)	600.0 (540.0, 650.0)
Mean (SD)	621.1 (85.0)	591.9 (79.6)
First Semester GPA		
Median (IQR)	3.1 (2.6, 3.6)	2.8 (2.4, 3.2)
Mean (SD)	3.1 (0.6)	2.8 (0.6)
Total Semesters		
Median (IQR)	8.0 (6.0, 10.0)	9.0 (6.0, 10.0)
Mean (SD)	7.9 (2.9)	8.2 (3.1)

Exploratory Analysis: Interactions

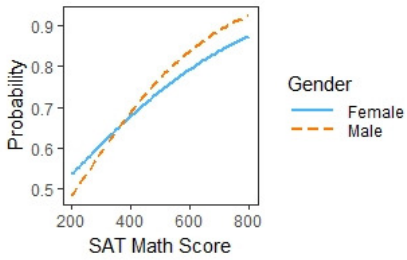
We suspected that gender and ethnic group might have interactions with GPAs, test scores, and whether students started in STEM. We also tested for an interaction between first semester category and first semester GPA. The plots below illustrate the interactions between the variables in terms of their effects on the probability of graduating with a STEM degree.

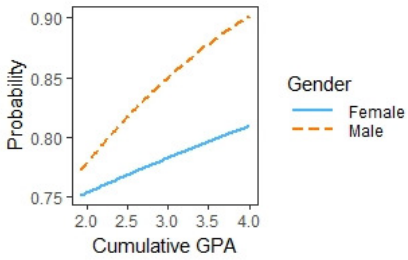


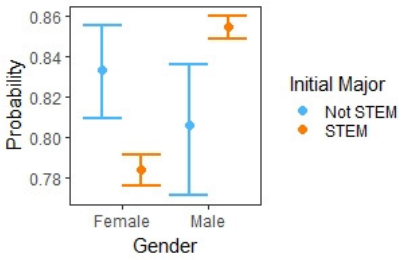
We found no significant interactions between gender and ethnic group. While males in each ethnic group were more likely to graduate in STEM majors than females, ethnic groups with higher probabilities of graduating in STEM had higher probabilities for both males and females. We also did not observe significant interactions between first semester GPA and either gender or ethnic group. As first semester GPA increased, the probability of graduating with a STEM degree increased at a similar rate for all students.



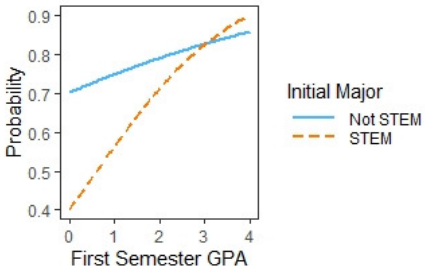
The significant interactions we observed with ethnic group were specific to one ethnicity. There was a significant interaction between ethnic group and SAT math score for Black students. The probability of graduating in STEM increased more quickly as SAT math score increased for Black students than for students of other ethnic groups. We also observed a significant interaction between cumulative GPA and ethnic group for Hispanics. There was very little change in the probability of finishing in STEM for Hispanics as cumulative GPA increased, but for the other ethnic groups, the probability of staying in STEM increased with cumulative GPA. Finally, we observed a significant interaction between starting in a STEM major and ethnic group for Asian students. For the other ethnic groups, the probability of graduating in STEM was roughly the same for those who started in STEM and for those who transferred to STEM later in their college careers. For Asians, the probability of ending up in a STEM major was higher for those who started in STEM than for those who switched to STEM later.







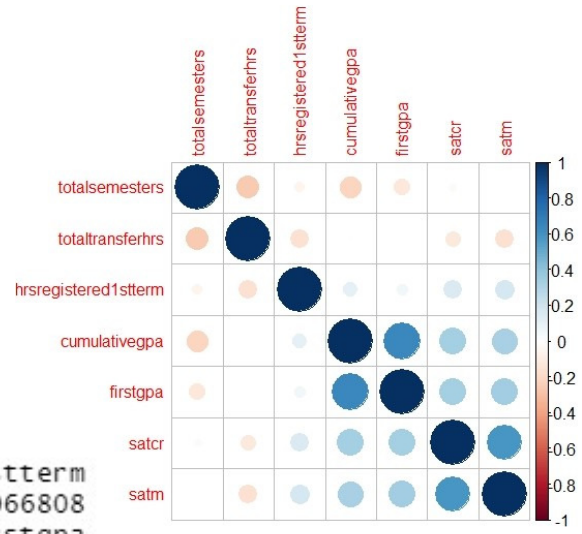
There were three significant interactions involving gender. The probability of staying in STEM increased faster with higher SAT math scores and with higher cumulative GPA for males than for females. Also, whether a student started in STEM had an opposite effect on males and females. For females, those who switched to STEM later had a higher probability of graduating with a STEM degree than those who started in STEM. For males, the opposite was true: starting with a STEM major increased the probability of finishing in STEM.



The final significant interaction we observed was between first semester GPA and initial major category. For those who started in STEM majors, a low first semester GPA was associated with a low probability of staying in STEM, and as first semester GPA increased, the probability of graduating in STEM increased dramatically. For those who transferred to STEM later, a low first semester GPA was only moderately less likely to result in a STEM degree than a high first semester GPA. This makes a lot of sense, because those who start in a STEM major and do poorly would be much more likely to leave STEM than those who transfer into STEM after doing poorly in a non-STEM major.

Models

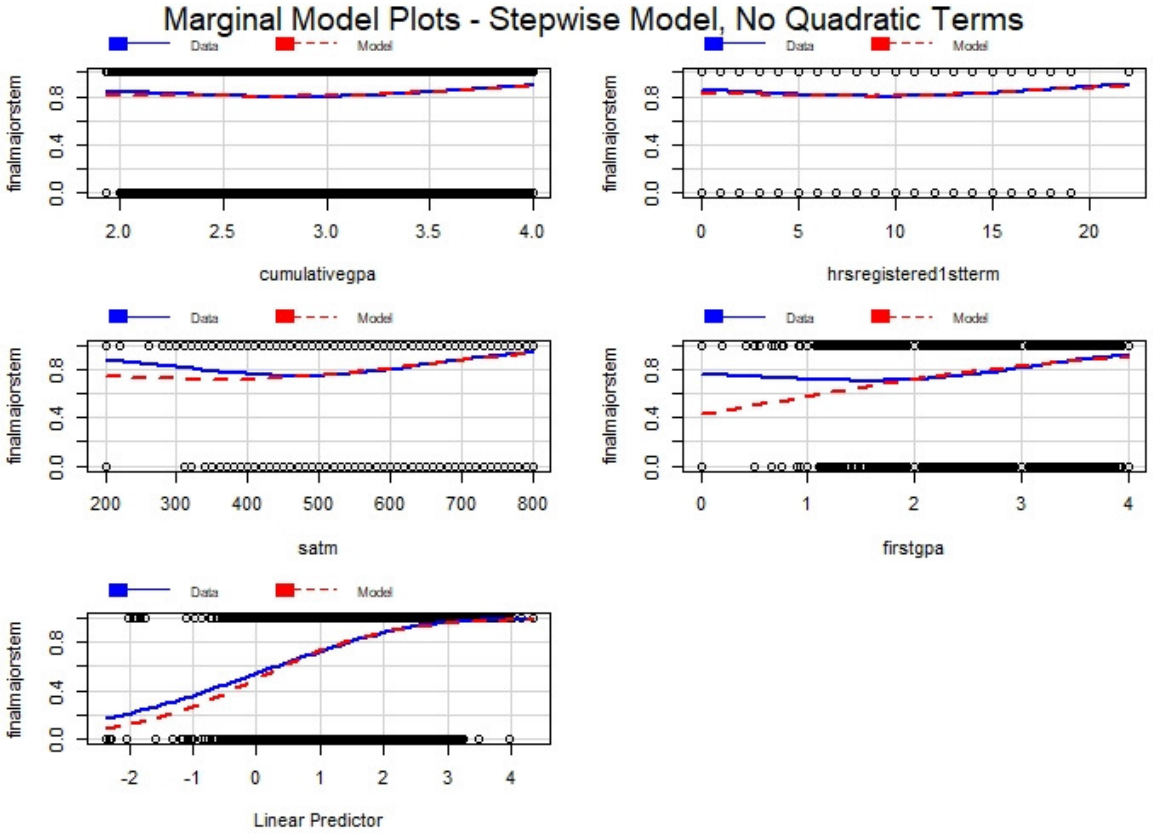
Before creating models, we checked for collinearity between the quantitative variables using a correlation plot and variance inflation factors. There were moderate correlations between first semester GPA and cumulative GPA and between SAT math score and SAT reading score, but the variance inflation factors were all well below 5, so none of the variables were correlated enough to cause issues with model development.

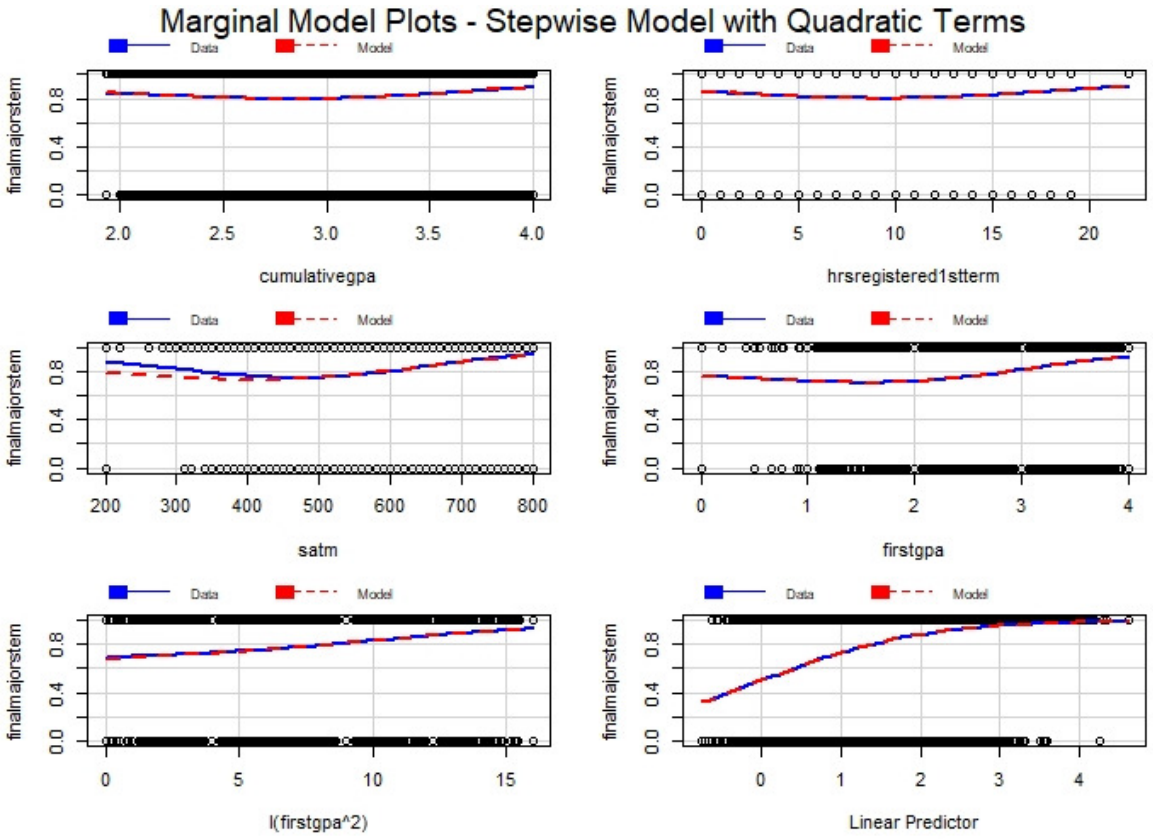



```
 cumulativegpa     totaltransferhrs hrsregistered1stterm
      1.785394             1.154519             1.066808
          satcr                 satm            firstgpa
      1.556345             1.579384             1.712929
totalsemesters
      1.137066
```

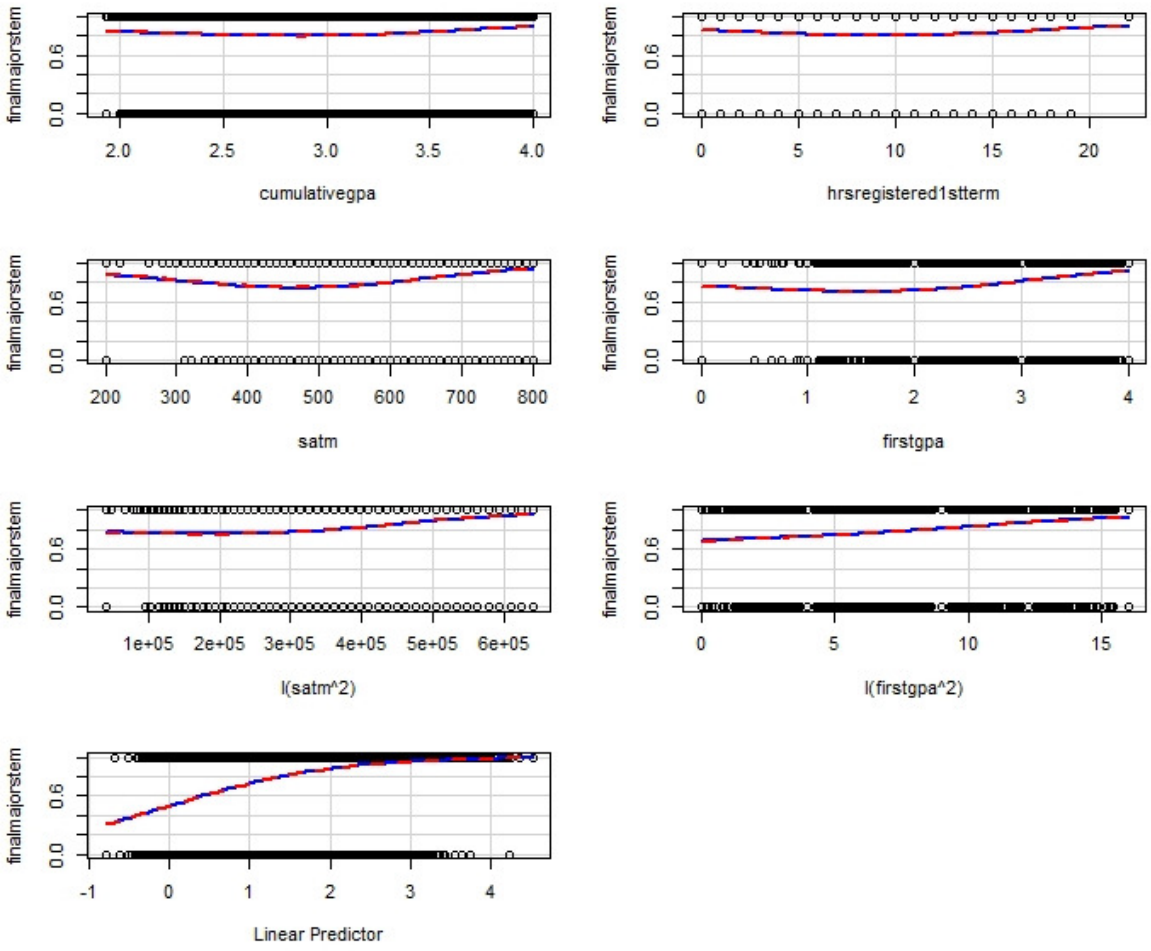
The data set was split into a training set and a testing set using a 75%/25% split. The training set was used to develop two different models. The testing set was then used to evaluate the models based on sensitivity, specificity, and balanced accuracy. The first model was developed using stepwise variable selection with the stepAIC function from the MASS package in R. Starting with a model containing all variables and interaction terms, variables were removed or readded to the model based on which removal or addition would lower BIC the most. BIC was chosen as the criterion because it penalizes more complicated models and simpler models are less likely to suffer for over-fitting. The second model was developed using a best-subsets approach with the bess function from the BeSS library in R. The best model was selected by minimizing BIC.

Marginal model plots were used to evaluate the fit of each model. The first attempt at creating a model showed that the models were a poor fit for SAT math score and first semester GPA. To remedy this, quadratic terms for these variables were added to the models (although the quadratic term for SAT math was selected out by the stepwise model). This addition greatly improved the model fits. The marginal model plots before and after the addition of the quadratic terms are shown on the following page.




Marginal Model Plots - Stepwise Model with Quadratic Terms
Data
Model
finalmajorstem
cumulativegpa
hrsregistered1stterm
satm
firstgpa
I(firstgpa^2)
Linear Predictor

Marginal Model Plots - Best Subsets Model with Quadratic Terms



Model Details

Similar variables were selected with stepwise selection and a best-subsets approach. A table comparing the variables selected by the two methods is shown below, followed by the model summaries for the two models. For the best subsets model, the bess function did not select the variables ethnicgroupHispanic, ethnicgroupAsian, or satm, but these variables were added back to the model because they were part of quadratic or interaction terms that were selected.

Variables Selected in Each Model

Term	Stepwise	Best Subsets
ethnicgroupHispanic	x	*
ethnicgroupAsian	x	*
ethnicgroupOther	x	
ethnicgroupBlack	x	
cumulativegpa	x	x
hrsregistered1stterm	x	x
satm	x	*
studentadmitcodeTransfer	x	x
genderMale	x	x
major1stemSTEM	x	x
firstgpa	x	x
I(satm^2)		x
I(firstgpa^2)	x	x
satm:genderMale		x
cumulativegpa:genderMale	x	x
genderMale:firstgpa	x	x
genderMale:major1stemSTEM	x	
ethnicgroupHispanic:major1stemSTEM		x
ethnicgroupAsian:major1stemSTEM		x
major1stemSTEM:firstgpa	x	x

*Added back to model due to inclusion in a quadratic or interaction term

Both models selected cumulative GPA, 1st semester hours, admission type, gender, initial major category, 1st semester GPA, (1st semester GPA)^2, the interaction between SAT math score and gender, the interaction between 1st semester GPA and gender, and the interaction between initial major category and 1st semester GPA. Variables unique to the stepwise model were ethnic group and the interaction between gender and initial major category, while the best subsets model uniquely selected (SAT math score)^2, the interaction between SAT math score and gender, and the interaction between ethnic group and initial major category for Hispanics and Asians.

Stepwise Model:

```
glm(formula = finalmajorstem ~ ethnicgroup + cumulativegpa +
    hrsregistered1stterm + satm + studentadmitcode + gender +
    major1stem + firstgpa + I(firstgpa^2) + cumulativegpa:gender +
    gender:firstgpa + gender:major1stem + major1stem:firstgpa,
    family = "binomial", data = stemTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9192   0.3445   0.5106   0.6679   1.4479

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               1.7555207  0.4921031   3.567 0.000361 ***
ethnicgroupHispanic       0.2605759  0.0544145   4.789 1.68e-06 ***
ethnicgroupAsian          0.4300761  0.0974628   4.413 1.02e-05 ***
ethnicgroupOther          0.2887690  0.1276282   2.263 0.023662 *
ethnicgroupBlack          0.2870310  0.1086781   2.641 0.008263 **
cumulativegpa            -0.9387839  0.0786689 -11.933  < 2e-16 ***
hrsregistered1stterm      0.0455339  0.0109813   4.147 3.38e-05 ***
satm                      0.0039739  0.0002769  14.350  < 2e-16 ***
studentadmitcodeTransfer  1.0295256  0.0648736  15.870  < 2e-16 ***
genderMale               -1.7219124  0.3036835  -5.670 1.43e-08 ***
major1stemSTEM           -1.8492151  0.3711230  -4.983 6.27e-07 ***
firstgpa                 -0.9127680  0.1992776  -4.580 4.64e-06 ***
I(firstgpa^2)             0.2622718  0.0294869   8.895  < 2e-16 ***
cumulativegpa:genderMale  0.6552810  0.1073406   6.105 1.03e-09 ***
genderMale:firstgpa      -0.2480338  0.0759434  -3.266 0.001091 **
genderMale:major1stemSTEM 0.6987934  0.1622813   4.306 1.66e-05 ***
major1stemSTEM:firstgpa   0.4945570  0.1222307   4.046 5.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19240  on 20557  degrees of freedom
Residual deviance: 17830  on 20541  degrees of freedom
AIC: 17864
```

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 1.76 + 0.26I_{\text{Hispanic}} + 0.43I_{\text{Asian}} + 0.29I_{\text{Other}} + 0.29I_{\text{Black}} - 0.94(\text{cumulative GPA})$$
$$+ 0.05(\text{1st semester hours}) + 0.004(\text{SAT math score}) + 1.03I_{\text{transfer}} - 1.72I_{\text{male}}$$
$$- 1.84I_{\text{original major STEM}} - 0.91(\text{1st semester GPA}) + 0.26(\text{1st semester GPA})^2$$
$$+ 0.66(\text{cumulative GPA})(I_{\text{male}}) - 0.24(\text{1st semester GPA})(I_{\text{male}})$$
$$+ 0.70(I_{\text{male}})(I_{\text{original major STEM}}) + 0.49(\text{1st semester GPA})(I_{\text{original major STEM}})$$

$\hat{p}$ = predicted probability of graduating with a STEM degree

$I_x = 1$ if student is in category x, 0 otherwise

Best Subsets Model:

```
glm(formula = finalmajorstem ~ ethnicgroupHispanic + ethnicgroupAsian +
    cumulativegpa + hrsregistered1stterm + satm + studentadmitcodeTransfer +
    genderMale + major1stemSTEM + firstgpa + I(satm^2) + I(firstgpa^2) +
    satm:genderMale + cumulativegpa:genderMale + genderMale:firstgpa +
    ethnicgroupHispanic:major1stemSTEM + ethnicgroupAsian:major1stemSTEM +
    major1stemSTEM:firstgpa, family = "binomial", data = XyTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9177   0.3394   0.5109   0.6710   1.4796

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)
(Intercept)                       3.554e+00  8.476e-01   4.193 2.75e-05 ***
ethnicgroupHispanic              -6.518e-02  2.283e-01  -0.286 0.775239
ethnicgroupAsian                 -7.126e-01  4.313e-01  -1.652 0.098516 .
cumulativegpa                    -9.176e-01  7.888e-02 -11.633  < 2e-16 ***
hrsregistered1stterm              4.639e-02  1.094e-02   4.239 2.24e-05 ***
satm                             -3.013e-03  2.575e-03  -1.170 0.241965
studentadmitcodeTransfer          1.011e+00  6.501e-02  15.547  < 2e-16 ***
genderMale                       -1.698e+00  3.508e-01  -4.840 1.30e-06 ***
major1stemSTEM                   -1.491e+00  3.614e-01  -4.124 3.72e-05 ***
firstgpa                         -8.256e-01  1.980e-01  -4.170 3.05e-05 ***
I(satm^2)                         5.255e-06  2.202e-06   2.386 0.017011 *
I(firstgpa^2)                     2.594e-01  2.963e-02   8.754  < 2e-16 ***
satm:genderMale                   1.521e-03  5.270e-04   2.886 0.003903 **
cumulativegpa:genderMale          6.053e-01  1.089e-01   5.560 2.70e-08 ***
genderMale:firstgpa              -2.927e-01  7.701e-02  -3.801 0.000144 ***
ethnicgroupHispanic:major1stemSTEM  3.103e-01  2.345e-01   1.323 0.185843
ethnicgroupAsian:major1stemSTEM    1.163e+00  4.426e-01   2.629 0.008564 **
major1stemSTEM:firstgpa            4.367e-01  1.222e-01   3.574 0.000351 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19240  on 20557  degrees of freedom
Residual deviance: 17835  on 20540  degrees of freedom
AIC: 17871
```

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 3.55 - 0.07 I_{\text{Hispanic}} - 0.71 I_{\text{Asian}} - 0.92(\text{cumulative GPA}) + 0.05(\text{1st semester hours})$$
$$- 0.003(\text{SAT math score}) + 1.01 I_{\text{transfer}} - 1.70 I_{\text{male}} - 1.49 I_{\text{initial major STEM}}$$
$$- 0.83(\text{1st semester GPA}) + 5.3\times10^{-6}(\text{SAT math score})^2 + 0.26(\text{1st semester GPA})^2$$
$$+ 0.002(\text{SAT math score})(I_{\text{male}}) + 0.61(\text{cumulative GPA})(I_{\text{male}})$$
$$- 0.29(\text{1st semester GPA})(I_{\text{male}}) + 0.31(I_{\text{Hispanic}})(I_{\text{initial major STEM}})$$
$$+ 1.16(I_{\text{Asian}})(I_{\text{initial major STEM}}) + 0.44(\text{1st semester GPA})(I_{\text{initial major STEM}})$$

$\hat{p}$ = predicted probability of graduating with a STEM degree

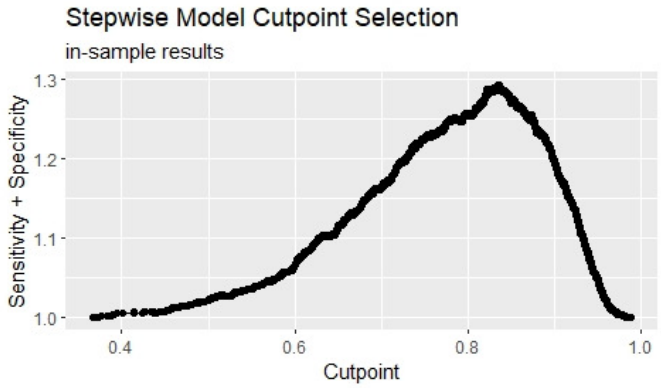
$I_x = 1$ if student is in category x, 0 otherwise

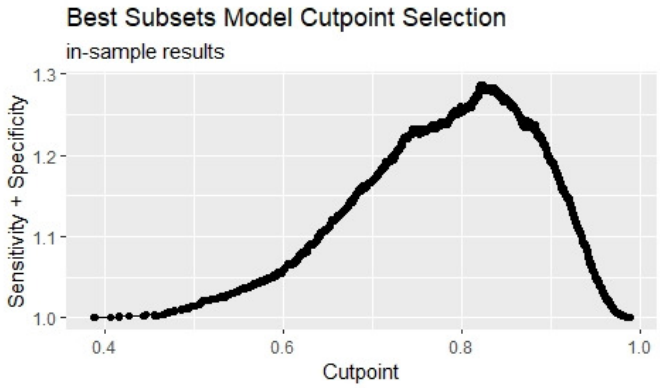
Model Performance

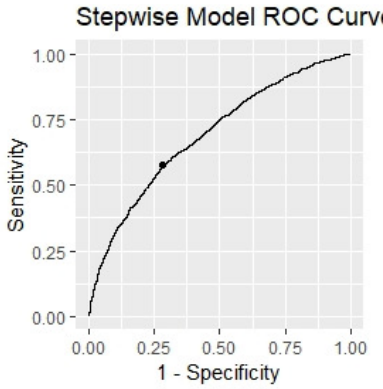
Initially, the models were used to predict whether students would graduate in STEM using a cutoff probability of 0.5. That is, those with a predicted probability of 0.5 or higher were predicted as STEM graduates, and those with predicted probabilities lower than 0.5 were predicted as non-STEM graduates. We quickly realized that this was not an effective method of classification, as it resulted in predicting that every student in the testing set would graduate with a STEM degree:

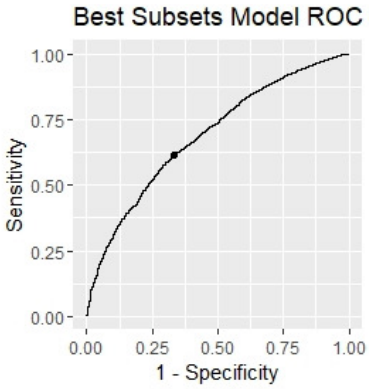
	Reference	
Prediction	STEM	Not STEM
STEM	5609	1186
Not STEM	0	0

While the STEM graduates were all classified correctly (100% sensitivity), none of the non-STEM graduates were (0% specificity). To address this problem, we decided to find the cutoff probability that would maximize the balanced accuracy (average of the sensitivity and the specificity). This was accomplished using the cutpointr package in R. This package produced the following curves, showing the cut points that maximize the sum of the sensitivity and the specificity (equivalent to maximizing the balanced accuracy) and the position of those cut points on ROC curves.









Model	Optimal_Cutpoint	Sensitivity	Specificity	Balanced_Accuracy	AUC
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Stepwise	0.8353734	0.5750799	0.7175698	0.6463248	0.6928640
Best Subsets	0.8229586	0.6173234	0.6699507	0.6436371	0.6908127

The two models are very similar in terms of optimal cut points, balanced accuracy, and AUC. While both models give some insight into factors influencing STEM retention, predictions from the models should not be considered very trustworthy because both models result in high rates of false positives and false negatives. The confusion matrices for the two models are shown below:

Stepwise model:

Prediction	Reference STEM	Reference Not STEM
STEM	3240	344
Not STEM	2394	874

Best subsets model:

Prediction	Reference STEM	Reference Not STEM
STEM	3478	402
Not STEM	2156	816

The stepwise model has a higher specificity and a lower sensitivity than the best subsets model. For students who graduate in STEM, the stepwise model makes a correct prediction 57.5% of the time and the best subsets model classifies the student correctly 61.7% of the time. For students who do not graduate with STEM degrees, the stepwise model is correct 71.8% of the time compared to 67.0% for the best subsets model. A common use of a model like this would be to identify students in danger of leaving a STEM major, which would place more importance on specificity than sensitivity. If the model is desired for this purpose, the stepwise model would be more useful than the best subsets model.

Model Interpretation

It is difficult to interpret model coefficients directly for the variables involved in quadratic and interaction terms. However, it makes sense to calculate and interpret odds ratios for the variables that only appear in one term. Odds ratios with 95% confidence intervals are shown in the tables below.

Odds Ratios for Stepwise Model

	OR	2.5 %	97.5 %
ethnicgroupHispanic	1.297677	1.167172	1.444718
ethnicgroupAsian	1.537374	1.274643	1.868246
ethnicgroupOther	1.334783	1.045705	1.725686
ethnicgroupBlack	1.332465	1.080400	1.654770
hrsregistered1stterm	1.046587	1.024154	1.069213
satm	1.003982	1.003438	1.004528
studentadmitcodeTransfer	2.799737	2.468355	3.183246

Odds Ratios for Best Subsets Model

```
                              OR    2.5 %   97.5 %
hrsregistered1stterm     1.047485 1.025111 1.070052
studentadmitcodeTransfer 2.747637 2.421902 3.125031
```

Based on the stepwise model, we are 95% confident that after controlling for other variables, the odds of graduating in STEM are:

- between 17% and 44% higher for Hispanics than for Whites
- between 27% and 87% higher for Asians than for Whites
- between 8% and 65% higher for Blacks than for Whites
- between 5% and 73% higher for other ethnic groups than for Whites
- between 2.5 and 3.2 times higher for transfer students than for freshmen
- between 2% and 7% higher for each additional 1st semester hour
- between 0.3% and 0.5% higher for each additional point on the SAT math section.

Based on the best subsets model, we are 95% confident that after controlling for other variables, the odds of graduating in STEM are:

- between 2.4 and 3.1 times higher for transfer students than for freshman
- between 3% and 7% higher for each additional 1st semester hour

The relative importance of each variable was determined by performing an ANOVA test comparing the complete model to a model missing all terms containing that variable. The difference in deviance between the two models was used to calculate a p-value for the null hypothesis that the terms do not improve the fit of the model. A small p-value indicates that the variable is important to the model. The relative importance of the variables in the two models is shown in the tables below.

Variable Importance – Stepwise Model

Variable <chr>	Deviance <dbl>	df <dbl>	p_val <dbl>
First Semester GPA	576.91700	4	0.000000e+00
Admission Type	290.41526	1	0.000000e+00
SAT Math Score	207.92050	1	0.000000e+00
Cumulative GPA	158.88306	2	0.000000e+00
Gender	94.68356	4	0.000000e+00
Ethnic Group	47.77108	4	1.053389e-09
First Major Category	37.83085	3	3.069407e-08
1st Semester Hours	16.70780	1	4.360144e-05

Variable Importance – Best Subsets Model

Variable <chr>	Deviance <dbl>	df <dbl>	p_val <dbl>
First Semester GPA	565.96281	4	0.000000e+00
Admission Type	279.75389	1	0.000000e+00
SAT Math Score	224.46499	3	0.000000e+00
Cumulative GPA	155.16649	2	0.000000e+00
Gender	80.45996	4	1.110223e-16
Ethnic Group	41.55778	4	2.060052e-08
First Major Category	26.45477	4	2.561795e-05
1st Semester Hours	17.45399	1	2.943456e-05

The order of importance of the variables is the same for the two models. Both identify 1st semester GPA, admission type, SAT math score, cumulative GPA, and gender as the five most important variables in predicting whether a student will graduate with a STEM degree. Ethnic group, first major category, and 1st semester hours are also significant, but slightly less important.

Conclusion

We experienced only moderate success in developing a model to predict whether a student involved in STEM will graduate with a STEM major or switch to a non-STEM major. There were only small differences observed between STEM graduates and non-STEM graduates in terms of the variables we considered. This resulted in models with only 64% - 65% balanced accuracy. We identified 1st semester GPA, admission type, SAT math score, cumulative GPA, and gender as the five most important variables in predicting STEM retention. However, the decision of whether to stay in STEM is clearly complicated, and likely depends on many variables that were not considered in this study. Possible data to consider in follow-up studies that may improve model accuracy could include information about family economic circumstances, student employment, and physical and mental health.