GPA as an Indicator of a Student's Graduation Chance

Jaeheuk Jung, Zi Qing Liang, Nishali Parikh, Jungmin Park
Group Hajim 1, DSC 383 University of Rochester
{jjung16, nparikh3, jpark178, zliang11}@ur.rochester.edu

1.Introduction

With the emphasis on higher education in society, a college degree has become essential in developing individuals' economic opportunity and America's competitiveness in the world. (U.S. Department of Education, 2015). Many more students, both conventional and non-conventional, are applying to postsecondary institutions for a degree that hopefully opens the door to two-thirds of job openings. In the Fall 2018 semester, approximately 20 million students were enrolled in a degree-seeking institution in the U.S (IPEDS, 2019). However, more than 40% of first-time full time college students in the nation do not graduate within 6 years. With parents overbearing the tuition burden, states providing students loans and grants, and the credential of an institution becoming tremendously weighted, University administrations face pressure to ensure a better college experience and future opportunities to prevent dropouts (Marcus, 2020).

Traditionally, GPA has always been deemed as an important metric to evaluate a student's performance and abilities, particularly with their first job applications. Some researchers have also found that First-year GPA offers a powerful indication of a student's chances of graduation (University of Illinois at Urbana-Champaign, 2016). To better understand if GPA can serve as an indicator of a student's graduation chance, we will be examining the correlation between GPA and graduation rate through the following aspects: (1) Can students receiving a disastrous semester (GPA <1.0) recover and successfully graduate? (2) Does double majoring or having additional minors affect a student's GPA and graduation chance? (3) How does the term-to-term GPA trend differ between students who graduate on time and those who don't? Ultimately, our goal is to identify whether GPA contributes to flagging students at potential risk of not graduating and develop a predictive model.

2. Dataset

2.1 Dataset Description

Our sponsor, the advising team from the University of Rochester College of Art and Science & Engineering, provided three separate datasets that describe the "demographics", "term", and "course" information about 20,389 individual students linked across by the unique "SubjectID".

The "Course" dataset [Table 1]includes details on the classes ("CRN" and "Subject Code" etc) that a student registered with their corresponding grades and credits earned for a specific semester. The semester is indicated by the attribute "Year Term ID" in the format of "year" followed by the semester code; 1 = Fall, 2 = Spring, 3 = Business School Winter, 4 = Summer. This dataset also recorded the dorm area that these individual students lived in.

The Demographics dataset [Table 2] gives an overview of the student's background with information on High School, SAT/ACT scores and admission type; it also included the finalized information on cumulative GPA, majors/minor/clusters for the individual's degree. To identify whether a student graduated successfully, it is indicated by the attribute "Degree Confer Date"; a non-NA entry means a student successfully graduated.

The Term dataset [Table 3], on the other hand, included all relevant features such as term hours and term points that go into the GPA calculation; it allows the analysis on how individual students perform on the term to term level.

Table 1. A preview of the dataset "Course"

	SubjectID	Year Term ID	Ps1 Timestat Code	Ps1 Ofcl Stat Flag	Ps1 Major1 Code	Ps1 Major2 Code	Ps2 Ofcl Stat Flag	Ps2 Major1 Code	Ps2 Major2 Code	Ps1 Class Code	CRN	College Code	Subject Code	Course No	Course Discipline Code	Parent CRN	Parent College Code	Parent Subject Code
0	172789163	20062.0	X	Y	PSY	.	NaN	NaN	NaN	2004	62723.0	1.0	SAB	397	OTHR	62723.0	1.0	SAB
1	172846340	20102.0	P	Y	PSC	.	NaN	NaN	NaN	2010	79703.0	1.0	PSC	202	SOCS	79703.0	1.0	PSC
2	172846340	20102.0	P	Y	PSC	.	NaN	NaN	NaN	2010	80012.0	1.0	PSC	291W	SOCS	80003.0	1.0	PSC
3	172857770	20162.0	F	Y	ENG	.	NaN	NaN	NaN	2016	38294.0	1.0	SAB	398A	OTHR	38294.0	1.0	SAB
4	172877093	19871.0	F	N	UNC	.	NaN	NaN	NaN	1988	16092.0	1.0	COG	201	NATS	16092.0	1.0	COG
5	172877093	19871.0	F	N	UNC	.	NaN	NaN	NaN	1988	27656.0	1.0	LIN	101	SOCS	27656.0	1.0	LIN

Table 2. A preview of the dataset "Demographics"

	SubjectID	Gender	Visa Code	Visa Desc	Citizen Type	Citizen Country Code	Citizen Country Name	First Generation Flag	2010 IPEDS Ethnicity Description	2010 Ethnicity Short Desc	Veteran Status Desc	Birthdate	HS City	HS State Prov	HS Postal Code
0	175693463	F	.	.	US CITIZEN	1NY	USA	N	White	WHITE - NO ETHNICITY REPORTED	NaN	1979-03-17	East Rochester	NY	14445
1	179236322	M	.	.	US CITIZEN	1MA	USA	N	White	WHITE - NO ETHNICITY REPORTED	NaN	1971-04-13	Lexington	MA	2421
2	181015370	M	.	.	US CITIZEN	1NY	USA	N	White	WHITE - NO ETHNICITY REPORTED	NaN	1972-11-11	Rochester	NY	14606
3	183070133	M	IM	IMMIGRANT PERMANENT RESIDENT	PERM RES	3GY	GUYANA	N	Unknown	NO RACE REPORTED - NO ETHNICITY REPORTED	NaN	1973-08-13	Brooklyn	NY	11203
4	183070133	M	IM	IMMIGRANT PERMANENT RESIDENT	PERM RES	3GY	GUYANA	N	Unknown	NO RACE REPORTED - NO ETHNICITY REPORTED	NaN	1973-08-13	Brooklyn	NY	11203
5	183070133	M	IM	IMMIGRANT PERMANENT RESIDENT	PERM RES	3GY	GUYANA	N	Unknown	NO RACE REPORTED - NO ETHNICITY REPORTED	NaN	1973-08-13	Brooklyn	NY	11203

Table 3. A preview of dataset "Term"

	SubjectID	Year Term ID	Term GPA	Term Hrs Earned	Term Hrs GPA	Term Points Earned	Cumul GPA	Cumul Hrs Earned	Cumul Hrs GPA	Cumul Points Earned	Term Transfer Hrs
0	172789163	20062.0	0.00	0.0	0.0	0.0	3.07	128.0	128.0	393.2	0.0
1	172846340	20102.0	3.30	8.0	8.0	26.4	2.70	128.0	77.0	207.8	26.0
2	172857770	20162.0	0.00	0.0	0.0	0.0	2.20	128.0	124.0	273.2	0.0
3	172857770	20171.0	0.00	0.0	0.0	0.0	2.20	132.0	124.0	273.2	4.0
4	172877093	19871.0	1.33	12.0	12.0	16.0	1.77	85.0	97.0	171.8	0.0
5	172877093	19872.0	0.00	0.0	0.0	0.0	1.77	85.0	97.0	171.8	0.0

2.2 Data Preprocessing

Data processing is required to conduct any analysis to avoid any misleading results - "trash in trash out"

1. Removal of invalid rows
    a. All three datasets have automatic generated random "SubjectID" for the empty rows, which are removed.
2. Dataset Merge
    a. The datasets are inner merged by "SubjectID"
3. Graduation classification
    a. Students are classified into 4 different graduating groups as the sponsor wants to explore the difference across the 4 groups.
    b. The first group - normal - are students who graduated within 8 semesters. The second group were those who graduated within 9~12 semesters and the third group were students who took 12+ semesters to graduate. Lastly, the fourth group is students who never graduated.
    c. The dataset is then condensed to include relevant columns such as SubjectID, Year Term, Term GPA, Cumulative GPA, Degree Awarded, and Degree Conferred etc.
    d. Classification is the backbone for data analysis in this project. We were able to find out the columns which were less valid through visualizing and testing the relationship between the classification and other data such as GPA or majors to better understand our data.

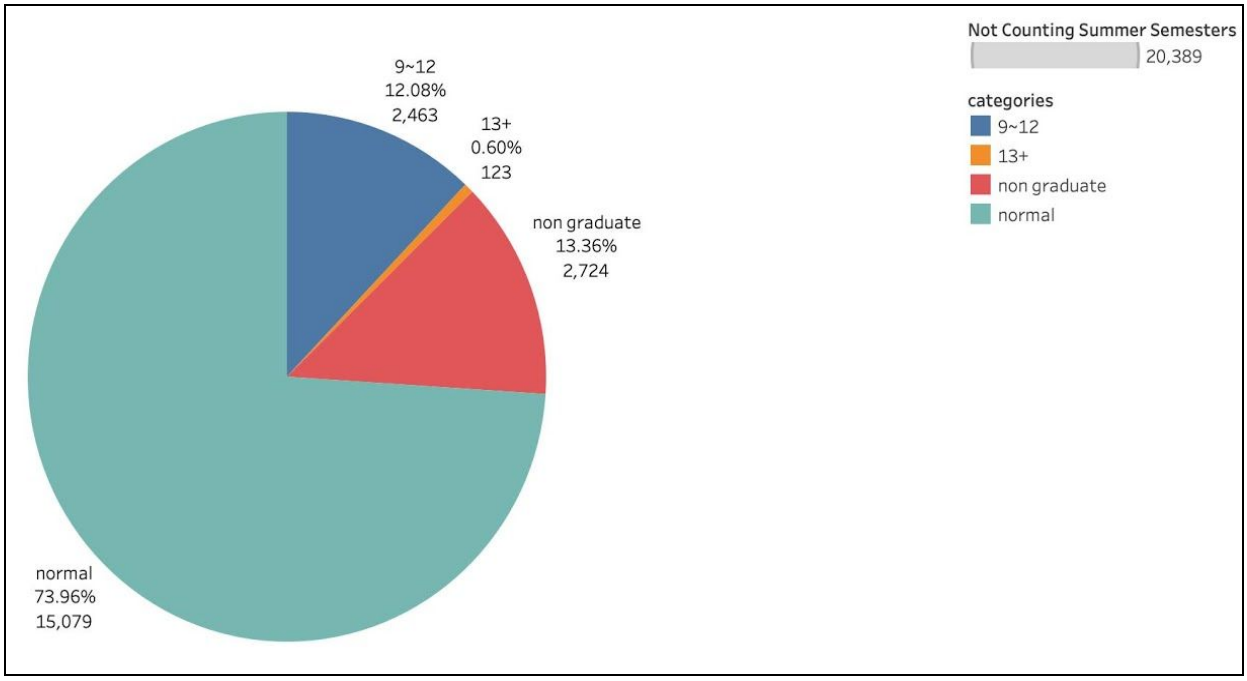
3. Exploratory Data Analysis

Our sponsor had strong interests in visualizing patterns surrounding features that are relevant to GPA and graduation classification to gain insights to potentially help with their advising strategies and to better identify the students at risk of not graduating on time. Therefore, data analysis with visualization is the particular emphasis for this project. We explored various features that our sponsors were most curious about to better understand how GPA relates to the graduating chance of a student - disastrous semesters, SAT/ACT scores, majors, and term-to-term trends etc.

3.1 Graduation Classification

After classifying students into 4 graduating groups in data preprocessing, the distribution is visualized in Chart 1.The normal student group consisted of 74% of the student population in the dataset, followed by 13% of non-graduate students. While the 9-12 semesters group took up 12%, the 12+ semesters group is less than 1%. From the breakdown, it is reasonable to suspect that if a student sees no potential in graduating sooner than 12 semesters, the student is very likely to not graduate instead. Additionally, these numbers aligned with the sponsor's initial knowledge on the graduation rate, acknowledging the accuracy of our classification.

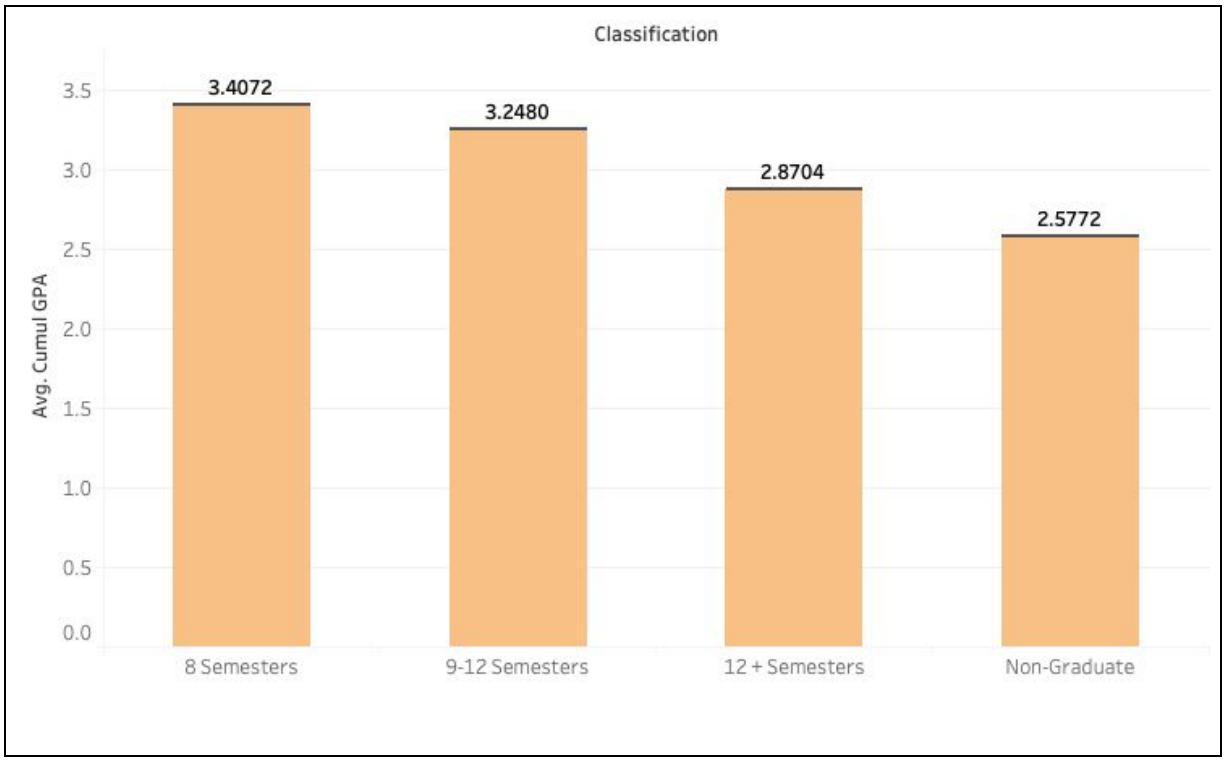
Chart 1: Pie chart showing the percentage of students belonging to each graduation group



3.2 Cumulative GPA

To explore whether GPA has an impact on the graduating rate, we visualized the average cumulative GPA of students belonging in each graduating classification. In Chart 2, a downward sloping in the GPA value is clearly observable as students moving from graduating on-time towards not-graduating. By conducting an ANOVA test and obtaining a p-value much lower than 0.01, we can conclude with 99% confidence level that the mean GPA among these 4 classified groups is statistically significantly different. The analysis also showed a positive relationship between GPA and graduating on-time (towards 8 semesters) with a coefficient of 0.74. This indicates that as a student's cumulative GPA goes up by 1, his/her chance of graduating within 8 semesters would go up by 0.74. This indicates that GPA has great importance and is very likely to contribute to predicting students' graduating group.

Chart 2. Average Cumulative GPA vs. Graduation Classification



3.3 SAT/ACT Scores

Standardized exam scores have always played an essential role in a student's college application process because it is a metric used by University admission to evaluate the academic potentials of students. But does it give any insights on a student's graduation chance? To find the answer, we provided a summary table and again ran an ANOVA test on the SAT/ACT scores with the graduating classification

According to Table 4, a difference of approx 30 points in Avg SAT and Median SAT between graduating 8 semesters and non-graduating is shown. With the statistical test, the coefficients (p-value < 0.01) show that if a student's SAT score goes up by 1, the chance of a student graduating within 8 semesters would go up by 0.0004, which is a very low correlation; if a student's ACT score goes up by 1, then the chance would go by 0.026 (ACT is much higher than SAT due to the scaling difference)..

Table 4. Summary of SAT/ACT scores based on graduation classification

< SAT/ACT>

Classification	Avg. ACT Composite	Avg. SAT Total	Median ACT Composite	Median SAT Total	Max. ACT Composite	Max. SAT Total	Min. ACT Composite	Min. SAT Total
8 Semesters	29	1,297	29	1,320	36	1,600	0	0
9-12 Semesters	28	1,282	29	1,300	36	1,600	13	0
12 + Semesters	28	1,265	29	1,290	34	1,540	14	0
Non-Graduate	28	1,264	28	1,290	35	1,600	14	0

In Chart 3 and 4, the distribution of scores among students in each classification group is very similar, not presenting any distinguishable patterns between those who graduated on time and those who did not. The peaks of each classification for ACT is around 30 and 1290 and 1350 for SAT, showing the range of exam scores that the University of Rochester likely admits. Technically speaking, there is a significant difference in the average SAT/ACT scores between the 4 graduating groups; but with such small differences and the trends observed in the charts, it is highly unlikely to help flag students graduating on-time or not graduate if it is solely through SAT/ACT scores.

Chart 3. Distribution graph of students in 4 classification groups with respect to SAT scores

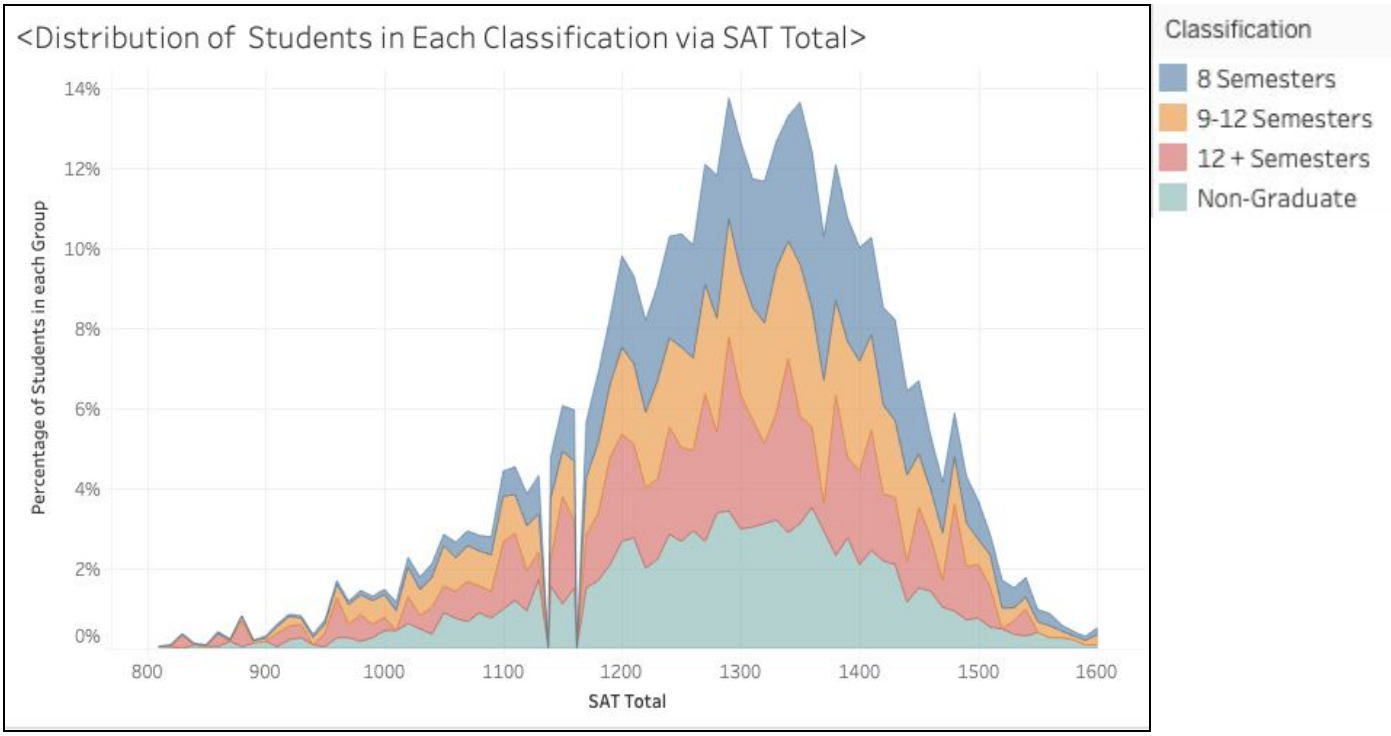
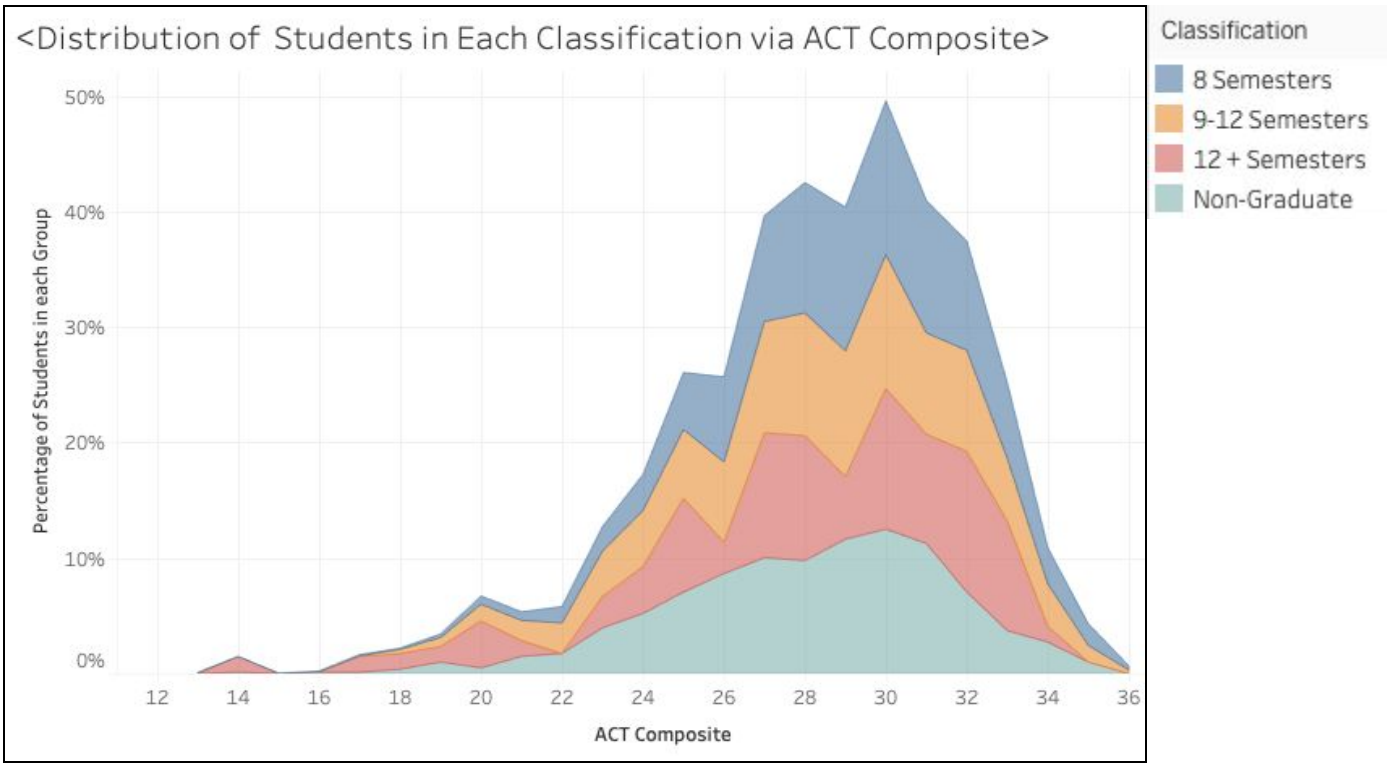


Chart 4. Distribution graph of students in 4 classification groups with respect to ACT scores



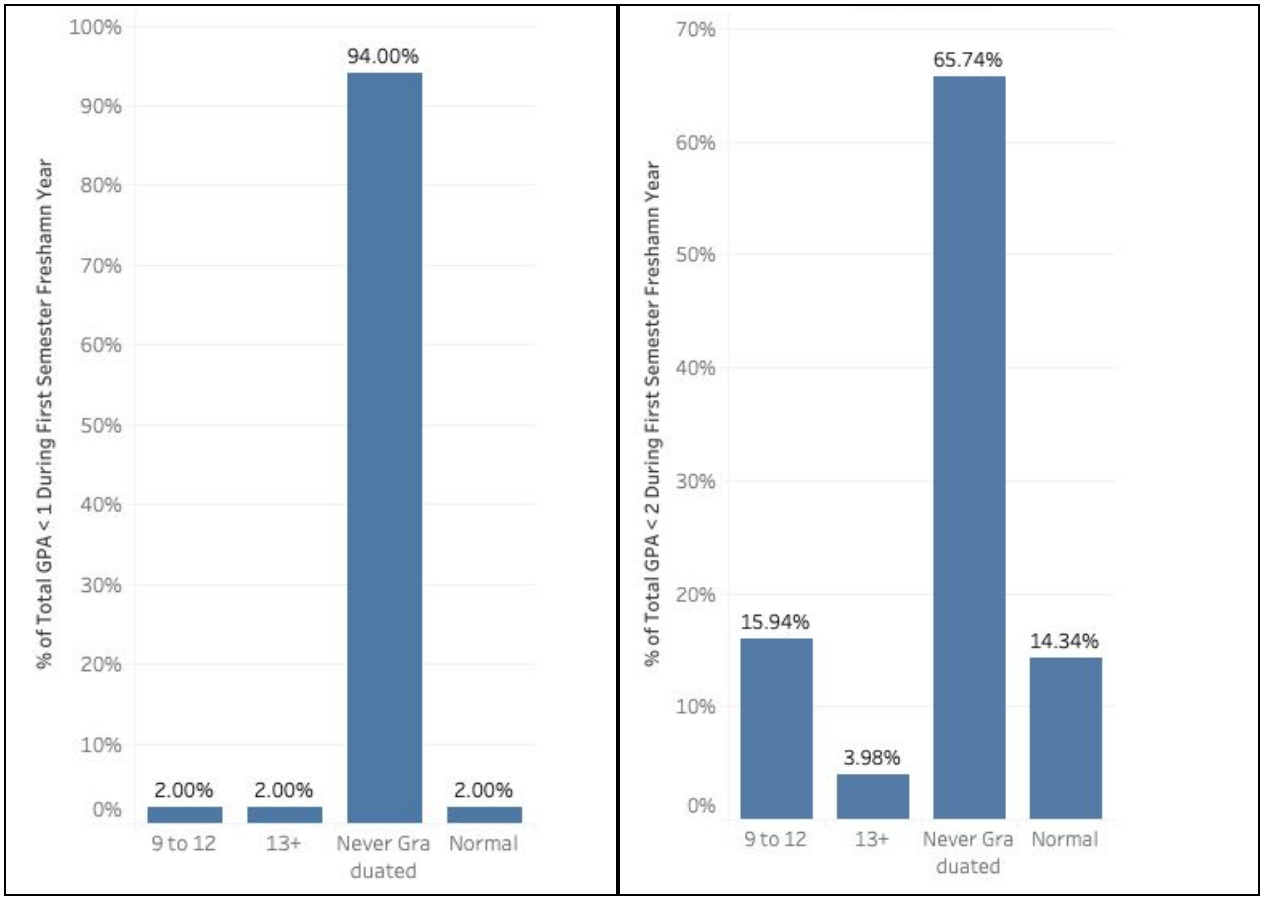
3.4 Disastrous Semester

A disastrous semester as categorized by the University of Rochester is a term GPA of less than 1.0 received by a student. Many students tend to struggle during their first year of college and this is the foundation of their college experience. We investigated whether first year first semester GPA could be an indicator of whether a student will graduate at the University or not. We looked into the impacts a disastrous semester or two disastrous semesters in a row could have on a students chances of graduating.

As one can see in Chart 5, 94% of first year first semester students who received a GPA of less than 1.0 did not graduate; that percentage drops to 70% if they received less than 2.0, which remains a huge dropout rate. Additionally, many students who had a GPA lower than a 2.0 may need to take a few extra semesters to graduate with approximately 16% of them graduated in 9-12 semesters.

From the overall perspective, 31% of all students who got a lower than 1.0 GPA in one semester didn't graduate; and if they experienced 2 or more disastrous semesters in a row, their chance of not graduating spikes up to 51%.

Chart 5. The graduating distribution of students experiencing disastrous semesters



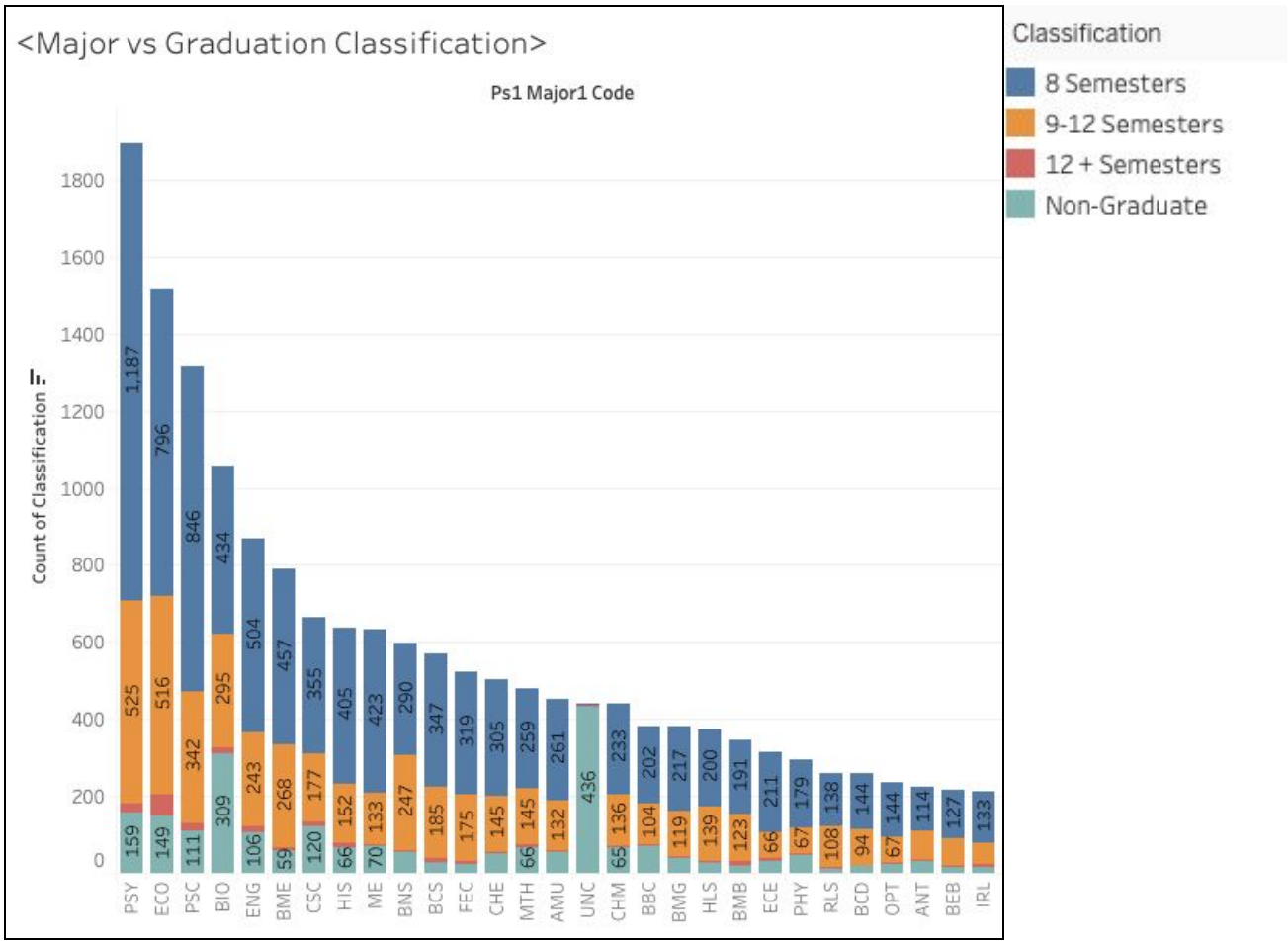
3.5 Majors vs GPA

3.5.1 Individual Major

To better understand the potential academic struggles a student is undergoing, our sponsor was interested in knowing if any majors have a different trendline than others. With the major distribution chart in Chart 6, it is clear that Psychology "PSY", Economics "ECO", Political Science "PSC", and Biology "BIO" are the most popular majors in the dataset. While the distribution between the 4 graduating groups seem similar across most majors, it is very noticeable that "UNC" - Uncertainty of Major, has a significantly higher number of non-graduates.

After running some statistical tests, we were able to discover certain majors that have significant correlation with graduating on time with p-values under 0.01. Such positively correlated majors are for example: BCS (Brain Cognitive Sciences), FEC (Financial Economics), HBS (Health, Behavior and Society), OPE (Optical Engineering), STT (Statistics), WS (Women's Studies). On the other hand, the ones with a higher indication of graduating late are CSM (Computer Science Mathematics) and CSA(Computer Science Applied Math) etc. (see Appendix I)

Chart 6. The graduation classification across individual majors



3.5.2 Multiple Majors/Minors

Exploring the effect of having double majors or more minors on the individual's cumulative GPA is part of the focus to help identify students at risk of not graduating on time. In Chart 7, the left hand side heatmap shows such a correlation. As the orange color goes deeper, it shows a higher average cumulative GPA along with greater academic responsibilities. The lower GPA for groups graduating with 12 + semesters and those never graduated further supports the negative correlation found between graduating late with GPA in Section 3.2.

On the right hand side (blue), it shows the number of students belonging to each classification. Single majors is certainly the largest student group in the dataset and most graduated in 8 semesters. By taking a horizontal look, except for "no majors", 60% of each major/minor grouping graduate within 8 semesters and a total of 96% graduate within 9-12 semesters. On the other hand, 99% of the non-graduate students are those who've never declared a major; this signifies that the students most likely in danger of dropping out tend to show signs by the end of sophomore year, especially if they still haven't declared a major, connecting to the foundings on the large number of non-graduates on "UNC" in chart 6.

To make sure the correlations that we explored are statistically significant, the ANOVA test is used to further analyze; we were able to find that the mean cumulative GPA across all various major/minor groups are significantly different with a 99% value since our P-value is under 0.01, supporting the view that GPA correlates to many aspects of a student's academic abilities.

Chart 7. Breakdown of GPA and graduation classification of students having different numbers of majors/minors.

<Average Cumul GPA by Group>

Majors/Minors	Classification: 8 Semesters	9-12 Semesters	12 + Semesters	Non-Graduate	Classification: 8 Semesters	9-12 Semesters	12 + Semesters	Non-Graduate
No Major				2.575				2,707
One Major	3.337	3.126	2.802	2.809	5,318	2,909	237	14
One Major One Minor	3.432	3.296	3.012	2.940	3,362	1,674	63	1
Two Majors	3.497	3.401	3.077	2.970	24	16		
One Major Two Minors	3.530	3.427	2.932	3.220	581	333	6	1
One Major Three Minors	3.520	3.498			1,568	846	27	1
Two Majors One Minor	3.586	3.536	3.073		389	251	7	
Two Majors Two Minors	3.581	3.584			3			
Two Majors Three Minors	3.593				22	27		

3.5.3 Changing Majors

As an undergraduate student, each student has opportunities to change their course of studies. In Table 8, the percentage of students who changed majors among the 4 graduating classifications is calculated. One can notice that, except non-graduates, students who changed majors once consisted at least 50% of the overall number. While the students who changed majors twice or three times only accounts for 5% of those who graduated in 8 semesters, 10% for 9-12 semesters, and 25% for 12+ Semesters. While 81% of the non-graduates never changed their majors, it is because, as previously explored, many of them never declared a major to begin with.

In Table 9, where the calculation is showing the percentage of students graduating on time if they changed their majors several times. More than 52% of the students who changed majors twice graduated with more than 8 semesters and 5% of them didn't graduate; that number jumps to 69% and 8.78% for students changing majors three times. Looking down the columns, as a student changes their major one more time, it is 15% more likely that they will graduate within 9-12 semesters, and doubling their chance of graduating more than 12 semesters later.

Therefore, we have a reason to believe that changing major once doesn't necessarily hurt a student's chance of graduating within 6 years, rather it could indicate that this student cares about their academic interest and graduating successfully. However, if students were to change majors again, starting from this point, there is a risk that they would need to take additional semesters to fulfill all requirements.

<Table 8. Percentage Down the Table>

	Classification			
Changed Majors	8 Semesters	9-12 Semesters	12+ Semesters	Non-Graduates
Never	33.75%	21.24%	19.71%	81.57%
Once	61.57%	68.18%	55.29%	15.57%
Twice	4.38%	9.15%	20.59%	2.39%
Three Times	0.29%	1.44%	4.41%	0.48%

<Table 9. Percentage Across the Table>

	Classification			
Changed Majors	8 Semesters	9-12 Semesters	12+ Semesters	Non-Graduates
Never	51.55%	17.43%	0.91%	30.12%
Once	59.40%	35.36%	1.61%	3.63%
Twice	41.76%	46.83%	5.92%	5.49%
Three Times	22.30%	58.78%	10.14%	8.78%

3.6 Term to Term Trends

Since our student gpa dataset was broad and contained lots of different data, we classified this dataset into several groups based on the number of semesters until their graduation. The categorized student groups are students who graduated on time (8 semesters), students who graduated in 9-12 semesters, students who graduated after 12 semesters, and students who never graduated. Using these classified groups, we explored the term to term gpa trend patterns to predict a student's likelihood of graduating.

Fig. 1 shows an overall gpa trend graph for students who graduated in 8 semesters. The red line in this graph indicates the trend throughout the students' entire semester. Obviously, it shows an overall stable trend, maintaining their gpa over 3.0 and showing a slight increase in their last semester. The violin plot of the term gpa right next to the trend plot represents that a high percentage of students who graduated on time has a gpa over 3.0 throughout their years in college. The color of dots specified here states the type of the semester as shown from the legend of the plot.

Fig. 1 Trend graph of students graduating in 8 semesters

The trend graph of students who graduated in 9-12 semesters is shown in Fig. 2. We can see that it shows a steadily increasing trend, keeping their grades up the most in the second half of their total semesters. However, their overall gpa throughout the semesters was mostly between 2.0 to 3.0, which is slightly lower than the students who graduated on time. Also, we can see that these students took a lot of summer semesters, which is the yellow dot, compared to the students who graduated on time.

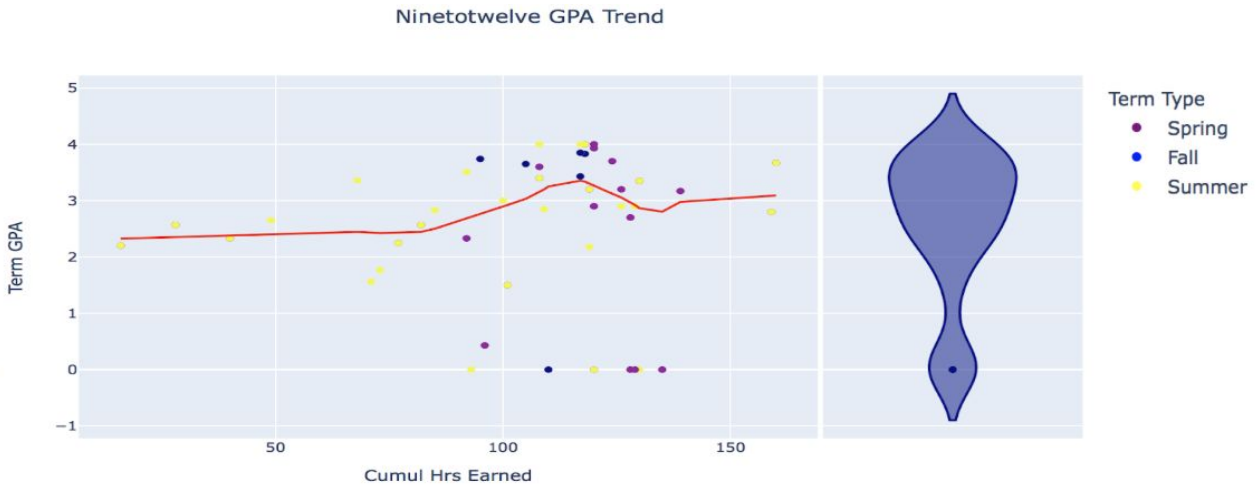


Fig. 2 Trend graph of students graduating in 9-12 semesters

The gpa trend graph of students who graduated in over 12 semesters is shown in Fig. 3. It shows a slowly falling trend with gpa drastically dropping in the students' last year in college. They mostly took only spring semesters and it is also shown that they tend to take more summer semesters at the end of their school years.

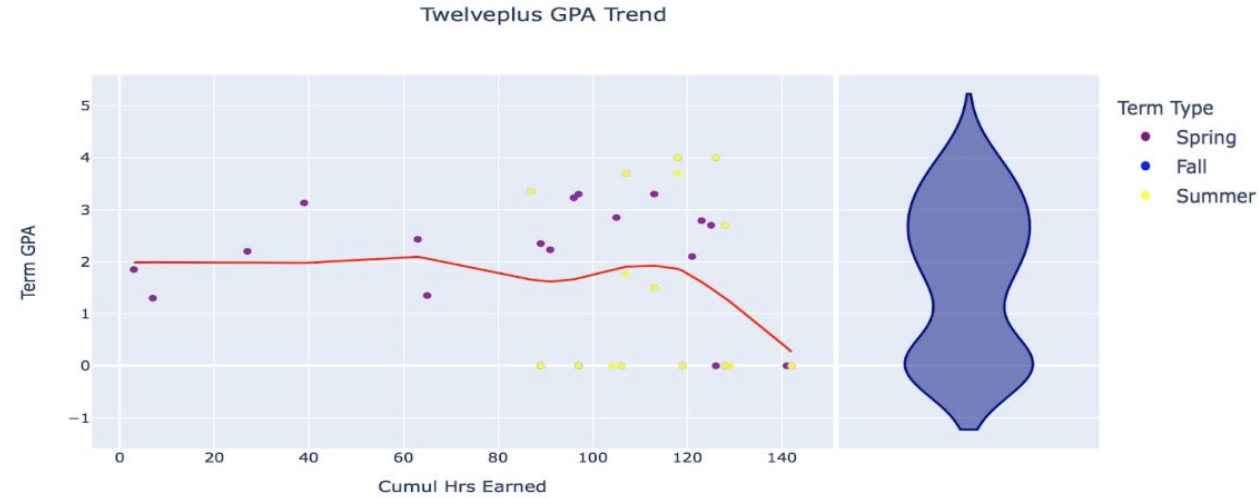


Fig. 3 Trend graph of students graduating in over 12 semesters

The trend graph for students who never graduated is shown in Fig. 4. Clearly, it shows a rapid downward trend for the first two semesters and the gpa remains very low for the rest of their semesters. From the violin plot right next to the trend graph, we can see that a high percentage of them mostly had a gpa around 0-1.0, which is very low compared to students who were able to graduate.

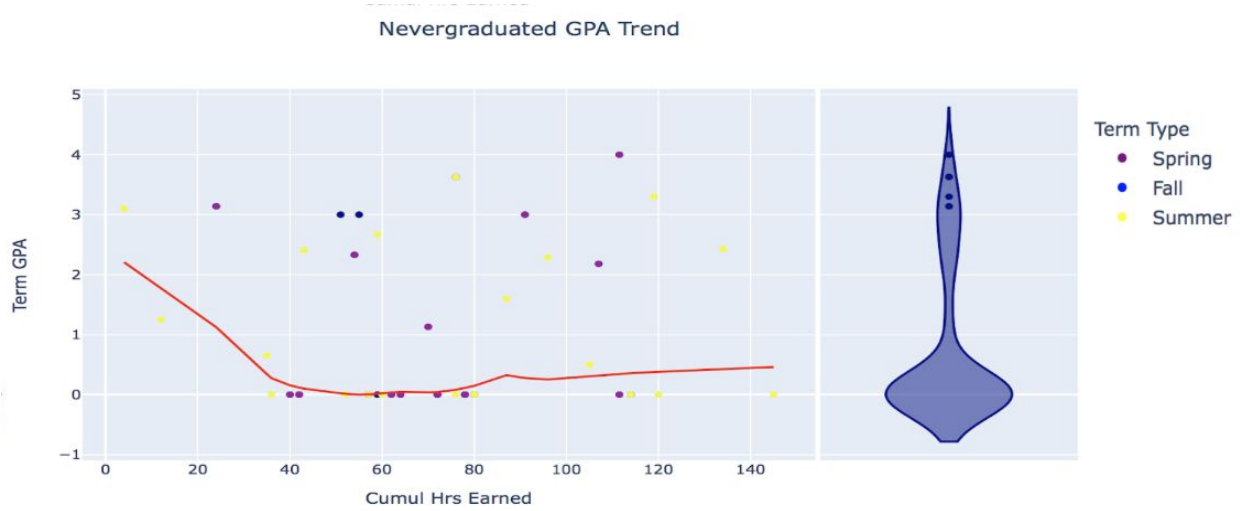


Fig. 4 Trend graph of students who never graduated

3.7 Summary

In this section, we obtained some key findings by analyzing the student gpa data based on three factors; disastrous semester, major, and gpa trend pattern. First, 31% of the students didn't graduate when they experienced one disastrous semester whereas 51% of the students didn't graduate when they had two or more disastrous semesters. Next, certain majors contribute to a higher chance of graduating or not graduating on time and also having multiple majors leads to higher average cumulative GPA. Lastly, the overall gpa trend of students who graduated on time and those who never graduated had an opposite pattern, respectively showing upward trend and downward trend.

4. Model Development

To verify that our findings could be in practical use for our sponsor, a prediction model based on the datasets given is developed. First, we built a deep learning model based on the full records of the students incorporated with our findings on the data analysis section. It was a multi-class classification model for the four classification groups with tensorflow's keras library. This model reached 85% to 90% accuracy depending on the number of labels that were in use. Considering the minimum accuracy (accuracy that could be reached by guessing that the student belongs to the largest classification group) being 55.2%, we believe that our model supports the significance of our findings. We also developed a model based on only the first four semesters of the records which showed 65% accuracy, with 56% base accuracy. This model would be helpful for our sponsor to flag students who potentially need extra care due to their higher risk of not graduating on time.

4.1 Modeling Setup

The aggregated data frame had 9058094 rows with 34 columns even after dropping columns that seemed to have low correlation to classification. Therefore, further simplification was needed. Due to the large size of the dataset, parallel processing and tensorflow-gpu library were utilized. Usual model building exercises such as one-hot encoding on categorical columns and normalization were performed after data cleaning.

4.1.1 Data Cleaning

It was impossible to feed in the original dataset with almost 10 million rows after labeling and one-hot-encoding, which extended the number of columns to over a hundred. We had to go through abstraction and simplification of the dataset based on our previous findings.

4.1.1 a) *Drop duplicates:*

We were able to drop columns after every time we were dropping each column since the data frame was a joined data frame based on Subject IDs (by its nature it creates duplicate rows for every unique column data)

4.1.1 b) *Normalize the number of column values based on frequency:*

Since there were different kinds of probations and it was impractical to uniquely label every type, the number of probation semesters a student experienced was normalized by creating a new column 'normalized probation' to contain the ratio of number of probation semesters/ number of all semesters for individual students

4.1.1. c) *Make multi-label columns into binary columns:*

This was frequently used for the cleaning process. For columns that had too many categorical values, for example, country of origin, the output is binarized such as international and domestic.

4.1.1 d) *Weed out the 'fake zeros':*

For our model to learn based on GPA, it was important to remove the 0.0 term

grades that are not due to academic failure (inactive status etc).

4.1.1 e) *Simplify the term-by-term columns:*

This process was the most important and difficult task. The main reason behind the large size of the dataframe is from the term-to-term information which included rows of individual classes that students registered for. This hindered our process of performing train-test-split since it is possible that different terms of the same student be randomly distributed which creates a lot of noise; instead, keeping each individual student as one slice of data in a single row is the ideal for the model to learn. We used our findings in the data analysis process - the number of times of changing majors, disastrous first year first semester, and double major or not. During this process, we also used the time when minimum and maximum gpa of a student occured (which semester) as an indicator to help better classify their graduating group.

4.1.2 Train-Validate-Test Splitting

The Scikit-Learn library's train-test-split function is used to split 90% of the dataset as training and 10% as testing data. We chose a relatively small proportion of the test set to better train the model.

4.1.3 Validation / Test Decoding Process

The 5-fold cross-validation method is used to prevent overfitting in the model

4.1.4 Making Prediction

Prediction was made by y_test and the scikit-learn library's prediction method with the training dataset by the model.

4.1.5 Evaluation Metric

We used loss and accuracy as our evaluation metrics on the model, as well as classification reports and confusion matrix to evaluate the model's performance of individual classification groups.

5. Model Specifications, Performance and Results

We built 3 different predictive models.

1) all columns that were left after cleaning out
2) with the columns that are considered most significant/ practical (columns like Academic standing were removed since probation obviously indicates poor performance of the student)
3) a model with only the first four semesters

The first model performed with a significantly high accuracy of 90% after training 400 ephoches. This clearly demonstrates that our analyzed data had overall significance for classifying the student groups. We limited our number of epochs to 500 to prevent overfitting. The loss function used was categorical cross entropy, with the output layer's activation function as softmax. Both optimize the performance of multi-label classification predictive models.

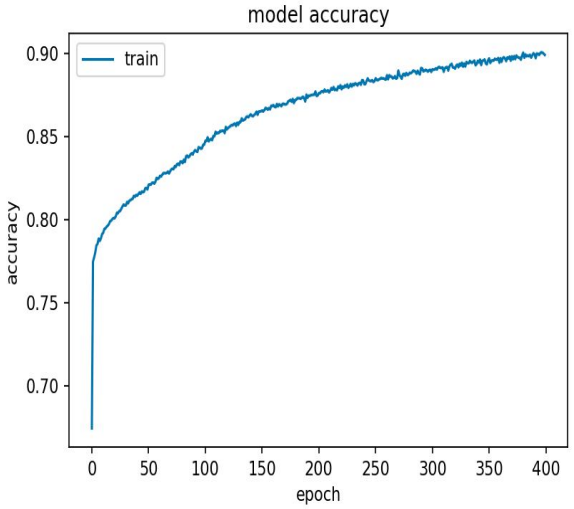
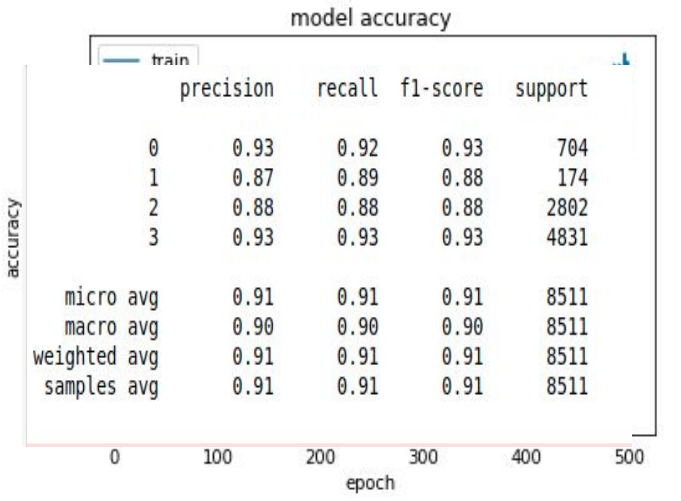


Fig. 5 Model accuracy based on all columns



	precision	recall	f1-score	support
0	0.93	0.92	0.93	704
1	0.87	0.89	0.88	174
2	0.88	0.88	0.88	2802
3	0.93	0.93	0.93	4831
micro avg	0.91	0.91	0.91	8511
macro avg	0.90	0.90	0.90	8511
weighted avg	0.91	0.91	0.91	8511
samples avg	0.91	0.91	0.91	8511

Fig. 6 Model accuracy based on most significant columns

The precision on the normal group and non-graduate groups were the highest, which demonstrates that it is easy to distinguish the two different groups from other groups. The 12+ semester group was identified as the hardest group to distinguish, which is understandable due to lack of data for the model to learn.

	precision	recall	f1-score	support
0	0.86	0.80	0.83	594
1	0.71	0.45	0.55	152
2	0.81	0.79	0.80	2251
3	0.88	0.91	0.89	3814
accuracy			0.85	6811
macro avg	0.81	0.74	0.77	6811
weighted avg	0.85	0.85	0.85	6811

The 12+ classification group further drags down the overall accuracy of the second model, with a 16% decrease in its precision.

	Non graduates	+12	9~12	Normal
Non graduates	476	1	52	65
+12	4	68	79	1
9~12	22	27	1782	420
Normal	54	0	297	3463

Confusion matrix reported that it is most difficult to distinguish between 9~12 semester groups and the normal group. This is a less significant issue since both groups are considered the "safe" groups according to our sponsor.

	Non graduates	+12	9~12	Normal
Non graduates	662	1	17	24
+12	7	154	13	0
9~12	13	21	2462	306
Normal	42	1	299	4489

The confusion matrix reported similar results for the second model as well.

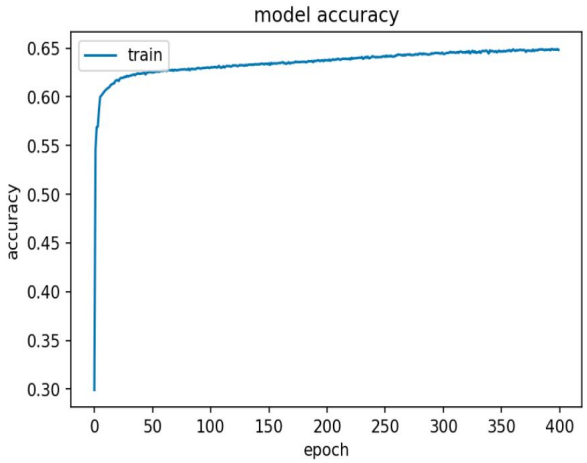


Figure 7 model learning graph: first four semesters

Figure above is the learning graph for the model with the first four semesters only. We implemented the same data cleaning methods. This showed that it is much more difficult to classify the group of students in their second year into classification groups than when having the whole record. It could be presumed that junior and senior year are more decisive on whether the students would graduate on time.

	precision	recall	f1-score	support
0	0.59	0.17	0.26	120
1	0.00	0.00	0.00	56
2	0.44	0.17	0.25	921
3	0.67	0.93	0.78	1883
accuracy			0.64	2980
macro avg	0.43	0.32	0.32	2980
weighted avg	0.59	0.64	0.58	2980

From the classification report, it was noticed that none of the students were classified as 12+ group by the model.

	Non graduates	+12	9~12	Normal
Non graduates	476	1	52	65
+12	4	68	79	1
9~12	22	27	1782	420
Normal	54	0	297	3463

Looking at the confusion matrix, this model has the most difficulty differentiating between the 12+ group and 9~12 group. The model relatively shows good precision on Non graduates and Normal groups which means that in the first four semesters, it is possible to find some indicators of the students who are not going to be graduating, but not so well who are going to take a longer time to graduate.

6. Conclusion & Future Works

In this project, we could find out that GPA is an attribute that significantly correlates with a student's chances of graduating on time after conducting an exploratory data analysis based on the number of disastrous semesters, the number of majors/minors they have, the influence of freshman GPA on graduation rate, lowest gpa a student got, and the term to term GPA trend patterns. Predictive model was also built using the data provided to identify attributes relevant to GPA that contribute to flagging students at higher risks of not graduating and showed a great performance with accuracy reaching to almost 90%.

There are some possible amendments and improvements we can make on our analysis. We can try training and implementing other machine learning techniques to further enhance our predictive model for better performance. Also, we can further identify potential pathways that would help increase a student's chance of graduating on time.

7. Advices and Suggestions

We have several pieces of advice that we could give for the students and the advising office.

1. It is very noticeable that the non-graduate students do not declare their majors. It would be best for the students to decide their academic path as soon as possible.
2. Clusters have their mean gpa significantly different to one another (with p value less than .1) therefore it would be beneficial for students to choose less challenging clusters (clusters like SOCI, MUSI, LING, ASLA had the best mean GPAs).
3. Take summer courses, more than 15% of the students graduate in time thanks to summer courses
4. Minimum GPA of the students had the most significant correlation between classification groups. If a student's gpa drops below 2 for a semester, advisors should take extra care for the student.

Reference

1. Marcus, J. (2020, March 30). Embattled colleges focus on an obvious fix: Helping students graduate on time. Retrieved May 06, 2020, from https://hechingerreport.org/embattled-colleges-focus-on-an-obvious-fix-helping-students-graduate-on-time/
2. U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), 12-month Enrollment component final data (2001-02 - 2017-18) and provisional data (2018-19).
3. U.S. Department of Education. (2015, July 27). Fact Sheet: Focusing Higher Education on Student Success. Retrieved May 06, 2020, from https://www.ed.gov/news/press-releases/fact-sheet-focusing-higher-education-student-success
4. University of Illinois at Urbana-Champaign. "First-semester GPA a better predictor of college success than ACT score." ScienceDaily. ScienceDaily, 2 February 2016. <www.sciencedaily.com/releases/2016/02/160202174021.htm>.

Appendix I. Coefficients of individual majors based on ANOVA statistical test. The * indicates the coefficient's significance based on a 95% confidence.

```
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.95652    0.19619  15.070  < 2e-16 ***
## `Ps1 Major1 Code`AH       0.39642    0.22114   1.793 0.073052 .
## `Ps1 Major1 Code`AME      0.12681    0.33505   0.378 0.705076
## `Ps1 Major1 Code`AMS      0.54348    0.50971   1.066 0.286321
## `Ps1 Major1 Code`AMU      0.36942    0.20113   1.837 0.066261 .
## `Ps1 Major1 Code`ANT      0.24884    0.20601   1.208 0.227113
## `Ps1 Major1 Code`APM      0.49603    0.21202   2.340 0.019314 *
## `Ps1 Major1 Code`ASL      0.46453    0.21865   2.125 0.033637 *
## `Ps1 Major1 Code`ATH      0.27205    0.25255   1.077 0.281401
## `Ps1 Major1 Code`BBC      0.18222    0.20201   0.902 0.367036
## `Ps1 Major1 Code`BCD      0.43730    0.20471   2.136 0.032677 *
## `Ps1 Major1 Code`BCS      0.53385    0.20010   2.668 0.007638 **
## `Ps1 Major1 Code`BEB      0.46744    0.20632   2.266 0.023487 *
## `Ps1 Major1 Code`BEN     -0.95652    0.96111  -0.995 0.319640
## `Ps1 Major1 Code`BET      0.37681    0.33505   1.125 0.260757
## `Ps1 Major1 Code`BIG     -0.06763    0.36993  -0.183 0.854937
## `Ps1 Major1 Code`BIO     -0.14857    0.19831  -0.749 0.453742
## `Ps1 Major1 Code`BMB      0.45336    0.20264   2.237 0.025278 *
## `Ps1 Major1 Code`BME      0.46141    0.19901   2.318 0.020434 *
## `Ps1 Major1 Code`BMG      0.39348    0.20204   1.948 0.051481 .
## `Ps1 Major1 Code`BNS      0.33159    0.19993   1.659 0.097230 .
## `Ps1 Major1 Code`BSB      0.47002    0.20660   2.275 0.022917 *
## `Ps1 Major1 Code`CGS     -0.38509    0.40614  -0.948 0.343056
## `Ps1 Major1 Code`CHE      0.44229    0.20061   2.205 0.027488 *
## `Ps1 Major1 Code`CHM      0.26671    0.20126   1.325 0.185115
## `Ps1 Major1 Code`CL       0.71014    0.43131   1.646 0.099682 .
## `Ps1 Major1 Code`CLS      0.34580    0.24306   1.423 0.154829
## `Ps1 Major1 Code`CLT      0.51014    0.31226   1.634 0.102333
## `Ps1 Major1 Code`CMP      0.26570    0.22922   1.159 0.246404
## `Ps1 Major1 Code`CSA     -1.95652    0.57756  -3.388 0.000706 ***
## `Ps1 Major1 Code`CSC      0.19686    0.19955   0.987 0.323887
## `Ps1 Major1 Code`CSM     -1.62319    0.57756  -2.810 0.004952 **
## `Ps1 Major1 Code`DMS      0.33977    0.26698   1.273 0.203147
## `Ps1 Major1 Code`DSC      0.04348    0.50971   0.085 0.932023
## `Ps1 Major1 Code`EBA     -0.55652    0.46426  -1.199 0.230650
## `Ps1 Major1 Code`EBS     -0.14073    0.24857  -0.566 0.571280
## `Ps1 Major1 Code`ECE      0.48335    0.20320   2.379 0.017383 *
## `Ps1 Major1 Code`ECO      0.33484    0.19767   1.694 0.090287 .
## `Ps1 Major1 Code`ECW      0.04348    0.69362   0.063 0.950020
## `Ps1 Major1 Code`EE      -0.09445    0.22060  -0.428 0.668537
## `Ps1 Major1 Code`EMC      0.04348    0.96111   0.045 0.963919
## `Ps1 Major1 Code`ENG      0.36260    0.19877   1.824 0.068130 .
## `Ps1 Major1 Code`EPD      0.47591    0.24983   1.905 0.056802 .
## `Ps1 Major1 Code`ES      -0.27319    0.23075  -1.184 0.236451
## `Ps1 Major1 Code`ESP      0.16848    0.23302   0.723 0.469671
## `Ps1 Major1 Code`EST      0.18634    0.40614   0.459 0.646389
## `Ps1 Major1 Code`EVS     -0.01535    0.21718  -0.071 0.943672
## `Ps1 Major1 Code`FEC      0.54730    0.20045   2.730 0.006331 **
## `Ps1 Major1 Code`FMS      0.32583    0.20904   1.559 0.119077
## `Ps1 Major1 Code`FR       0.08348    0.22426   0.372 0.709719
```

```
## `Ps1 Major1 Code`GEO  0.43295    0.21865   1.980 0.047704 *
## `Ps1 Major1 Code`GER  0.26570    0.25116   1.058 0.290109
## `Ps1 Major1 Code`GSM -1.95652    0.32647  -5.993 2.10e-09 ***
## `Ps1 Major1 Code`GSW  0.04348    0.96111   0.045 0.963919
## `Ps1 Major1 Code`HBS  0.62243    0.20896   2.979 0.002899 **
## `Ps1 Major1 Code`HIS  0.45228    0.19970   2.265 0.023536 *
## `Ps1 Major1 Code`HLP  0.63607    0.26698   2.383 0.017205 *
## `Ps1 Major1 Code`HLS  0.41176    0.20216   2.037 0.041684 *
## `Ps1 Major1 Code`IDE -0.62319    0.36993  -1.685 0.092081 .
## `Ps1 Major1 Code`IPA  0.51304    0.21491   2.387 0.016985 *
## `Ps1 Major1 Code`IRL  0.49372    0.20660   2.390 0.016872 *
## `Ps1 Major1 Code`JCM  0.50502    0.26933   1.875 0.060793 .
## `Ps1 Major1 Code`JCP -1.95652    0.69362  -2.821 0.004796 **
## `Ps1 Major1 Code`JPN  0.26453    0.21865   1.210 0.226354
## `Ps1 Major1 Code`LIN  0.44225    0.20957   2.110 0.034848 *
## `Ps1 Major1 Code`ME   0.48177    0.19972   2.412 0.015867 *
## `Ps1 Major1 Code`MEG  0.14693    0.26271   0.559 0.575978
## `Ps1 Major1 Code`MEI  0.29545    0.21321   1.386 0.165857
## `Ps1 Major1 Code`MEV  0.02387    0.23632   0.101 0.919544
## `Ps1 Major1 Code`MST -0.03986    0.25116  -0.159 0.873918
## `Ps1 Major1 Code`MTH  0.28982    0.20084   1.443 0.149021
## `Ps1 Major1 Code`MUA  0.47205    0.40614   1.162 0.245140
## `Ps1 Major1 Code`MUS  0.26458    0.20721   1.277 0.201665
## `Ps1 Major1 Code`OPE  0.69110    0.21661   3.191 0.001422 **
## `Ps1 Major1 Code`OPT  0.44854    0.20549   2.183 0.029059 *
## `Ps1 Major1 Code`PAS  0.19882    0.21699   0.916 0.359540
## `Ps1 Major1 Code`PBP -1.36829    0.30094  -4.547 5.48e-06 ***
## `Ps1 Major1 Code`PHL  0.22062    0.20868   1.057 0.290425
## `Ps1 Major1 Code`PHY  0.32051    0.20367   1.574 0.115577
## `Ps1 Major1 Code`PSC  0.50251    0.19789   2.539 0.011114 *
## `Ps1 Major1 Code`PSY  0.49015    0.19737   2.483 0.013023 *
## `Ps1 Major1 Code`RLS  0.47809    0.20468   2.336 0.019511 *
## `Ps1 Major1 Code`RST  0.49348    0.28767   1.715 0.086277 .
## `Ps1 Major1 Code`RUS  0.19732    0.24736   0.798 0.425045
## `Ps1 Major1 Code`SA   0.45419    0.21539   2.109 0.034983 *
## `Ps1 Major1 Code`SP   0.35067    0.21042   1.667 0.095621 .
## `Ps1 Major1 Code`STT  0.65638    0.22971   2.857 0.004275 **
## `Ps1 Major1 Code`TH   0.24348    0.27184   0.896 0.370448
## `Ps1 Major1 Code`UNC -1.93158    0.20124  -9.599  < 2e-16 ***
## `Ps1 Major1 Code`WS   0.80538    0.28398   2.836 0.004572 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9409 on 20295 degrees of freedom
## Multiple R-squared:  0.1457, Adjusted R-squared:  0.1419
## F-statistic: 38.04 on 91 and 20295 DF,  p-value: < 2.2e-16
```

Appendix II. Median minimum gpa among classification groups.

```
dataframe.groupby('Classification')['Min_GPA'].median()
```

```
Classification
1    1.62
2    1.50
3    2.58
4    3.00
Name: Min_GPA, dtype: float64
```

Appendix 3. The clusters' mean gpa difference p value with one sample t test.

```
In [103]: from scipy import stats
          stats.ttest_1samp(values,0)
```

```
Out[103]: Ttest_1sampResult(statistic=82.21126459116245, pvalue=1.057951015882456e-40)
```