Predicting Student Success of Finishing Their Major

ST 841 Final Report

Prepared by:

Katherine Allen, Andrew Giffin, and Robert Pehlman

Adviser:

Dr. Consuelo Arellano

NCSU Statistics Department

Submitted to:

Dr. Nancy Floyd

Director, Office of Institutional Analytics

December 8, 2016

Contents

1 Introduction

The Office of Institutional Research and Planning at North Carolina State University would like to identify courses or course combinations that have a "gatekeeper effect" in determining whether a student at North Carolina State University will be able to finish their major – or, needs to consider changing majors. The intended use of this information will be to assist student advisers to help students make academic career plans. Ideally, this modeling project would provide student-individualized predictions of success probability in various majors, based on past academic history as well as other pertinent factors.

The data have been gathered by The Office of Institutional Research and Planning on undergraduate students' majors, major switches, grades, and coursework from all colleges at North Carolina State University for the years 1970 to 2016, although data is sparse for earlier years. The dataset includes information on all students in attendance during those years.

The research objectives are as follows:

1. Calculate probabilities of a student finishing a degree based on the grade awarded in courses (allowing for multiple attempts of the same course).
2. Calculating comparative probabilities of succeeding in alternative majors based on grades awarded in courses.

2 The Data

2.1 Data Cleaning

As given by the client, the core dataset is stored in 3 main files:

- a large file with information at the student/course level (3.7 million rows)
- a file with information at the student/term level (500k rows)
- a smaller file with details of which students graduated, and in which major (70k rows)

Each file has between 5-60 variables. The majority of these variables have values that are encoded, and the 16 dictionary files that came with the three main files are needed to parse the values into human-readable format. (E.g., for the variable "Student Major", the value of "14CSCMR" needs to be translated to "Computer Science").

Pre-processing has involved incorporating the dictionary file information into the three main files, and doing basic cleaning on the main files (e.g., specifying missing data, deleting junk variables, enforcing variable-name continuity across data files, and putting character-date variables into R-recognized date format.) Any analyses we do will require more data pre-processing to successfully implement. This additional work is documented in section 3, Analysis.

2.2 Variables of Interest

Each of the three main data files contains important information and variables for our analysis:

"Student Course" file: this file contains the key course-level grade information that will be the major predictor in our analysis. Specific variables that will be essential are:

- "Course": (Department/Course number)
- "Final Grade"
- "Course Date"

- "Student ID"
- "Term ID"

(The last two will be necessary for merging with other files.)

"Student Term" file: this file contains term-level information on students, to help understand where students are in their academic careers, when they take specific courses.

Specific variables that will be essential are:

- "GPA"
- "Current Major"
- "Academic Standing" (Good/On Probation/etc)
- "Student ID"
- "Term ID"

(The last two will be necessary for merging with other files.)

"Graduated Students" file: this file contains students-level data for students that have graduated and with which major(s). The information in this dataset will go into creating our "response" variable. Note that there is no variable for whether students graduated or not. A student being in this file implies graduation; their not being on this file implies non-graduation. Thus, a binary variable will need to be created that incorporates this information. Specific variables of interest in this file are:

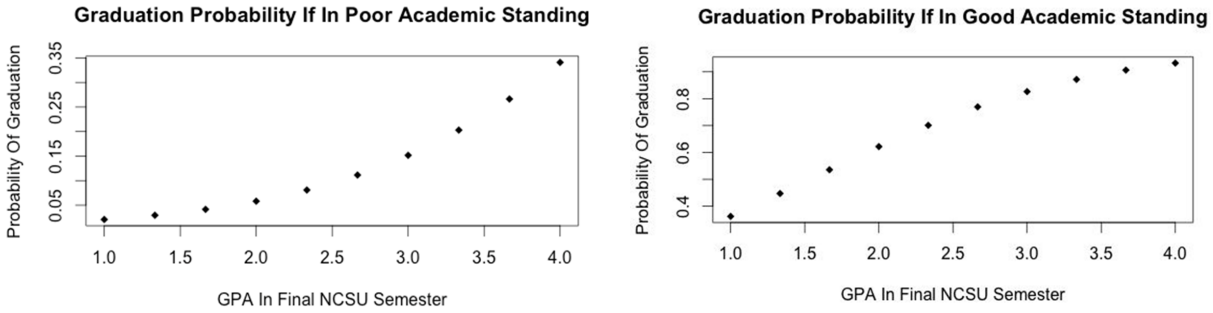
- "Major at Graduation"
- "Student ID"

- "Term ID"

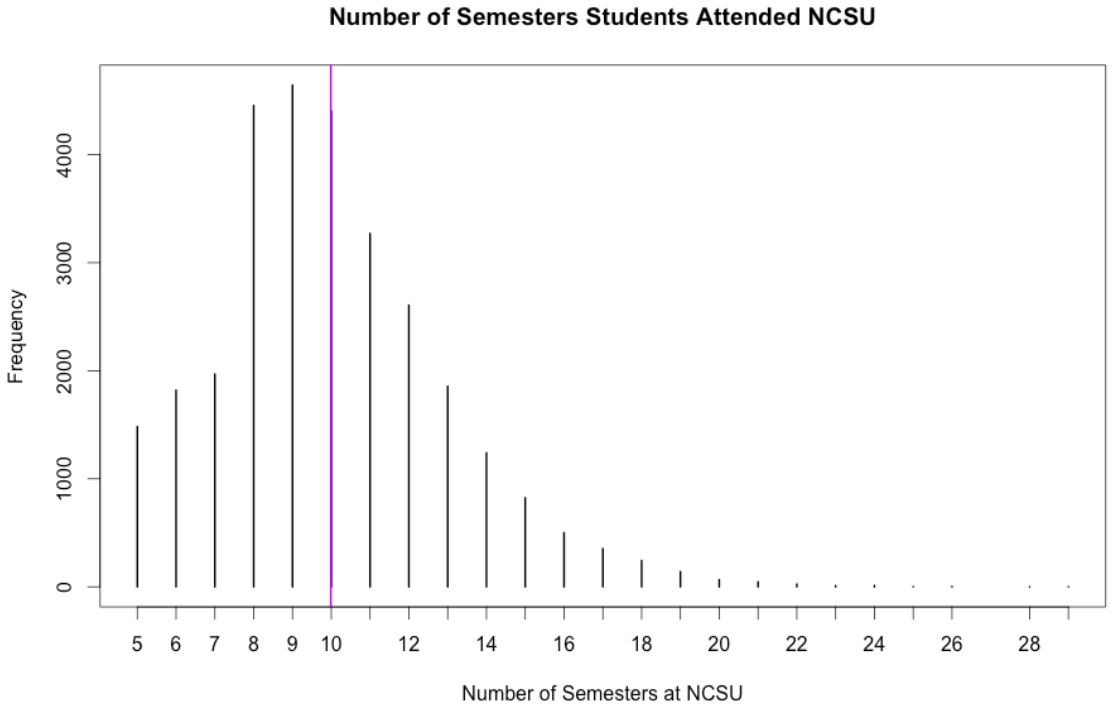
(The last two will be necessary for merging with other files.)

2.3 Exploratory Data Analysis

Looking the "Student Term" dataset, we created flags to determine whether the student had graduated and whether they were in good academic standing. We cleaned the data and examined only students whose last term at NC State was between 2006 and 2013. We then built a logistic regression model to determine probability of graduating based on academic standing and GPA in terminal semester. We ensured that there were students that were in every combination of academic standing and graduation status so that the logistic regression would be interpretable. Below are the results, demonstrating how probability of graduation varies as a function of GPA and academic standing.



For students in the data range we examined (Fall 2001 - Fall 2013), we wanted to make note of the number of semesters students attended NCSU, before either graduating, or not attending another course between Fall 2013 and the present Fall 2016 semester.



Both the mean and median number of semesters students in this date range attended NCSU are 10 (represented by the purple line in the graph). Knowing how long students typically attend NCSU will be of use in our later analysis, when we run a logistic regression based off of Student Term data.

3 Analysis

Our strategy for tackling this problem was to approach from two angles: Our main angle is to use "matrix completion", to use individualized student grades for the classes they have taken, to "fill in" estimates for what grades they would have achieved in classes which they have not taken. This highly individualized approach matches students with similar students, to create student-customized estimates. While this approach is promising, it is a relatively new technique with different tuning parameters, as well as nuance-required in its interpretation.

We also plan to supplement matrix completion with logistic regression, the standard technique when required to estimate probabilities using a binary response rate. However, logistic regression will only provide broad, more generalizing results.

3.1 Logistic Regression

3.1.1 Introduction and Interpretation

Logistic Regression is the gold standard technique for working with a binary response rate (in our case, if a student graduated or did not graduate in a certain time period), and outputting probabilities of an explanatory variable affecting that response rate. In our case, logistic regression answers the question: what is the probability a student will graduate given their 1. first term GPA and College or 2. fourth term GPA and Major. Do these factors have any affect on graduation?

3.1.2 Data Pre-processing and Subsetting

For this analysis, we subsetted on students in our date range (Fall 2001 - Fall 2013), who had attended NCSU for most than five terms. From there, we created a data set of each students' fourth term and each students' first term GPA. For the first term data, we found the top five colleges for first year students in our subsetted data to be Engineering, Agriculture and Life Sciences, Humanities and Social Sciences, Undergraduate Academic Program, and Management. All other colleges were grouped into one and given the name "Other," to keep the number of coefficients in our logistic regression down.

For the fourth term, we found the top 20 majors: Business Management; Department of Biology; Mechanical and Aerospace Engineering; Communication; Civil Engineering; Psychology; Electrical and Computer Engineering; First Year College; Political Science; Parks,

Recr and Tourism Mgmt; Textile and Apparel Management; Accounting; Animal Science; Computer Science; Sociology ; College of Engineering Deans Office; Chemical Engineering; English; Forestry; Biochemistry. All other majors were set to "Other."

Notice that First Year College is still a viable major on this list, despite that fact that we're subsetting by fourth term. A few reasons for this may be, a. when a student who started in the fall takes summer term courses after their first year, their fourth term would be the Summer 2 term, at which point they may still be a member of the First Year College. Another theory is that, although we assume that students at NCSU are required to switch out of the First Year College major after their second year, there may be cases where a student may remain in the college for longer; perhaps a student might take less than a full course load each semester.

3.1.3 Methodology

A logistic regression was performed on the college by first year GPA interaction, and the major by fourth year gpa interaction, using the R function 'glm.'

3.1.4 Results

Results from the regression on College by First Year GPA:

Coefficient	Estimate	Standard Error	Z-Value	P-value
1st Semester GPA	0.95	0.05	17.964	<.0001
Engineering College	-0.95	0.22	-4.337	<.0001
Humanities and Social Sciences College	0.39	0.25	1.581	0.11
College of Management	0.29	0.33	0.876	0.38
Other	0.99	0.21	4.542	<.0001
Undergrad Academic Program	-1.32	0.27	-4.972	<.0001
GPA and Engineering College	0.28	0.08	3.764	0.0002
GPA and Humanities and Social Sciences College	0.003	0.09	0.032	0.97
GPA and College of Management	0.11	0.11	0.965	0.335
GPA and Other Colleges	-0.33	0.07	-2.539	<.0001
GPA and Undergrad Academic Program	0.45	0.09	4.761	<.0001

Examining the results of our logistic regression, we see that there is a significant interaction between GPA of students in their first semester and if these students where in the Engineering College, Undergraduate Academic Program, or 'Other' College on graduation rate. Because these interaction terms are significant, we do not care about the main effects of GPA, and these colleges on graduation rate.

Thus, for every one unit change in GPA (i.e. 2.0 to 3.0), the log odds of graduating for a first year student in the Engineering College is 0.28 (Thus, a student in the Engineering College with a first term GPA of 3.0 is 1.3 times more likely to graduate than a student with a 2.0 GPA. A student with a GPA of 4.0 in their first term is thus more than 2 times more likely to graduate than a peer with a 2.0 GPA in their first term). Thus, having a better GPA does matter for Engineering College students' graduation rate.

For every one unit change in GPA, the log odds of graduating for a first year student in the Undergrad Academic Program is 0.45 (a student with a GPA of 3.0 in their first term is 1.6 times more likely than a student with a GPA of 2.0 in their first term to graduate). As most students in the Undergrad Academic Program most likely go on to enter other majors,

I would refrain from thinking too much about this result before investigating the second logistic regression.

For every one unit change in GPA, the log odds of graduating for a first year student in the "Other" College categorical is -0.33. It is difficult to interpret what these decreasing log odds mean, as the idea of lowering your GPA improving your graduating rate makes no intuitive sense. Thus, my conclusion is that this "Other" category is too broad: including most of the Colleges in one grouping has altered our results. One would have to run another regression on each College individually to better understand what is occurring here. However, this would yield too many coefficients. If an adviser is particularly interested in the significance of their college by GPA interaction on graduation rate, it is recommended they run a logistic regression solely on the college of interest.

There are no significant effects involving the College of Management or the College of Humanities and Social Sciences. The most straight-forward interpretation of this is that improving first term GPA by one unit in either of these colleges has such a small effect on graduation rate, it is nearly negligible. Another interpretation is perhaps there is less variation between first term GPAs of students in these colleges, so it's more difficult for our logistic regression to find any results.

Regression on Fourth Term GPA by Major Results:

Coefficient	Estimate	Standard Error	Z-Value	P-value
Fourth Semester GPA	1.63	0.27	6.10	<.0001
Animal Science	-1.31	0.96	-1.37	0.17
Biochem	-1.65	1.01	-1.62	0.10
Business Management	-1.14	0.92	-1.25	0.21
Chemical Engineering	-1.34	1.25	-1.08	0.28
Civil Engineering	-1.96	0.99	-1.98	0.047
College of Eng. Dean's Office	-2.59	0.94	-2.78	0.005
Communication	-1.46	1.09	-1.35	0.18
Computer Science	-1.48	0.97	-1.53	0.12
Biology	-0.89	0.85	-1.05	0.29
Electrical and Computer Eng	-2.04	0.94	-2.167	0.03
English	-1.26	1.01	-1.17	0.24
First Year College	-1.81	0.89	-2.05	0.04
Forestry	-0.19	1.02	-0.19	0.85
Mechanical and Aerospace Eng	-2.27	0.90	-2.512	0.01
Other	-0.72	0.80	-0.90	0.37
Parks, Recr, and Tourism Mgmt	-1.05	1.02	-1.029	0.3
Political Science	-0.64	0.98	-0.652	0.51
Psychology	-0.68	0.97	-0.702	0.48
Sociology	-0.31	1.04	-0.307	0.76
Textile and Apparel Mgmt	-1.85	0.07	-1.71	0.09

Coefficient	Estimate	Standard Error	Z-Value	P-value
GPA and Animal Science	0.25	0.34	0.744	0.46
GPA and Biochem	0.26	0.35	0.76	0.45
GPA and Business Mgmt	0.55	0.32	1.73	0.45
GPA and Chemical Engineering	0.49	0.43	1.13	0.08
GPA and Civil Engineering	0.73	0.35	2.06	0.27
GPA and College of Eng. Dean's Office	0.55	0.34	1.61	0.04
GPA and Communication	0.69	0.39	1.784	0.11
GPA and Computer Science	0.30	0.34	0.91	0.36
GPA and Biology	0.13	0.29	0.44	0.66
GPA and Electrical and Computer Eng.	0.55	0.33	1.67	0.09
GPA and English	0.31	0.38	0.83	0.41
GPA and First Year College	0.30	0.32	0.96	0.34
GPA and Forestry	0.07	0.37	0.19	0.85
GPA and Mechanical and Aerospace Eng	0.73	0.32	2.31	0.02
GPA and Other	0.09	0.27	0.34	0.73
GPA and Parks, Recr, and Tourism Mgmt	0.44	0.37r	1.21	0.23
GPA and Political Science	0.18	0.35	0.51	0.61
GPA and Psychology	0.17	0.34	0.50	0.62
GPA and Sociology	0.11	0.37	0.29	0.77
GPA and Textile and Apparel Mgmt	0.69	0.39	1.73	0.08

This second logistic regression, despite now using fourth term GPAs, echos what we saw with our regression on first term GPAs: more or less, the only significant terms involve students in Engineering Majors (our analysis looked at Chemical, Civil, Electrical and Computer, and Mechanical and Aerospace Engineering and all but Civil yielded a significant interaction with GPA. However, Civil Engineering has a significant main effect). However, we also see that GPA affects the graduation rate in Textile and Apparel Management significantly at an $\alpha = 0.09$ level (p-value = 0.08). In addition, the main effect of being in the First Year

College on graduation rate is significant at an $\alpha = 0.05$ level (p-value = 0.04). Although, I do not know enough about why students might be in the First Year College in their fourth term, and will leave that up to the reader to interpret.

A Chemical Engineering Major with a GPA of 3.0 in their fourth term is 1.6 (log odds 0.49) times more likely to graduate than their peer with a fourth term GPA of 2.0 in their same major. A student in Electrical and Computer Engineering is 1.7 times more likely to graduate if their fourth year GPA is 3.0 vs. 2.0

A Mechanical and Aerospace Engineering Major with a GPA of 3.0 in their fourth term is 2.1 (log odds 0.73) times more likely to graduate than their peer with a fourth term GPA of 2.0 in their same major. GPA appears to greatly affect graduation rate in this major. Similarly, a student in Textile and Apparel Management with a GPA of 3.0 is also around 2 times more likely to graduate than their peer with a GPA of 2.0 (log odds 0.69).

Our conclusion is that having a higher GPA in the majors listed above significantly increased the odds of graduating (the Coefficient associated with each major by GPA interaction represents their log odds of graduating). Perhaps students in these majors with a low GPA should consider transferring to a different major as early as their fourth term.

These logistic regression are meant to be more of an exploratory analysis into which aspects of a student's academic career affect their graduation rate are demonstrated. It seems both regressions overwhelmingly indicate that, from their first term, students in Engineering majors with a higher GPA are much more likely to graduate than their peers in the same major. For other Colleges and majors, lacking a significant interaction with GPA, or a significant College main effect, I would recommend analysis on a course by course level, or a student-by-student level to piece out why a higher GPA does not necessarily lead to a higher graduation rate among these majors, as that seems counter-intuitive.

3.2 Matrix Completion

3.2.1 Introduction and Interpretation

Matrix completion is a technique that takes a matrix with some missing values, and attempts to fill those values in, such that the matrix can be approximated in with another *low rank* matrix. This is typically done by minimizing the nuclear norm of the approximation matrix.

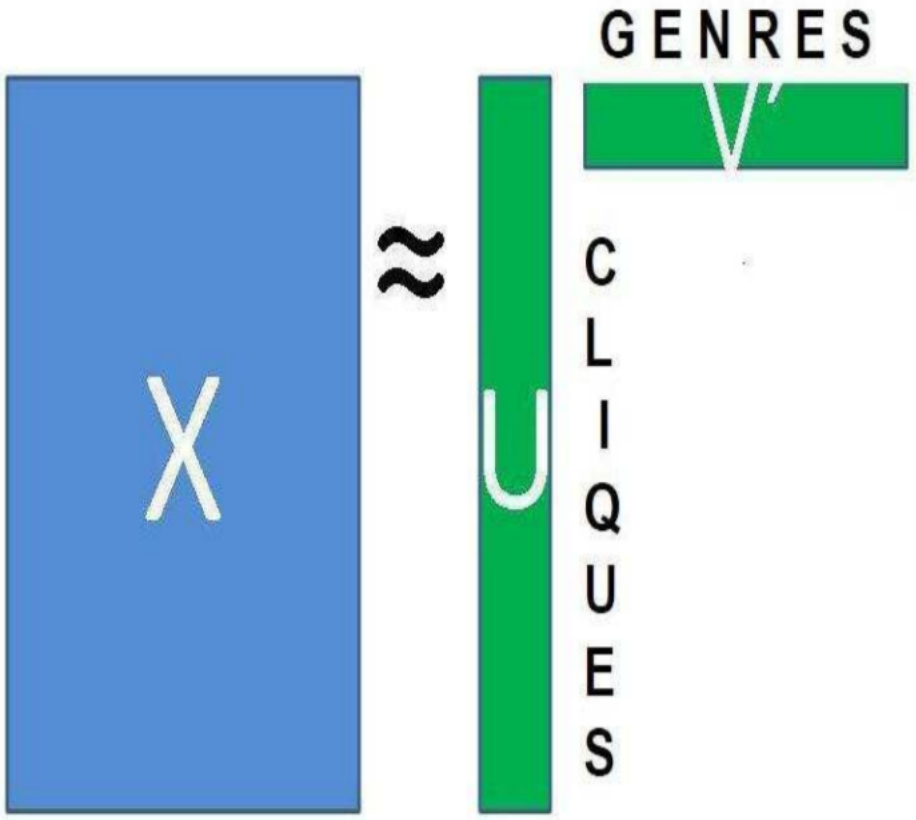
This method was popularized recently in the "Netflix Problem" – which essentially predicted missing entries in a matrix of user ratings for movies, where the rows were users and the columns were movies. This set up is instructive as it closely lines up with our model; that problem predicted reviews based off user and movie from past review data; our problem predicts grades based off student and class from past grade data.

The method essentially works by grouping similar users/students and movies/courses. The below figure shows how the matrix X is taken from Trevor Hastie's exposition of the Netflix Problem. [1] Similar users/students are grouped into cliques; similar movies/courses are grouped in to genres.

Formally, consider a matrix of observed values with missing values, $\underset{n\times m}{X}$. Let Ω represent the set of pairs of indices of the form (i, j) where $X_{i,j}$ is observed. We assume that if all of the values in the matrix X could be observed, they would have the form $X_{i,j} = Z_{i,j} + \epsilon_{i,j}$, where $\epsilon_{i,j} \sim iidN(0, \sigma^2)$ random variables and Z is a matrix of with a small rank, typically $rank(Z) = r << min(m, n)$. We will further constrain the sum of the singular values of Z (referred to as the nuclear norm of Z, $||Z||_\star$) to cause the problem to be convex (and therefore easier to optimize and solve). Our estimate for the low-rank approximation of X is found by optimizing the following expression: $Z^{opt} = \underset{\{Z:rank(Z)=r\}}{argmin} \sum_{(i,j)\in\Omega} (X_{i,j} - Z_{i,j})^2 + \lambda||Z||_\star$

We used an algorithm called Soft-Impute to perform this task. Soft-Impute uses iterative optimization to provide successively better and better approximations to the low rank matrix

Z until it has converged and outputs Z.



3.2.2 Data Pre-processing and Subsetting

To do "matrix completion," the data had to be further preprocessed so as to form a large matrix (with many empty entries). Each row corresponds to a unique student in the dataset; each column corresponds to a unique course in the dataset. Individual elements in which a student had actually taken a given class were filled in with the grade they received; individual elements in which a student had *not* taken a given class were left missing ("NA") – to be filled in with the completion technique.

Moreover, in order for matrix completion to be both computationally feasible, and have relatively accurate predictions, we need to subset to specific sub-populations within the the

student body. Our analysis chooses to focus on Engineering students: specifically, students who have taken at least 1 course in the Engineering department. Moreover, within this sub-population, the range of courses are restricted to the 200 most common engineering courses. With our population so restricted, we have far fewer missing values in our matrix to complete, and so we are much better able to impute grades.

3.2.3 Methodology

The matrix completion algorithm used the industry-standard "SoftImpute" package in R [2], which iteratively minimized the nuclear norm of the approximation matrix.

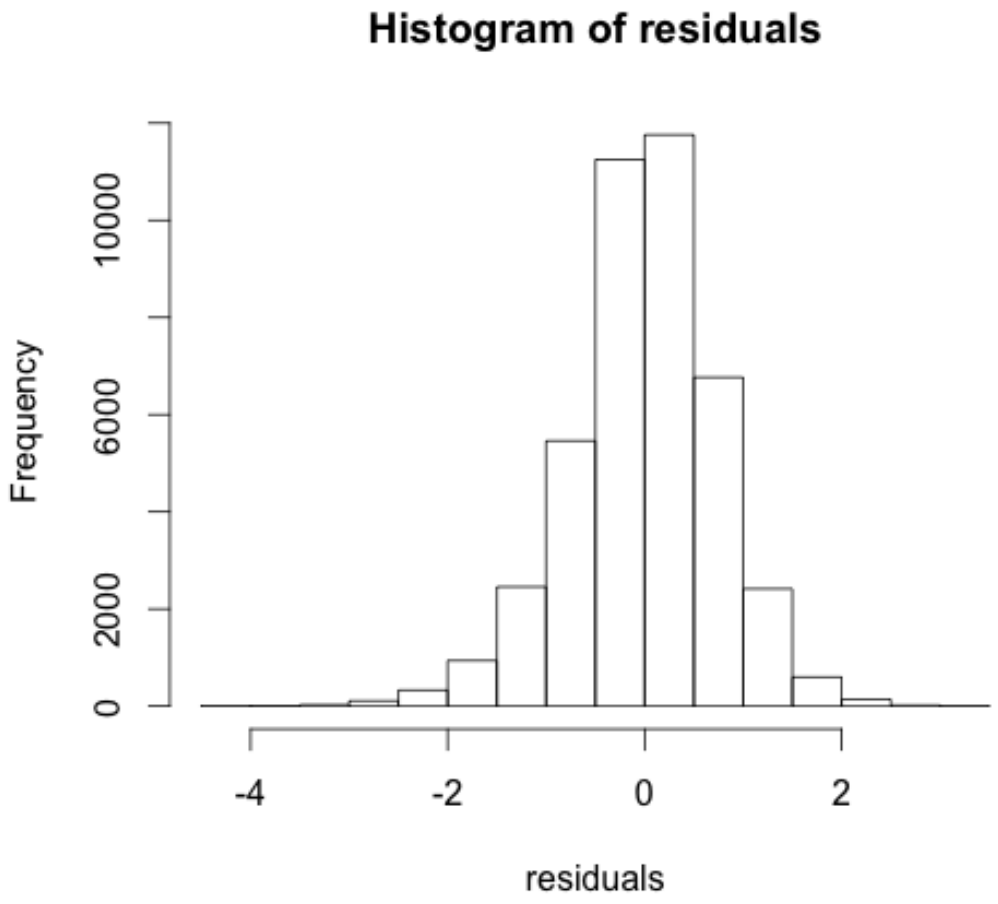
To assess our model, we selectively left out roughly 25% of the observed data, to test against our model predictions. (Since each row – i.e., student – had different number actual class grades, this was done row by row so as not to have all missing data for a given row or column.)

As an example, we'll pick a student at random, and look at some of the observed and predicted values for them (i.e., in their row).

	Observed Grade	Predicted Grade
ENG 101	2.3	
ENG 208	2.7	
ENG 209		2.86
ENG 261		2.81
ENG 282		3.05
ENG 331	3.3	
MA 242	2.7	
MA 302		2.69
MA 305		1.73
MA 341	1	
MA 401		2.05
MA 405		2.13
ST 311		2.21
ST 370	2.3	
ST 371		2.66

For this particular student, we see that they had a range of different grades over these classes: they seem slightly stronger in the Engineering courses, but decidedly weaker in Math and Statistics. Our model seems to give them reasonable predicted grades: they seem to do best in Engineering, and weaker in Math (particularly in the upper level course.)

For the data that we left out to test against, our residuals are displayed below. Note that the residuals have a bell-shaped curve – a feature that we did not impose on the data. Thus, our assumptions about normality are justified.

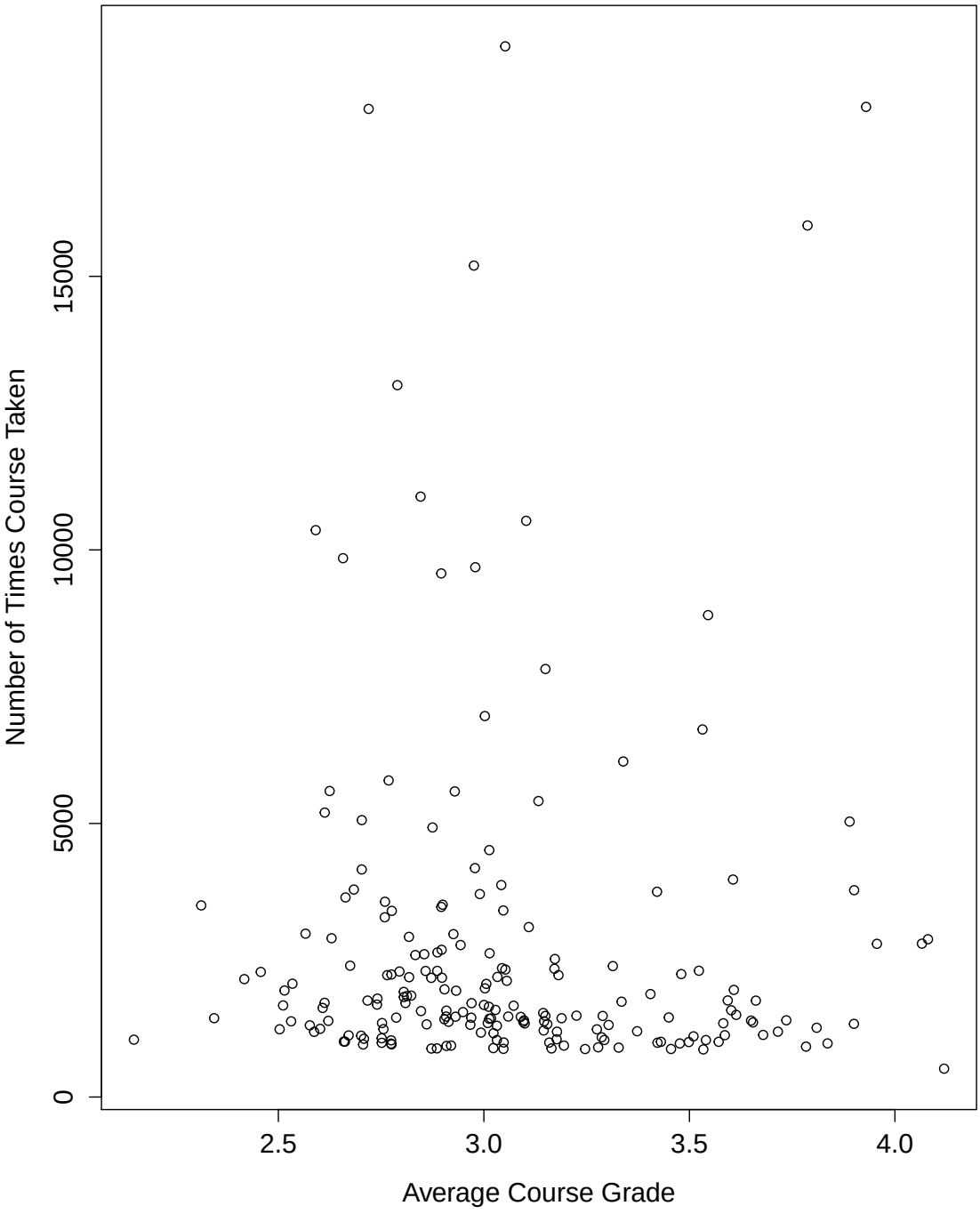


3.2.4 Results

One concern with our resulting completed matrix, is that the average class grade is significantly lower than the average class grade of the observed class grades. This could be a problem with the methodology, but it could also be accurate: perhaps students tend to take the classes that they will do well in, and don't take the classes that they would do worse in.

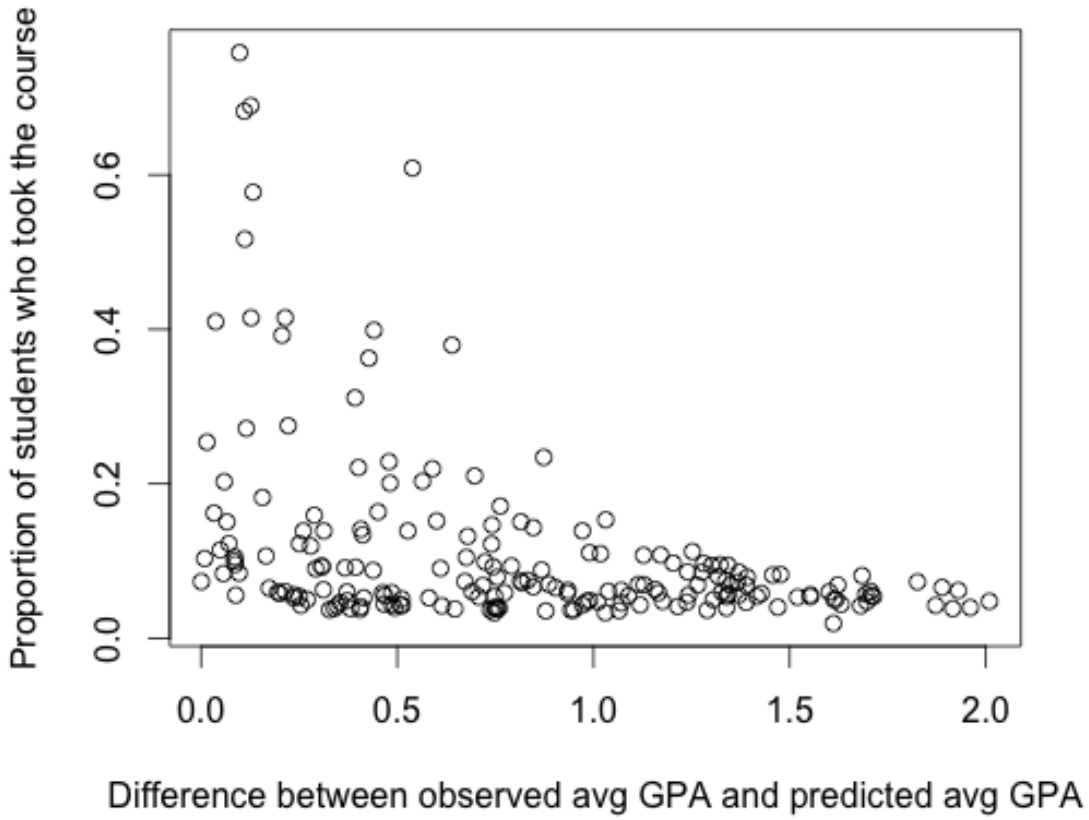
To investigate this, we looked at the average actual/imputed grade level for each course – which is a good proxy for course difficulty. Specifically, we wanted to see whether the rate of students taking courses was related to their difficulty. E.g., perhaps more students opt

to take the easier courses. If this was the case, that would validate our model giving lower values. The below plot shows this distribution of Frequency of students taking courses across course difficulty, and shows that this doesn't *not* appear to be taking place. Students tend to take the harder courses and easier courses at similar rates. Thus, the lower average values for our imputed grades is still a cause for concern.

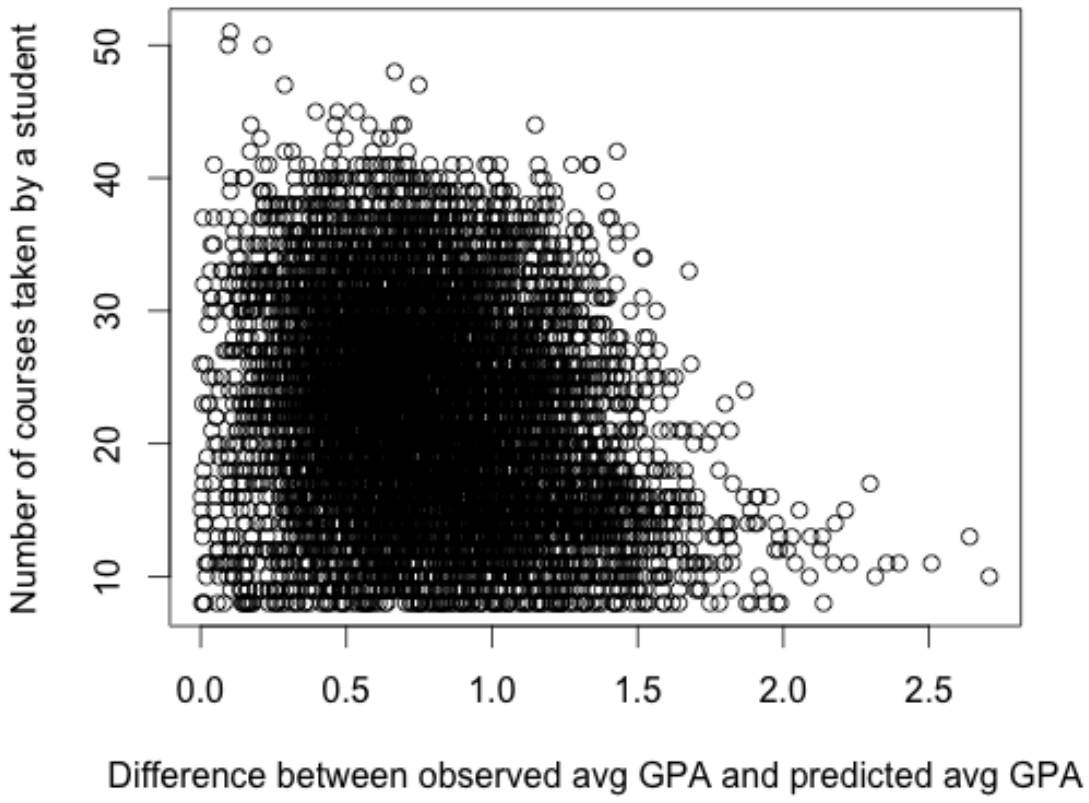


We further investigated this problem, by looking – for each class – at how the difference between average over observed grades compared and the average over predicted grades.

What we found was encouraging: classes that have more observed data tend to give more accurate predictions, as shown below.



This plot would seem to suggest the for courses that have less than 20% observed data, our model does not perform well. Similarly, we also looked at the difference between average over observed grades compared and the average over predicted grades, within individual students.



We further compared our model to a model that just uses the row means, as well as a model that just uses the column means, and model using one grand mean – to verify that we are actually making an improvement over these very simple models. Encouragingly, we found that our model handily beat all of these:

- Our model had an MSE = $.577 \times MSE_{\text{grand mean}}$
- Our model had an MSE = $.690 \times MSE_{\text{column means}}$
- Our model had an MSE = $.828 \times MSE_{\text{row means}}$

4 Limitations

- Large amount of data broken into a number of data sets: there was significant pre-processing required.
- A number of latent confounders that are too numerous to reasonably take into consideration (e.g. effect of professor on course grade, curriculum changes over the years, students leaving major or school for nonacademic reasons, etc.). Plan: it would be impossible to even think of every confounder. Thus, we simply will have to urge anyone using our results to advise students that this is simply one tool, not an end-all solution to an individual student's academic issues
- We will leave the bulk of interpretation of these results to Dr. Floyd, as we lack familiarity with the undergrad courses and curriculum.

5 Recommendations for Client

Matrix Completion is a relatively new and fast-growing field. Should the Matrix Completion approach be of interesting to the Office of Institutional Research, the next step for this project would be to have a statistician doing research in the field of Missing Data, or interested in Matrix Completion, to elaborate and improve on the methods presented in this report. Such a statistician could possibly create an application in R for academic advisers to plug in student's grades and output completed matrix of grades in other courses of interest, with confidence intervals for each class.

6 References

[1] Hastie, T (2012) "Matrix Completion and Large-scale SVD Computations." *(talk)* http://web.stanford.edu/~hastie/TALKS/SVD_hastie.pdf

[2] Mazumder, R, Hastie, T, Tibshirani, R. (2010) "Spectral Regularization Algorithms for Learning Large Incomplete Matrices." *Journal of Machine Learning Research.* 11, 2287-2322.