**Carnegie Mellon University**

**Dietrich College of Humanities and Social Sciences**

# Predicting Undergraduate Passions:

# An Analysis of Major Migration at Carnegie Mellon

---

Kelsey Dietz – Senior Honors Thesis – Spring 2015
Advisors: Baruch Fischhoff and Russell Golman

*Special thanks to Mr. John Papinchak and Dr. Gloria Hill for their support in providing the data for my research.*

## Abstract

With six undergraduate colleges, Carnegie Mellon University offers a plethora of choices for primary and additional majors, as well as interdisciplinary programs. Such variety has the potential to assist or hinder students' undergraduate careers, dependent upon how often their interests change, and how quickly they find their fit at CMU. This paper discusses the potential reasons behind major migration at CMU, and whether or not we are able to predict the likelihood that a given student will switch their major before graduation. By looking at the Class of 2015 cohort over twelve semesters, we run linear discriminant analysis and logistic regression on a representative sample of 1136 students. Through these methods, we determine that predicting student migration yields high error rates when working with a small subset of binary demographic variables, yet there is potential for a stronger prediction algorithm with more data and more robust variables such as cumulative QPA and socio-economic status. We also focus on the predictive capability of Dietrich College first-year survey data, and find that additional variables such as the number of interests incoming students have and which interests become their graduating majors are significant in classifying potential migrators. In both cohort and survey data, we find that the time it takes to initially declare is an important variable in determining whether or not a student will switch majors. The likelihood of switching is greatest if one initially declares in the spring of their first year, and drops off thereafter. This suggests that advisors should work closely with students to determine if it is the right time to declare. Declaring early can give students access to classes within the primary department, and declaring too late could leave students stuck in a major they are not truly passionate about.

## Research Question

Undergraduate majors have an invaluable impact on academic experience, field of study, and eventual career or graduate school decisions. The choice belongs to the student, so I hypothesize that we can characterize groups of students by when they choose their major, what major they choose, and demographic factors, in order to predict whether they change their major after they have declared. Major declaration and retention are two commonly studied topics, yet too often they are studied separately. A number of studies discuss how to increase retention in various departments, and others look at how high school and first-year students can make a better transition into their university major programs. In my research I hope to capture an aggregate analysis of major declaration and retention over time at Carnegie Mellon. My goal is to help those departments where declaration occurs very late, or departments with large attrition rates, to determine what practices they can implement to give their students a better understanding of the major so that students feel more informed before making their decisions and departments will see lower migration rates. I am also interested in determining if a few key demographic variables can be powerful predictors of major migration.

## Existing Research

The current research on major declaration and retention can be divided into two distinct categories: analyzing and providing solutions to attrition problems in a specific program, and how general student welfare affects academic performance and major choice. There have been numerous studies at Carnegie Mellon with regards to attracting women to Computer Science and retention of women in the major. Fisher and Margolis (2002) did a study on Women and Computing at Carnegie Mellon University that shows the impact of academics and the industry on women choosing to major in computer science and sticking with that major. It discusses the stark differences in women versus men studying Computer Science and the various attempts to close that gap. Other studies have shown that pre-college exposure to computer science affects the number of men versus women that enter the field. Lastly, a study that J. McGrath Cohoon (2001) ran at the University of Virginia categorizes the various factors that affect attrition and switching in men and women, including faculty support, same-sex support in the field, and mentors in the field. All these studies focus primarily on one demographic group in one major.

Another study of specific departmental attrition at Carnegie Mellon University focused on how the structure of the Fundamentals of Mechanical Engineering course affects the number of students who remain in the major (2013). The study focuses on how to design a course that combines lab work and lecture to keep students actively engaged. Both the research on specific demographic groups and specific major courses fail to examine where students go after they leave the specific major. While some research has noted that students should not be considered a part of an attrition study if they are leaving to pursue a stronger interest as opposed to a negative experience, these studies still do not examine the individual student's path from major to major, nor do they analyze whether students are all switching to the same new major. My research will highlight these factors as well.

The second branch of research is concerned with the general satisfaction of students, mainly first-years, in their academic experience and choice of major. A study done by Eric Jamelske (2009) shows how creating a satisfying first-year experience academically and socially can greatly affect a students' retention and satisfaction in a university program. This research only focuses on first-year students. Much of the research involving student satisfaction is survey-based and qualitative. For example, Chase and Keene (1981) performed observational research on how the length of time it takes for a student to declare a major affects his or her motivation. They found that the longer a student waits to declare, the less motivated he or she is. This research was not conducted at Carnegie Mellon and therefore did not consider that in some colleges, like CFA, students declare before entering their first year, while Dietrich College students are not allowed to declare until their spring semester at the earliest. Such differences could affect motivation, satisfaction, and retention. Overall, there is a wide variety of research on major satisfaction and retention, but my research will expand the currently available research in two key areas.

## Key Differences

There are two major differences between the research that exists for major declaration and retention, and my proposed thesis. Carnegie Mellon differs from the traditional university in that it acts as a system of interconnected colleges. Carnegie Mellon has seven distinct colleges that

have within-group differences (ranging from Economics to Creative Writing in Dietrich, for example), between-group differences (Chemical Engineering in CIT versus Vocal Performance in CFA), and connecting networks (SHS, QSSS, and BXA, to name a few). Trying to study the flow within colleges as well as across colleges is a key factor of my research. We assess not just one department's retention success, but also multiple departments across multiple colleges. My aggregate analysis looks at all of the information I can query from the registrar database and focuses on a few specific departments or groups of interest. Another primary difference of my research is that it does not stop tracking the student once they leave a specific college. The studies on women in Computer Science or Mechanical Engineering majors in CIT focus on how and why students enter and leave that one college or major, but my research will also focus on where they go when they leave. For example, if a woman in computer science leaves to study physics or mathematics (also male-dominated fields), then the reason for leaving might be very different from a woman who leaves SCS to study Psychology or even Statistics. These are important factors of major flow and retention that should be tracked in order to potentially influence the way schools market themselves, or the policies they have for students' abilities to work across colleges.

Through my research, I hope to be able to address departments with high attrition rates and offer insight into which students are leaving. I have taken a quantitative approach to assessing major migration to set my research apart from the qualitative analysis that exists in the field. For example, I hypothesize that certain factors such as initial department and presence of additional majors can lead to an increase in the probability that a student will change majors over the course of his or her undergraduate career. Lastly, I analyze how initial interests of incoming freshmen predict their ultimate major upon graduation. These are the primary goals of my research and my main contributions to the Carnegie Mellon community.

## Overview

Through my research, I studied how a standard cohort at Carnegie Mellon behaves from arrival through graduation. I examined the Class of 2014 cohort and how individuals' primary colleges, departments, and majors changed from year to year, along with consistent demographic factors and student traits. The three questions I look to answer are:

1. **LDA/Logistic Regression/Random Forests:** What is the risk of attrition/migration associated with each department and major (and what predictor variables influence that risk)?
2. **Additional Majors/Dietrich College Survey Data:** Where do students migrate when they do? (What pairs of majors tend to have more connections, including additional majors)?
3. **Dietrich College Survey Data:** Are predictors such as Dietrich College Surveys and Introductory Courses associated with graduating major?

Through my analysis of these three questions, I built a case for classifying students who are likely to switch, which has the potential to be employed by advisors, departments, and colleges at Carnegie Mellon to ensure greater retention among students when possible, and to allow students to find their fit sooner.

## Data

The data I chose to look at is from the most recent completed cohort of CMU students, those who graduated in the spring of 2014, ranging twelve semesters. This allowed me to look at the class of 2014, but also class of 2013+, in order to get an accurate representation of the Architecture major, which is a five-year program. I observed which factors comprise the initial distribution of majors upon entry or first declaration, and how these traits might affect the migration rates out of various majors and departments. I make use of demographic variables of race and gender, as well as students' U.S. Citizenship status. Other predictor variables of interest included Greek affiliation, Athletics affiliation, and F1/J1 status.

Lastly, using data from the Carnegie Institute of Technology (CIT) and the Dietrich College of Humanities and Social Sciences (DC), I look at the initial major preference of students who enter Carnegie Mellon undeclared. For CIT, I define my instrument to be the first semester introduction to engineering course the student takes. For example, if a student enters CIT and begins his or her first semester by taking Introduction to Mechanical Engineering, I will assume his or her intention was to be a Mechanical Engineering major. For Dietrich College, students submit an initial survey on majors of interest. I look at how all these choices potentially influence major upon graduation.

## Tools

I analyze major migration and risk of attrition through statistical classification and logistic regression. By using linear discriminant analysis, I assess the predictive power of my demographic variables and other qualifiers, including length of time until first declaration, and their ability to predict whether a student will switch their primary department during their undergraduate career.

I can also answer my second question by changing this response variable to the number of students who graduated in a particular department divided by the number of students who initially declared in that department, in order to assess the pull that a particular major has. The downside to this analysis is that there are such a wide variety of majors and departments at Carnegie Mellon that I have sparsity issues that affect the quality of my predictions. Looking at entire departments as opposed to individual majors helps to combat this sparsity issue, so my statistical models will have less variation.

By bringing together data from multiple sources, including registrar data from six years prior to the graduation of the class of 2014, I have both visually and analytically classified major migration at Carnegie Mellon. Through linear discriminant analysis, logistic regression, and random forests, I classify students and identify significant predictors for major migration. I also visualize the data in Tableau, allowing me to piece together once disparate datasets to create a full picture of each individual student. This included aggregation of semester major and department information, additional major data, majors of interest for students in Dietrich College, and numbers of students in various Carnegie Institute of Technology introductory courses.

## Initial Sample

The queried data includes 1228 students who graduated in the year 2014 (Mid-Year, Spring, and Summer). The breakdown of gender is 42.3% Female and 57.7% Male. 88.0% of the population has U.S. Citizenship and 12.0% have Visas. 32.3% of students in the 2014 cohort participated in Greek life. As we can see in the figures below, when we look at graduating majors, the most common major was Computer Science, followed by three departments in the Carnegie Institute

of Technology: Electrical and Computer Engineering, Mechanical Engineering, and Chemical Engineering. All these statistics are representative of the expected distributions from the Institutional Research Analysis for a Carnegie Mellon Cohort.

Looking at the number of students in each department over the years, we notice that while Mechanical Engineers, Electrical and Computer Engineers, and Business Administration Students remain relatively constant, the amount of students graduating in the department of Computer Science rises dramatically over the course of the standard four-year graduation period. I would hypothesize that exposure to Carnegie Mellon's computer science program attracts more students towards the major. It is also possible that moving into the Computer Science department is easier once one is at Carnegie Mellon than it is when one first applies. This would imply that students trying to enter the major through other departments would have an easier time doing so than students who initially applied to Carnegie Mellon. Also, students who are in the School of Computer Science as a college are expected to graduate as Computer Science majors, thus they may not need to declare at any point before graduation. I also notice a steady rise in the number of students who enter the Bachelor of Humanities and Arts department, thus it appears this combination gains popularity as students move throughout their careers.

The top left graph on the next page displays the majors that had the highest number of graduates. The graph on the top right displays the departments with the highest number of graduates, including the male to female ratio. As we can see, the Carnegie Institute of Technology graduates 64.3% males, while Dietrich College has a more even split, with 45.3% male graduates. Lastly, the bottom graph displays the number of students in each department on a semester basis, to give you a sense of when students are declaring and how the overall landscape changes. You can see the steady rise in computer science majors declaring over time. Engineers and Business Administration stay relatively constant over time.

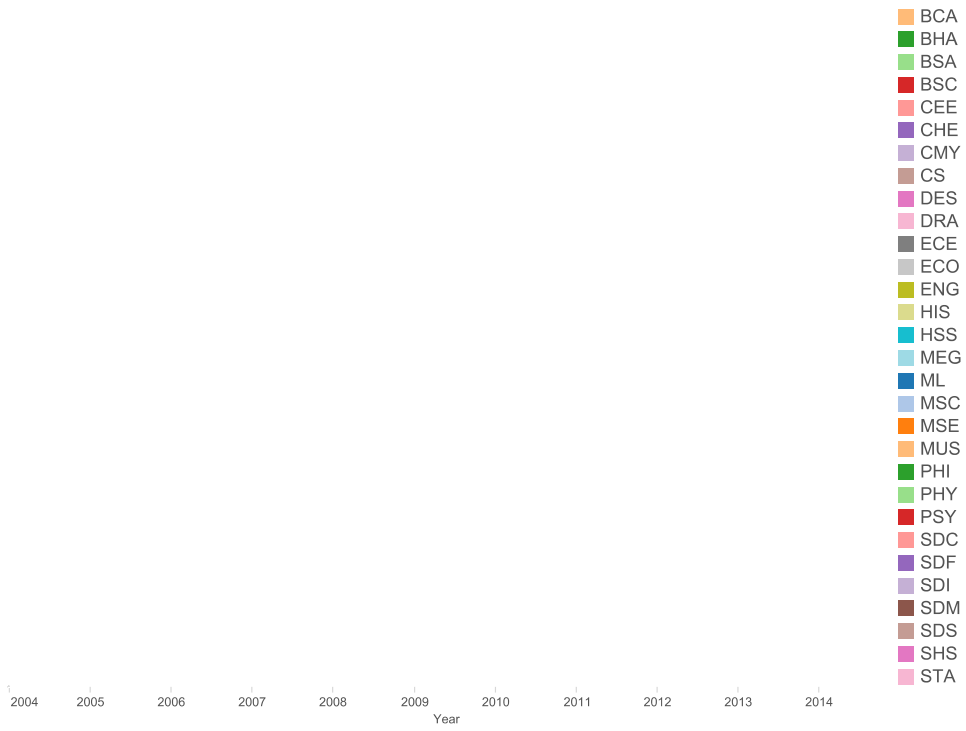



**Figure 1: Students in each major, department, and department over time. A tabular version of department over time is available in the appendix.**

My initial hypothesis was that the Chemical Engineering major would show greater attrition than its engineering counterparts over time, primarily due to the Introduction to Chemical Engineering Course. This hypothesis stemmed from a qualitative survey I designed and administered via social media to seniors in CIT. The survey asked the class of 2015 with the assumption that the cohorts behave similarly, and was designed to determine why students in CIT switched majors, where 50% of the sample switched out of Chemical Engineering after taking the Introduction to Chemical Engineering Course. However, it is difficult to assess whether the correlation between Introduction to Chemical Engineering and leaving the Chemical Engineering major exists given the inability to link individual students to the introductory course. While there does appear to be slight drop-off in the number of students who major in Chemical Engineering over semesters, we cannot say how many students who were undeclared CIT in the fall were considering Chemical Engineering and then chose a different major after their freshman

introductory course. Thus, it is important to look at the quantitative data for CIT freshman introductory courses in order to gain more insight into the potential attrition of those students.

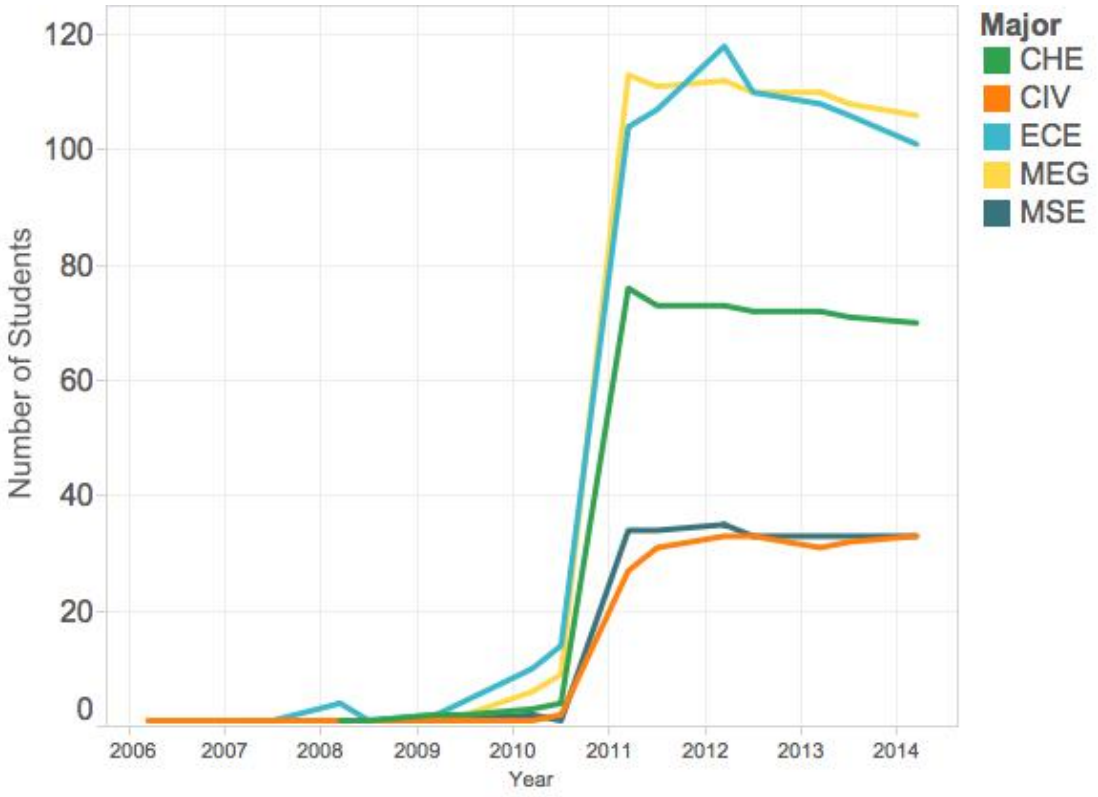



**Figure 2: CIT students declared in each primary department over time (no additional department data available for EPP and BME)**

Though I do not have the ability to connect Introductory Courses to specific students, I do have the total number of students enrolled in each of the freshman engineering courses at Carnegie Mellon in Fall 2010.

| Course Number | Department | Students | Grads |
|---|---|---|---|
| 12-100 | Civil and Environmental Engineering (CEE) | 68 - 35 | 33 |
| 06-100 | Chemical Engineering (CHE) | 87 - 17 | 70 |
| 18-100 | Electrical and Computer Engineering (ECE) | 140 - 39 | 101 |
| 27-100 | Materials Science Engineering (MSE) | 45 -12 | 33 |
| 24-101 | Mechanical Engineering (MEG) | 136 - 30 | 106 |

As we can see, there are a roughly equivalent number of students enrolled in ECE and MEG, which is concurrent with the distribution of majors found in the graph above. We also see that there was an over 50% decrease in the number of students who graduated with Civil and Environmental Engineering Degrees, compared with less than 30% reductions in all other majors. One large confounder is that the current data available to me includes all enrollees in the Introduction to Engineering Courses. Carnegie Institute of Technology expects students to take two introductory engineering courses, so it is quite possible that portions of the enrollees are not first-year students. It is necessary to look at the data for enrollment in Introduction to Engineering Courses by class year so that I can run more accurate analysis on major attrition in the department based on these intro-level courses. The revised table is below.

| **Course Number** | **Department** | **Students (First-Years)** | **Grads** | **Retention Rate** |
|---|---|---|---|---|
| 12-100 | Civil and Environmental Engineering (CEE) | 68 (52) | 33 | 63.5% |
| 06-100 | Chemical Engineering (CHE) | 87 (80) | 70 | 87.5% |
| 18-100 | Electrical and Computer Engineering (ECE) | 140 (123) | 101 | 82.1% |
| 27-100 | Materials Science Engineering (MSE) | 45 (36) | 33 | 91.7% |
| 24-101 | Mechanical Engineering (MEG) | 136 (130) | 106 | 81.5% |

Now we can see that the majority of students in these introductory courses are first-years, and the highest retention appears to be in MSE, followed by CHE, which goes against my initial hypothesis that their would be sharp decline in Chemical Engineering majors after first semester. The lowest retention is in CEE, and it would be interesting to go into more detail as to why this department had so much potential attrition. It could be that this introductory course has a reputation for being particularly engaging or easy, which could attract more first-year students who are not in the major to it. We could also assess the number of additional majors in CEE to determine if there are any students who took the course to qualify for the additional major in the department.

As we can see from the graph below, the majority of students who remained undeclared after their first year at Carnegie Mellon ended up majoring in Computer Science. This is mirrored by the steady rise in Computer Science Majors over the years, as shown in the previous graph of major declaration. Following that are Drama, ECE, and Information Systems. It is possible that the large number of undeclared Computer Science students is due to the fact that the School of Computer Science has one undergraduate major, thus students in SCS do not need to declare if they are expected to graduate as CS majors. This could be a potential confounder in determining attrition risk and the pull into particular majors, thus it will be important to take into account the differences in the School of Computer Science.

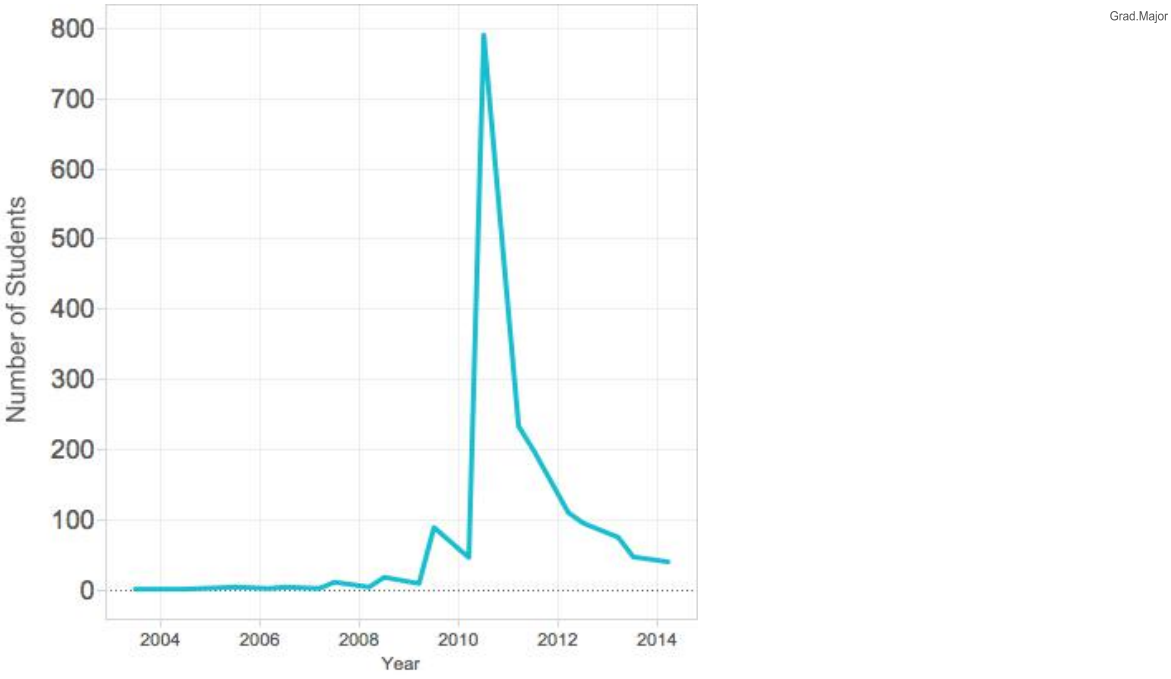



**Figure 3: Undeclared students over time (left), undeclared students after year 1 by department (right)**

Lastly, I note that I have data for athletic participation available to me, however, I find that only 21 students, or 1.71% of my population, was listed as involved in athletics. This data appears to be incompletely coded and will therefore not be included in my analysis.

I attempt to assess the risk associated with department change based on gender, U.S. Citizenship, and Greek affiliation, and whether some departments tend to pull from specific other departments more often. I use linear discriminant analysis and logistic regression to analyze the

probability that a student will switch majors based on initial status, initial declared major, and the demographic variables addressed above.

Using the attrition rates for all departments, we can look at how attrition rate is associated with original department, graduating department, and other characteristics such as gender and visa status. For example, the Electrical and Computer Engineering Department starts out with a larger proportion of males than females declaring, and attracts mostly males from Computer Science and undeclared CIT. Social and Decision Sciences starts out with a slightly more females, yet pulls in slightly more males from a wide variety of other departments, included Art, Business Administration, Psychology, and Mechanical Engineering. Thus, we could hypothesize that The probability of switching to the ECE department is greater if you are a male in Computer Science, while The risk associated with entering Social and Decision Sciences is uncorrelated with original department or gender.


**Figure 4: ECE and SDS departments: bars show initial departments for males and females who graduated in ECE (left), and SDS (right). This shows students migrate to ECE from two distinct locations, while SDS pulls students from all over into the department.**

## Additional Majors

Another important consideration in assessing major and department migration is additional major, department, and college, if a student graduated with an additional major. By linking student ID from additional major data to the original dataset, I was able to examine students who graduated with two or more majors. The two main objectives of this analyses were to determine if certain pairs of majors are more likely to go together, and if certain areas where I had originally designated a student as having switched a major were due to the primary major becoming the additional major. For example, the table below shows a student in the cohort who graduated with a degree in Business Administration in Spring 2014, though in Spring of 2012 she was majoring in International Relations and Public Policy. In my original analysis, this was categorized as a change of major and department, and was equivalent to all other major changes.

| Student.ID | Major.S12 | Major.Grad | Major.Grad.Add |
|---|---|---|---|
| 0613425* | INTRELP | BA | INTRELP |

*Student.ID is randomized to protect anonymity

Upon joining the additional major data, I found that while the student did change his or her primary major, the additional major was equivalent to the primary major designated in the Spring of 2012. This is a key piece of my analysis, because the student continued on with the International Relations and Public Policy courses, just not as a primary major. This distinction is significant, as students who switch majors and abandon the initial are different from those who switch primary majors but continue on with the original major as a secondary. They are also different from students who take on an additional major that they treat as their primary, regardless of how the two are listed on their graduating degree. When I run regression analysis on the risk of switching out of SDS, students like 0613425 should be viewed differently from students who do not hold an additional major in their original department.

Additional major data also presented interested findings when it came to looking for common pairs of additional majors. For example, the graph below displays the College CIT, with panes representing graduating departments and additional major departments within each graduating department. What I found was that students in Chemical and Materials Science Engineering were far more likely to have additional majors, most often in BME. Mechanical Engineers also have a

high percentage of additional majors in BME, but have many EPP additional majors as well. Civil and Environmental Engineers most commonly earn additional majors in EPP. These connections are important to our analysis because they represent common pairs to look out for when students migrate within or between departments. They could also help explain why students switch primary departments (potentially because of interest in an additional major that has more overlap with a different department than their original).

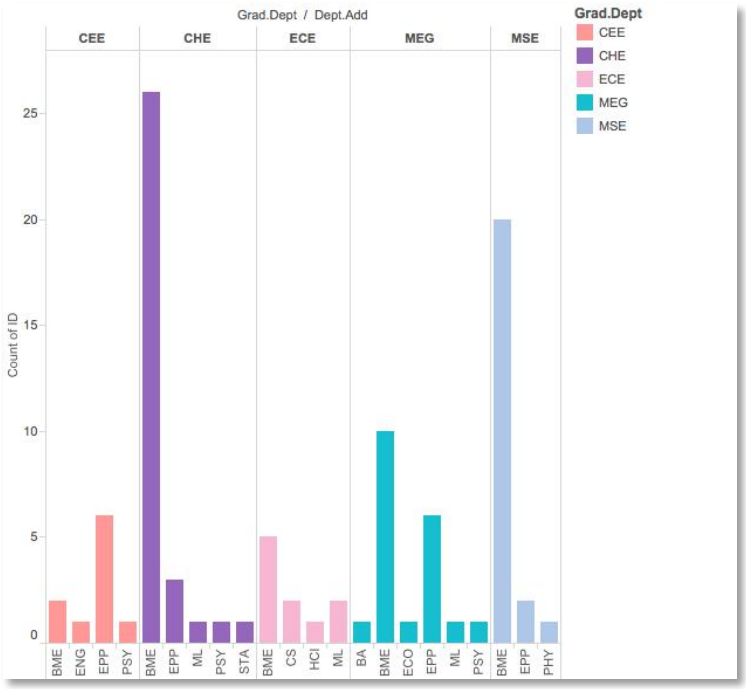



**Figure 5: Additional majors held in each department upon graduation (CIT). The most common additional majors are BME and EPP, with Civil and Mechanical Engineers most likely to hold an additional major in EPP, and Chemical and Materials Science Engineers most likely to hold an additional major in BME.**

This is especially important when comparing additional majors outside the original department (for the purpose of assessing partial versus full migration). For example, CFA's additional major pairings are almost exclusively from HCI, outside the department and college. This pairing is especially common for design students.

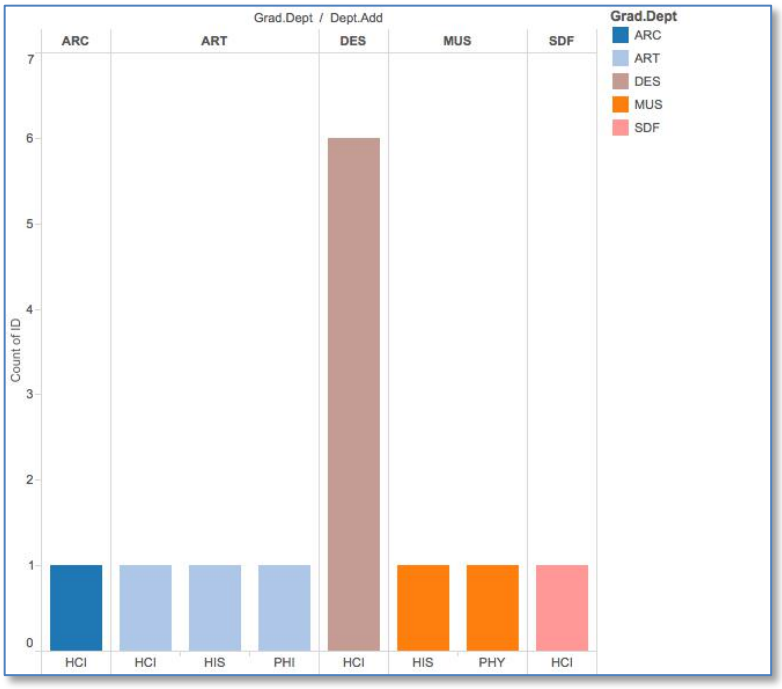


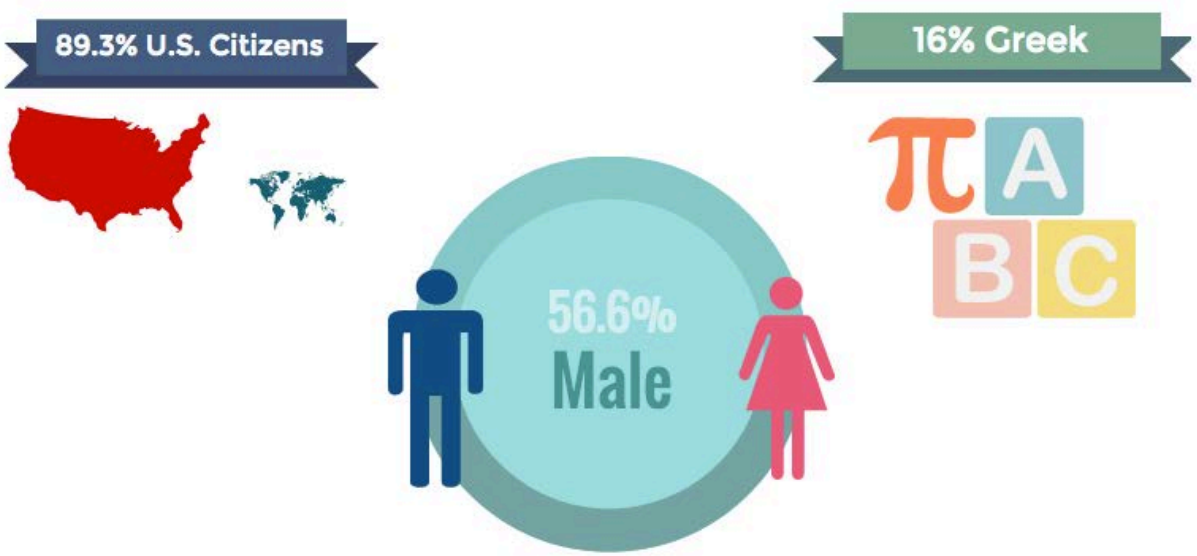

**Figure 6: Additional majors held in each department upon graduation (CFA). The table above shows only additional majors outside of the College of Fine Arts, with the most common additional major being HCI.**

## Definition of New Variables

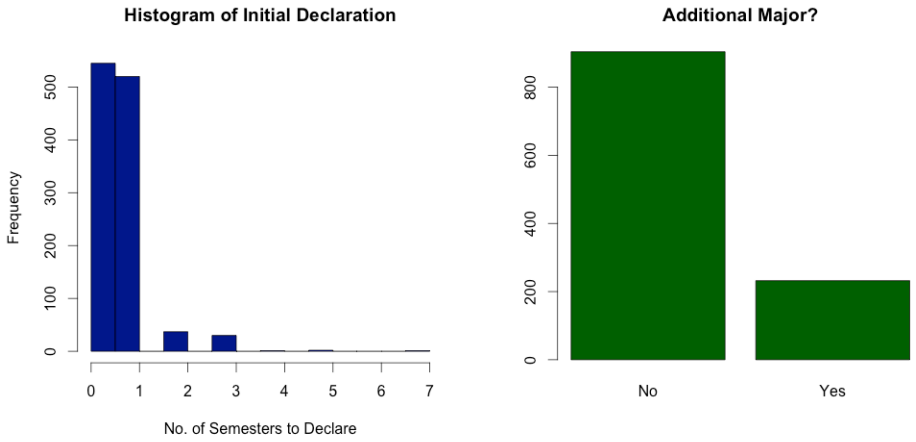
In order to analyze the probability of migration we must first define what it means to migrate from a major or department. I have chosen to classify department switches into three distinct categories: no switch, partial switch, and full switch. We define the no switch category to contain students whose initial declared department is the primary department upon graduation. A partial switch is defined as a student who switched their primary department but kept an additional major in their initial department. This could be the case when a student chooses to switch his or her primary major, but has completed enough of the initial major that obtaining an additional major does not take much more effort. Lastly, a full switch means that the primary department changed and no additional major was kept in the original department. For the purposes of our data, we code no switch as 0, partial switch as 1, and full switch as 2.

## Preliminary Data Analysis

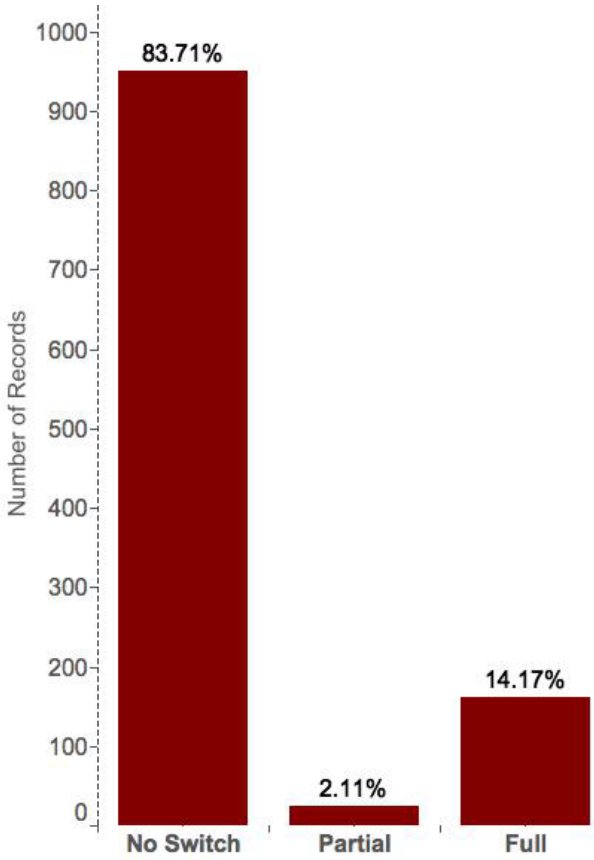
After cleaning the data, we end up with 1136 rows of student data indicating initial declared department, graduated department, our demographic variables, and additional qualifiers such as the number of semesters before a student first declared a major. Initial analysis of the variables shows a similar breakdown to the whole dataset of gender, citizenship, and participation in Greek life.




We can use our data on additional majors and initial primary departments to create two additional indicator variables. The first is a binomial indicator of whether or not a student holds an additional major. The second is a quantitative variable of the number of semester it took a student to declare their initial primary major. The summary of these variables is displayed below.



As we can see, the majority of students declared their first major in their first or second semester, and 20.42% students graduated with an additional major.




We also have access to a new variable, "switch," which determines no, partial, or full switches for each of our 1136 students. As we can see, 14.17% of students fully switched departments prior to graduation, and only 2.11% of students partially switched.

This means that approximately 1 in 6 students switched their primary department at some point in their university career. My goal is to classify the groups of students that switch and determine if it is possible to predict the class a student will end up in, or the probability that a student will switch their major prior to graduation, given the exploratory variables available to me.

We can also look at the departments that attracted the most individuals. Social and Decision Sciences and Computer Science attract the most total switches (full and partial). We also see many students from Economics switching into Business Administration and many Electrical and Computer Engineering students switching into the Computer Science department. Now that we know which departments attract and which are losing students, we can move on to our first method of classification.

| Grad Dept<br>First Dept | BA | BHA | CEE | CS | HSS | MSC | PSY | SDS | STA |
|---|---|---|---|---|---|---|---|---|---|
| ARC | 1 | 3 | 5 | | 1 | | | 3 | 1 |
| BA | 81 | | | | 1 | 3 | | 1 | 2 |
| ECE | | | | 13 | 1 | 1 | | 1 | |
| ECO | 12 | | | 1 | 1 | 1 | | 4 | 7 |
| MSC | | | | 7 | | 38 | | | 1 |
| In Total | 17 | 8 | 7 | 32 | 9 | 12 | 8 | 22 | 14 |

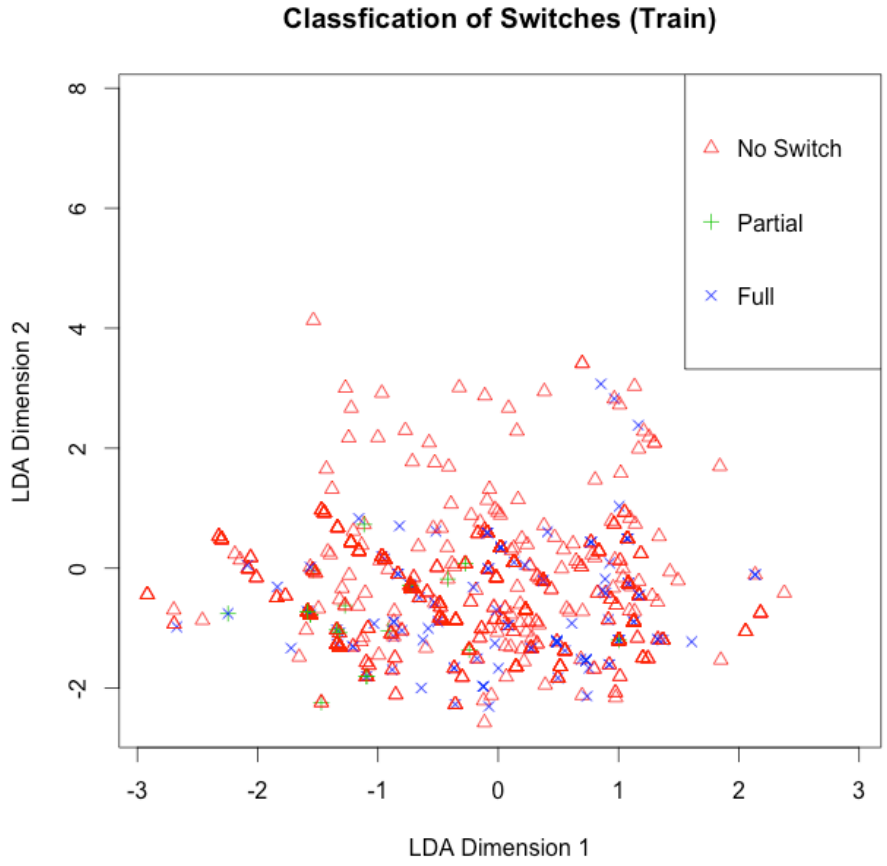
**Figure 7: Total full and partial switches from initial declared department to graduate department.**

## Linear Discriminant Analysis

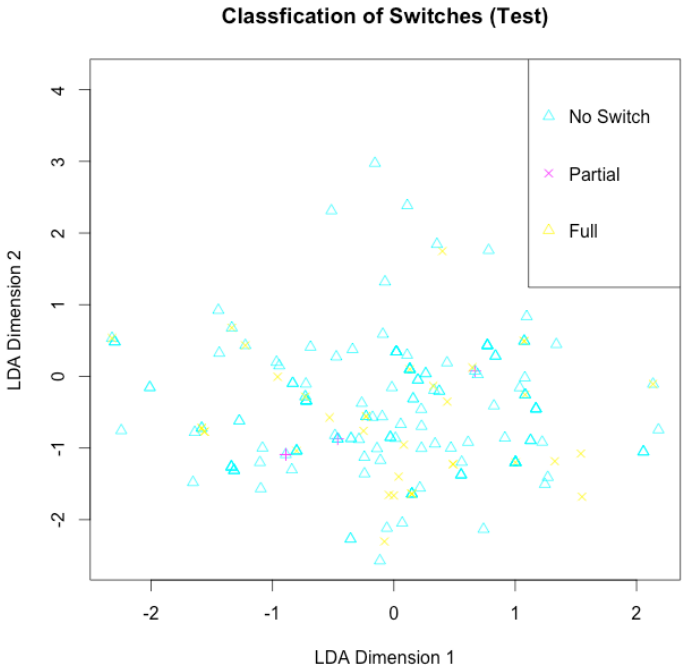
Our first approach is to use linear discriminant analysis (LDA) in order to classify students as full, partial, or no switch, based on the demographic variables we are given, as well as some we have created. These additional variables include *Has.Add*, which tracks whether a student has an additional major, and *When.Declare*, which is a quantitative variable indicating the number of semesters (Fall and Spring) it took a student to declare his or her initial primary department. We randomize our data using resampling of rows, and divide the data into a training and testing set, so that we can explore the predictive capabilities of LDA on data that we already have.

Linear discriminant analysis attempts to divide the data into the three separate classes we have defined, using linear boundaries in a dimension-reduced space. After running LDA initially on our training set, focusing on gender, U.S. Citizenship, participation in Greek life, initial department, time took to initially declare, and presence of additional major, we find the following map of classification for switches. The following graphs take our multi-dimensional data and reduce it into a two-dimensional space based on how different each datum is. This should create clusters of points that are similar based on similar predictive features.

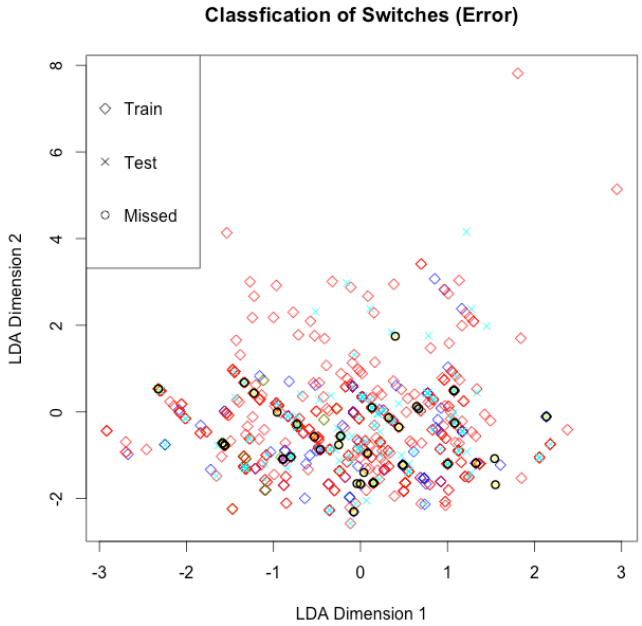
LDA graphs are a great way to look at the Euclidian distance between multi-dimensional vectors of points. Here, we take our multivariate data and compress it into two arbitrary dimensions. To properly interpret these graphs, do not look at where each point lies on dimension 1 and dimension 2. Instead, look at the closeness of points relative to each other, or specific clusters of points that are far away from the rest of the data. For example, in the graph below we see a dense cluster of points in the bottom center of the graph, and a less dense, wide cluster above that. There are also two small outlier groups on the left and right hand sides of the dense center cluster. We could hypothesize that students are very similar in general (potentially due to the large number of U.S. citizens and smaller subsets of students with additional majors or who participate in Greek life.




As we can see, there is a lot of overlap between our three classes, and there does not appear to be any distinction between where “No Switch” students cluster, and where the majority of “Full Switch” students cluster. The misclassification rate on our training set is 15.8%, which is no better than if we randomly guessed that no students switched departments. We can also determine if this algorithm for classification would work when presented with new data. Thus, we give it the same variables from our test set, and once again we graph our LDA on a two-dimensional scale. We created our test set by randomizing the rows in our data and pulling out a sample of 20% of the rows. These random rows are meant to represent new data we might be presented with in future cohorts of students. We create our algorithm on our training data, and test how well it works on data it has not seen before. This is an indicator of how robust our prediction function is, and whether we will be able to externalize it to future groups of Carnegie Mellon students.




The graph above shows that our test data appears to have the same general mapping as our training data. If we overlay our two graphs we can see where they match up and where our errors were in classifying switches.



In the graph above, the majority of errors occur in that central cluster where there exists a lot of overlap between the variables. We often misclassify those points, which are very similar along the predictor dimensions we have used. This is a difficulty with having a large number of binomial categorical variables: there is little variation in our predictor variables, which leads to a smaller set of criteria on which to classify our students. Calculating the testing error, we find that it is slightly higher than our training error.

| **Training Error** | **Testing Error** |
|---|---|
| 15.8% | 18.1% |

Our training and test error are both near equivalent and unfortunately they perform worse than the classification error we would achieve by simply guessing that all students stayed in their primary department (because 15.8% of the population switched departments). As we can see from our linear discriminant analysis, there is room for improvement. When training and testing error are very close, we are typically concerned with high bias or data sparsity, which could be improved by changing the method of analysis we use to analyze the data or obtaining more data. In this case, we have data on an entire cohort of students, so adding more data would not be a feasible solution. However, it is also possible that the low number of variables we are working with (and the loss of incomplete variables like athletics) is affecting the accuracy of our predictions.

Our concern with high bias could be solved if we move from linear discriminant analysis to quadratic discriminant analysis (QDA), which attempts to cluster the data into non-linear fields. When looking at our clusters we notice that they appear to be Gaussian in nature, with a large cluster of “No Switch” data points in the center of our mapping, and tails on either end. This sort of mapping could benefit from QDA. However, because the majority of our variables are binomial factors, in order to run QDA we need to remove low-rank variables from our calculations. The fewer variables we have to analyze leads to higher error rates, so there is a trade-off between using powerful variables and reducing bias in our analysis. When we remove the low-rank variables from our data, our QDA classification error becomes 18.1%, which is no different from our LDA misclassification rate for our training set (though it does achieve the same level of accuracy with fewer variables). These high error rates mean we have to look to other methods if we want to improve our predictive capabilities.

## Logistic Regression: Assessing the Probability of Any Switch

One method that could yield better results is logistic regression, by which we would determine the probability of a student switching out of their primary department given the demographic variables we were given and the response variables we created. Logistic regression would not take into account partial switches, as it deals with a binomial response, but it could prove to be a stronger indicator of the likelihood that a student will switch majors over the course of his or her time at Carnegie Mellon.

In order to run logistic regression, we first need to reclassify our response variable, Switch, as binomial 1, 0. We consider partial switches to be a switch, as they do involve a change in primary department, which is the area we are most invested in analyzing. After creating a model for our logistic function, we can analyze the training error by predicting the probability each student has of switching, and then sampling from a random binomial distribution with that probability in order to determine if they do. We then calculate our misclassification rate. First we can look at the summary statistics of our logistic regression to determine if any of our predictor variables are meaningful. Meaningful variables on our training set could give us insight into what predictors will be important in assessing the probability that future students at Carnegie Mellon will switch their primary departments.

| Coefficient | Estimate | Std. Error | P-Value | Significance |
|---|---|---|---|---|
| Intercept | -1.74 | 0.737 | 0.018 | * |
| Gender | -0.035 | 0.186 | 0.852 | |
| US.Citizen | -0.049 | 0.304 | 0.871 | |
| Greek | -0.086 | 0.243 | 0.723 | |
| First.Dept | 0.017 | 0.012 | 0.173 | |
| First.College | 0.050 | 0.048 | 0.292 | |
| When.Declare | -0.414 | 0.157 | 0.0083 | ** |
| Add.Major.Yes | -0.245 | 0.245 | 0.317 | |

We find that time took to declare is a significant variable in determining the probability that a student has switched primary department. This is good news considering the number of semesters taken to declare is a quantitative variable, so we can interpret its significance. In this case, each additional semester it takes one to declare will decrease the additive log odds that one

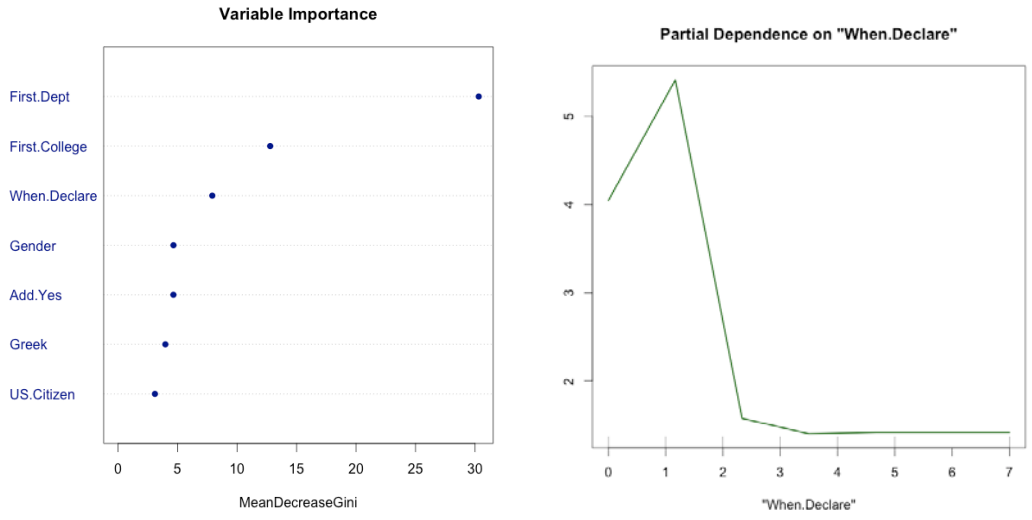
will switch majors by approximately 40%. However, just because When.Declare is quantitative does not mean it is linear. We will see through further analysis that When.Declare is actually quadratic. If we represent When.Declare as $1/(1+\text{When.Declare})^2$ and run another logistic regression, we find that the variable becomes even more significant (p-value of 9.48e-5*** for alpha < 0.001). We can also calculate the testing misclassification rate for this logistic function, which turns out to be approximately 25.1%. This is a higher error rate than our original LDA classification. Because logistic regression is determining a logistic relationship between data points, it suffers when we use many unordered categorical variables. LDA is much better at handling categorical indicators, and thus performs better in this case. We can also determine the cross-validated error of our log function to get a more accurate reading for how our logistic function would predict when given new values. This cross-validated mean error is 15.84%, which is a much better error, but still no better than LDA. Thus, we may need to look to one more method of classification to give us better predictive power. Logistic regression may not have given us a better method of prediction, but it did provide us with a significant variable to pay attention to in our future methods.

## Random Forests

A random forest is another classification technique that we can use to test the importance of specific variables in our dataset on classifying whether or not an undergraduate will switch his or her primary department. They allow us to create trees that split the data along small subsets of independent predictors (for example, the presence of an additional major). If we run this function on our training set, we find the out of bag (OOB) estimate of error to be 14.41%, which is lower than LDA and logistic regression, and also a better predictor than random guessing. If we then predict our test set using this random forest function, we find the following classification table. This random forest has a misclassification rate on our testing set of 17.2%, the lowest error we have seen thus far. It often classifies a student who switched as a “No Switch,” which is an error we have seen in all of our classification techniques. All of our classification techniques have overrepresented the “No Switch” category, which means that a portion of students who did switch are very similar to those did not. This could mean we do not have enough variables to distinguish that subset as different, or it could be telling us that those students would have been more typical and perhaps even better off had they stayed in their initial department.

| Predicted / Actual | No Switch | Switch |
|---|---|---|
| **No Switch** | 185 | 1 |
| **Switch** | 38 | 3 |

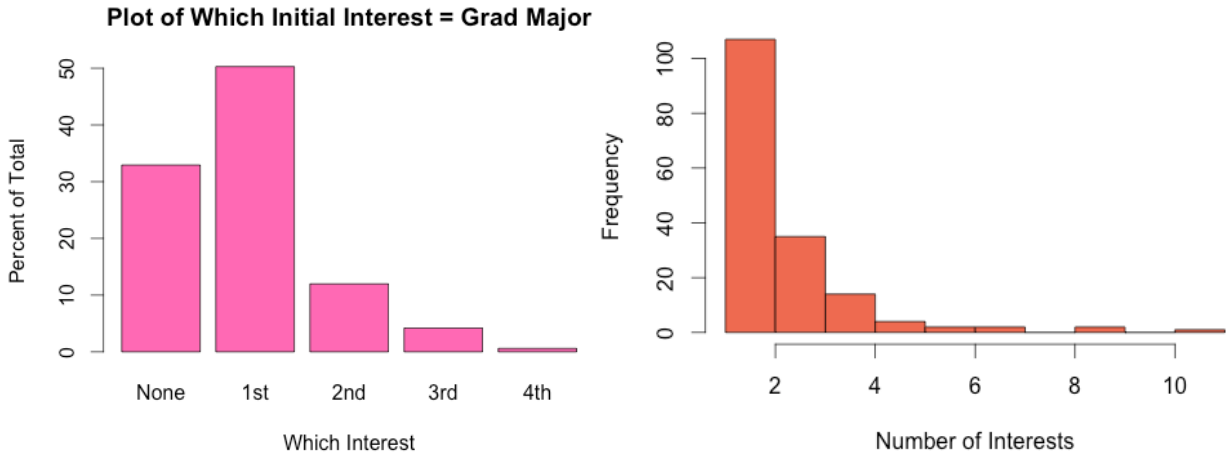
We can then look at the importance of each variable in our random forest in classifying major migration. As we can see, initial college, initial department, and time to declare are the three variables that contribute most to our classification. Once again, we notice that time took to declare is a significant factor in predicting the likelihood a student will switch departments. Thus, we can look at the partial dependence of our quantitative variable, “When.Declare”, to see how it effects the classification of students. As we can see, initially taking an extra semester to declare leads to a higher chance of being classified as a switch student, yet if a student declares after two or more semesters, the more likely he or she is to stick with that major. This could be because students who come in already declared have applied to a specific program like Information Systems or Architecture, and thus already have a vested interest in their primary department. The idea that being fast to declare could lead to more switches down the road might be a sign for advisors to work more closely with students to determine if it is the right time to declare. Declaring early can be beneficial as it gives students access to classes within the primary department, and declaring too late could leave students stuck in a major their heart is not set on.



It appears that with more data and more quantitative variables, random forests could be an excellent way of classifying student migration. They provide us with a good tool for predicting migration and indicate which variables have the strongest pull on our classification. While we do not have more quantitative variables for our entire data set, we do have some available for Dietrich College, which we were able to acquire through Dietrich College surveys.

## Dietrich College Survey Data

The last portion of my research focused in on Dietrich College, where students have the opportunity to fill out a survey of initial interest in various majors before they arrive as first-years. I was able to link this survey with the data I collected from the class of 2014 cohort, in order to gain some insight into whether factors such as number of initial major interests and which interest ultimately became the graduating major were predictive of major migration of students in Dietrich. While approximately half of all students who took the survey graduated in the first department they specified interest in, over 30% ended up in a different major from any of those they had initially specified. Students ranged from indicating 1 to 11 interests, with the majority expressing 1 or 2. As we can see, both our graphs are unimodal with right skew.




We can also look at our response variable in this subset of the data. As we can see, with the Dietrich College survey data, approximately 24% of students switched their primary departments, which is 8% higher than we saw in our full dataset. For comparison, we look at the

percent of switches in each of the five other undergraduate colleges. As we can see, CIT and SCS have very low switch rates compared to their overall population, though CIT has the largest number of switches overall.

| College | CFA | CIT | MCS | SCS | TSB |
|---|---|---|---|---|---|
| **Switches (%)** | 36 (16.7%) | 38 (11.1%) | 24 (15.0%) | 10 (10.8%) | 10 (13.5%) |

This higher rate for Dietrich College survey data could be explained by the fact that these switches are based on intended initial major, and it requires no work to initially declare intention. The increase could also be due to the fact that our Dietrich College survey data looks at major switches, which may be more common than department switches.

| No Switch | Partial Switch | Full Switch |
|---|---|---|
| 127 (76.05%) | 5 (2.99%) | 35 (20.96%) |

With this new data we can once again try to classify students switching departments by using gender, U.S. citizenship, participation in Greek life, initial department, length of time to initially declare, and our two new variables. When we run LDA on our training and testing sets for the survey data, we find that our errors are higher than our previous LDA classification, but still lower than the null hypothesis, which would be classifying all students in the "No Switch" category. Our training and testing error for LDA on our survey data is shown in the table below.

| Training Error | Testing Error |
|---|---|
| 23.9% | 27.3% |

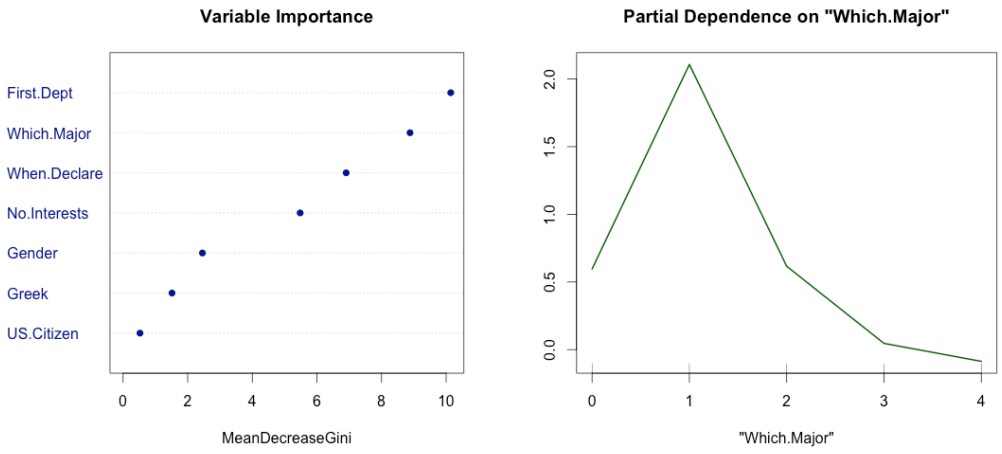
In this case our testing error is much higher than our training error, so we would assume that our data has high variance and would probably want to collect more data before considering other algorithms for classification. However, we can attempt to use random forests as an alternative method given their low error rates in our previous analysis.

Running a random forest classification function on our Dietrich College survey data, we begin by finding an out of bag (OOB) estimate of error rate of 17.91%. Our confusion matrix for misclassification for this random forest is shown below. As we can see, we never predict a switch when there is no switch, but we do still err when predicting no switches for students who

have switched. We can also use this random forest to predict major migration in our test set. We find an error rate of 15.15% on the test data, which is a very good error rate considering 23.95% of students surveyed actually switched their primary departments.

| Train Predicted Actual | No Switch | Switch | Test Predicted Actual | No Switch | Switch |
|---|---|---|---|---|---|
| No Switch | 97 | 5 | No Switch | 25 | 0 |
| Switch | 19 | 13 | Switch | 5 | 3 |

We are also able to create a variable importance plot for our survey data which shows us that, once again, first department and time to declare are important predictors for classification, as are our new variables which major and number of initial interests. Looking more closely at the "Which.Major" variable, we see that the majority of switches occur for students who end up graduating with the major that was their first choice. This is unexpected, as you would assume students would choose to declare in the major they initially expressed the most interest in. If they are not doing so, it is possible that they have forgotten their survey data or that Dietrich College academic advisors could be using this data more to their advantage when offering guidance to students on which departments to try out in their first semesters at Carnegie Mellon.




Thus, we have once again shown that random forests have the potential to be an excellent technique for classifying major migration at Carnegie Mellon, and they also have implications for prescriptive methods to be used by advisors in helping Dietrich College students follow their

initial interests from the start. With more data, and by sharing this data with advisors and students, we may be better equipped to handle the changing passions of students as they work through their undergraduate years.

## Discussion and Implications

The three main questions I hoped to answer through my research were:

- What is the risk of attrition associated with given departments/majors, and what are predictors associated with that risk?
- Where do students migrate to when they migrate?
- Do predictors such as Dietrich Surveys and Introductory Courses act as indicators of future major?

Through various statistical and analytical methods, I was able to determine the top predictors associated with classifying students as “at risk of migration”. Those top predictors were initial department, time to declare their initial major, and with Dietrich College survey data, which initial majors students had interest in. While these three variables are all valuable tools for classifying major migration, I also hypothesize that data such as cumulative GPA would be a very significant indicator in determining major or department attrition. Unfortunately, due to privacy concerns, this data was not available to me over the course of my research. However, I believe that adding this quantitative variable to my collected data could have a great impact on the ability to classify students via LDA, logistic regression, and random forests. I hope that I can share this information with advisors and administrators who have access to such data, and give them my analysis so that they can add variables like QPA to the training sets to see if those variables yield more impactful results. Another variable of importance is socio-economic status, and if that were added in as an additional predictor, it might also lower the classification error rates determined in my analysis.

## What's At Stake

The current methods of analysis employed by Carnegie Mellon include the Institutional Research Analysis Factbooks that are presented on a yearly basis. These documents include a large amount of tabular information on student degrees granted, the breakdown of demographics, finances, and other invaluable information to the university community. My analysis builds on this information by offering an easily digestible way for departments to determine how they are doing in terms of major retention. It also attempts to classify current students by the probability of switching their primary departments. This information could allow for departments to create action plans to increase information for students when they first enter the department, and lead to changes in advising strategies that could increase retention.

Obtaining more data is an essential part of providing a more robust analysis, and access to informative, clean data can open up a world of possibilities for analysis of major migration at Carnegie Mellon. Using a larger number of predictor variables, creating a visual analysis, and reporting those finding to advisors, particularly in Dietrich College, could help students find their passions sooner, as early as their first semester. The implications of this are invaluable, as it means students can take more courses that they are truly interested in, save money by graduating on time with all the necessary courses, and earn more out of their four to six-year undergraduate careers.

A key takeaway of my research is that at Carnegie Mellon, programs like BXA, QSSS, and SHS are not the only means of connecting colleges and the university. While our interdisciplinary programs and colleges provide ample opportunities for our *students* to connect with a wide variety of departments and interests, we could take the same "interdisciplinary approach" with our *data*. Data is an extremely important connection that drives many of our decisions and predictions on the quality of student education and life. Without the data we collect on students across all semesters, my work would not be possible. But I also believe there is room for improvement in the way we share data across our campus.

Through my research I found multiple connections between data I obtained from the registrar and data obtained from Dietrich College. These two disparate spreadsheets, when joined, yielded

results that had important implications for assessing the migration of students in Dietrich College. Without knowing the initial primary departments of students and their initial interests based on survey data, my classification would be much less significant than it turned out to be. Yet these two important datasets have not been looked at together before. In order to improve the way we guide our undergraduate careers, we must be able to look at all the data available to us. This means being able to connect data across multiple administrations in a way that is meaningful and can provide for the university as a whole. My hope is that this research can be the start of a discussion on creating a hub for data from all colleges and specific departments, as well as the university as a whole. That way, if an advisor in CIT wants to know why his new student switched departments from SDS, he can look at the initial survey data and understand that this student's primary passion was Environmental Policy. Without an interconnected hub, the advisor would never be able to see that survey data, and would lose some part of his understanding of his student.

My research is merely a catalyst for something much bigger. By adding in quantitative variables such as QPA and SES, and connecting data across multiple departments, the administration could use my prototype as a way to begin further analysis of major migration at Carnegie Mellon, and could make huge strides in helping students to find their place at the university as soon as possible, so that they can grow and contribute in a meaningful way throughout their undergraduate careers.

# *Works Cited*


Ambrose, Susan A., and Cristina H. Amon. "Systematic Design of a First-Year Mechanical Engineering Course at Carnegie Mellon University." The Research Journal for Engineering Education, 2 Jan. 2013. Web. 24 Sept. 2014.

Biggers, Maureen, Anne Brauer, and Tuba Yilmaz. "Student Perceptions of Computer Science: A Retention Study Comparing Graduating Seniors with Cs Leavers." *ACM SIGCSE Bulletin - SIGCSE 08* 40.1 (2008): 402-06. *Student Perceptions of Computer Science*. ACM, Mar. 2008. Web. 28 Sept. 2014.

Chase, Clinton I., and John M. Keene. "Major Declaration and Academic Motivation." *Eric.ed.gov*. Journal of College Student Personnel, Nov. 1981. Web. 24 Sept. 2014.

Cohoon, J. McGrath. "Toward Improving Female Retention in the Computer Science Major." *Communications of the ACM* 44.5 (2001): 108-114. *Toward Improving Female Retention in the Computer Science Major*. ACM, May 2001. Web. 28 Sept. 2014.

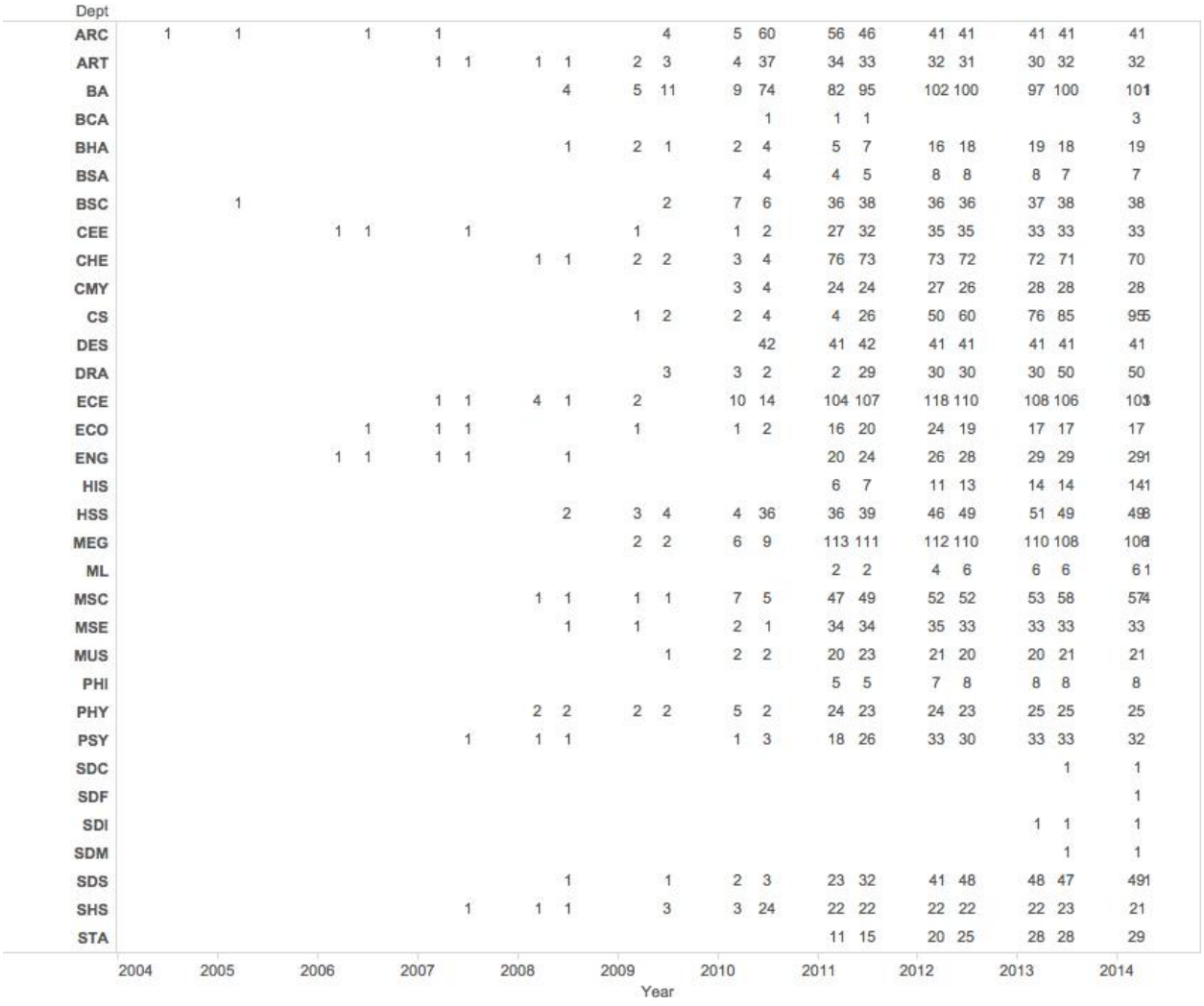
Fisher, Allan, and Jane Margolis. "Unlocking the Clubhouse: The Carnegie Mellon Experience." *ACM SIGCSE Bulletin - Women and Computing* 34.2 (2002): 79-83. *Unlocking the Clubhouse*. ACM, June 2002. Web. 24 Sept. 2014.

Hogan, Marjorie J., Jennifer M. Eastabrook, Amber Oke, and Laura M. Wood. "Emotional Intelligence and Student Retention: Predicting the Successful Transition from High School to University." *Personality and Individual Differences* 41.7 (2006): 1329-336. Elsevier, Nov. 2006. Web. 28 Sept. 2014.

Jamelske, Eric. "Measuring the Impact of a University First-year Experience Program on Student GPA and Retention - Springer." *Higher Education* 57.3 (2009): 373-91. *Measuring the Impact of a University First-year Experience Program on Student GPA and Retention - Springer*. 01 Mar. 2009. Web. 28 Sept. 2014.

# Appendix

## Figure 1: Department over Time Matrix



| Dept | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ARC | 1 1 | 1 | 1 | 1 | | 4 | 5 60 | 56 46 | 41 41 | 41 41 | 41 |
| ART | | | | 1 1 | 1 1 | 2 3 | 4 37 | 34 33 | 32 31 | 30 32 | 32 |
| BA | | | | | 4 | 5 11 | 9 74 | 82 95 | 102 100 | 97 100 | 101 |
| BCA | | | | | | | 1 | 1 1 | | | 3 |
| BHA | | | | | 1 | 2 1 | 2 4 | 5 7 | 16 18 | 19 18 | 19 |
| BSA | | | | | | | 4 | 4 5 | 8 8 | 8 7 | 7 |
| BSC | | 1 | | | | 2 | 7 6 | 36 38 | 36 36 | 37 38 | 38 |
| CEE | | | 1 1 | 1 | | 1 | 1 2 | 27 32 | 35 35 | 33 33 | 33 |
| CHE | | | | | 1 1 | 2 2 | 3 4 | 76 73 | 73 72 | 72 71 | 70 |
| CMY | | | | | | | 3 4 | 24 24 | 27 26 | 28 28 | 28 |
| CS | | | | | | 1 2 | 2 4 | 4 26 | 50 60 | 76 85 | 95 |
| DES | | | | | | | 42 | 41 42 | 41 41 | 41 41 | 41 |
| DRA | | | | | | 3 | 3 2 | 2 29 | 30 30 | 30 50 | 50 |
| ECE | | | | 1 1 | 4 1 | 2 | 10 14 | 104 107 | 118 110 | 108 106 | 103 |
| ECO | | | 1 | 1 1 | | 1 | 1 2 | 16 20 | 24 19 | 17 17 | 17 |
| ENG | | | 1 1 | 1 1 | 1 | | | 20 24 | 26 28 | 29 29 | 29 |
| HIS | | | | | | | | 6 7 | 11 13 | 14 14 | 14 |
| HSS | | | | | 2 | 3 4 | 4 36 | 36 39 | 46 49 | 51 49 | 49 |
| MEG | | | | | | 2 2 | 6 9 | 113 111 | 112 110 | 110 108 | 106 |
| ML | | | | | | | | 2 2 | 4 6 | 6 6 | 6 |
| MSC | | | | | 1 1 | 1 1 | 7 5 | 47 49 | 52 52 | 53 58 | 57 |
| MSE | | | | | 1 | 1 | 2 1 | 34 34 | 35 33 | 33 33 | 33 |
| MUS | | | | | | 1 | 2 2 | 20 23 | 21 20 | 20 21 | 21 |
| PHI | | | | | | | | 5 5 | 7 8 | 8 8 | 8 |
| PHY | | | | | 2 2 | 2 2 | 5 2 | 24 23 | 24 23 | 25 25 | 25 |
| PSY | | | | 1 | 1 1 | | 1 3 | 18 26 | 33 30 | 33 33 | 32 |
| SDC | | | | | | | | | | 1 | 1 |
| SDF | | | | | | | | | | | 1 |
| SDI | | | | | | | | | | 1 1 | 1 |
| SDM | | | | | | | | | | 1 | 1 |
| SDS | | | | | 1 | 1 | 2 3 | 23 32 | 41 48 | 48 47 | 49 |
| SHS | | | | 1 | 1 1 | 3 | 3 24 | 22 22 | 22 22 | 22 23 | 21 |
| STA | | | | | | | | 11 15 | 20 25 | 28 28 | 29 |

Year

## Switch Matrix 1

| First.Dept | ARC | ART | BA | BCA | BHA | BSA | BSC | CEE | CHE | CMY | CS | DES | DRA | ECE | ECO | ENG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARC | 42 | | 1 | 1 | 3 | 1 | | 5 | | | | | | | 1 | 1 |
| ART | | 30 | | | 2 | 1 | | | | | | | | | | 1 |
| BA | | | 81 | | | | | | | | | | | | 1 | |
| BCA | | | | 1 | | 1 | | | | | | | | | | |
| BHA | | | | | 10 | | | | | | | | | | | |
| BSA | | | | | | 2 | | | | | | | | | | |
| BSC | | | | | | | 32 | | | 2 | 1 | | | | | |
| CEE | | | | | | | | 24 | | | | | | | | |
| CHE | | | | | | | 1 | | 70 | 1 | 2 | | | | | |
| CMY | | | | | | | 1 | | | 22 | 1 | | | | | |
| CS | | | | 1 | | | 1 | 1 | | | 88 | | | | 1 | 1 |
| DES | | | | | | | | | | | | 41 | | | | 1 |
| DRA | | | | | 1 | | | | | | | | 47 | | | |
| ECE | | | | | | | | | | | 13 | | | 85 | 1 | |
| ECO | | | 12 | | | | 1 | | | | 1 | | | | 12 | 1 |
| ENG | | | | | 1 | | | | | | | | | | 1 | 21 |
| HIS | | | | | | | | | | | | | | | | |
| HSS | | | 1 | | | | | | | | 1 | | | | | |
| MEG | | | | | | | | | | | 3 | | | 1 | | |
| ML | | | | | | | | | | | | | | | | |
| MSC | | | | | | | | | | | 7 | | | | 1 | |
| MSE | | | | | | | | | | 1 | | | | 1 | | |
| MUS | | | | | 1 | 2 | | | | | | | | 1 | | |
| PHI | | | 1 | | | | | | | | | | | | | |
| PHY | | | | | | | | | | | | | | 1 | | |
| PSY | | | | | | | 1 | | | | | | | | | |
| SDF | | | | | | | | | | | | | | | | |
| SDS | | | 2 | | | | | 1 | | | | | | | | 1 |
| SHS | | | | | | | | | 1 | | 3 | | | | | |
| STA | | | | | | | | | | | | | | | | |

## Switch Matrix 1 (Continued)

| First.Dept | HIS | HSS | MEG | ML | MSC | MSE | MUS | PHI | PHY | PSY | SDF | SDI | SDM | SDS | SHS | STA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ARC | | 1 | | 1 | | | | | | | | | | 3 | | 1 |
| ART | | | | | | | | 1 | | 1 | | | | 1 | | |
| BA | | 1 | | | 3 | 1 | | | | | | 1 | | 1 | | 2 |
| BCA | | | | | | | | | | | | | | | | |
| BHA | | | | 1 | | | 1 | | | 2 | | | | | | |
| BSA | | | | | 2 | | | | | | | | | | | |
| BSC | 1 | | 1 | | | | | | | | | | | 2 | | 1 |
| CEE | | | | | | | | | | 1 | | | | 1 | | |
| CHE | | | | | | | | | | | | | | 1 | | |
| CMY | | | | | 1 | | | | | 1 | | | | | | |
| CS | | 1 | | | | | | | 1 | | | | | 2 | | 1 |
| DES | | | | | | | | | | | | | | | | |
| DRA | | | | 1 | | | | | | | | | | | | |
| ECE | | 1 | | | 1 | | 1 | | | | | | | 1 | | |
| ECO | 2 | 1 | | | 1 | | | 1 | | | | | | 4 | | 7 |
| ENG | | | | | | | | | | | | | | | | |
| HIS | 9 | | | | | | | | | | | | | | | |
| HSS | | 37 | | | 1 | | | | | | | | | 2 | | |
| MEG | | 1 | 104 | | | 1 | | | 1 | 1 | | | | 2 | | |
| ML | | | | 1 | | | | | | | | | | | | |
| MSC | | | | | 38 | | | | 1 | | | | | | | 1 |
| MSE | | 1 | | | 1 | 30 | | | | | | | | | | 1 |
| MUS | | 1 | | | | | 19 | | | 1 | | | | | | |
| PHI | | | | | | | | 4 | | 1 | | | | | | |
| PHY | | | | | 2 | | | | 19 | | | | | | | |
| PSY | | | | | | | | | | 26 | | | | 1 | | |
| SDF | | | | | | | | | | | 1 | | | | | |
| SDS | 3 | 1 | | 1 | | | | 1 | | | | | 1 | 24 | | |
| SHS | | | | | | | | | | | | | | 1 | 19 | |
| STA | | | | | | | | | | | | | | | | 12 |

## R Code

```
data <- as.data.frame(read.csv("Final_Thesis_Data.csv",header=T))
head(data)
nrow(data)
ncol(data)

# Only 21 Athletes for some reason so not included

attach(data)
summary(data)

# create has additional major variable
has.add <- ifelse(data$Add.Dept == "None",0,1)
data <- cbind(data,has.add)

# First, LDA
library(MASS)
# Pull out random test set
rand <- sample(1:nrow(data))
n_test <- round(nrow(data)/5)
n <- nrow(data)

head(data)
with(data,hist(When.Declare,xlab="No. of Semesters to Declare",
            main="Histogram of Initial Declaration",col="darkblue"))
with(data,barplot(table(has.add),names=c("No","Yes"),col="darkgreen",
              main="Additional Major?"))
with(data,table(has.add)/nrow(data))
names(data)
test <- data[rand[1:n_test],]
train <- data[rand[(n_test+1):length(rand)],]
head(train)

x.train <- train[,c(3,4,5,10,11,12,14)]
head(x.train)
x.train$Gender <- ifelse(x.train$Gender == "M",0,1)
x.train$US.Citizen <- ifelse(x.train$US.Citizen == "Yes",1,0)
x.train$Greek <- ifelse(x.train$Greek == "Yes",1,0)
x.train$First.College <- as.numeric(x.train$First.College)
x.train$First.Dept <- as.numeric(x.train$First.Dept)
x.train <- as.matrix(x.train)
y.train <- train$Switch
train.lda <- lda(x.train,y.train)
```

```
# Plot Training Data
dim(x.train)
dim(train.lda$scaling)
z <- x.train %*% train.lda$scaling
plot(z,pch=(y.train+2),col=(y.train+2),xlab="LDA Dimension 1",ylab="LDA Dimension 2",
    main="Classfication of Switches (Train)")
legend("topright",c("No Switch","Partial","Full"),pch=c(2,3,4),col=c(2,3,4))

# Calculate Training Error: 15.4%
train.mistakes <- sum(ifelse(predict(train.lda)$class == y.train,0,1))
train.misclass <- train.mistakes/nrow(x.train)
train.misclass

# 15.4% Switch
(length(which(y.train==2))+length(which(y.train==1)))/length(y.train)

# Create Testing Data
x.test <- (test[,c(3,4,5,10,11,12,14)])
x.test$Gender <- ifelse(x.test$Gender == "M",0,1)
x.test$US.Citizen <- ifelse(x.test$US.Citizen == "Yes",1,0)
x.test$Greek <- ifelse(x.test$Greek == "Yes",1,0)
x.test$First.College <- as.numeric(x.test$First.College)
x.test$First.Dept <- as.numeric(x.test$First.Dept)
x.test <- as.matrix(x.test)
y.test <- test$Switch

# Predict Classes for Test Data
test.lda <- predict(train.lda,newdata=x.test)

z2 <- x.test %*% train.lda$scaling
plot(z2,col=(y.test+5),pch=(y.test+2),xlab="LDA Dimension 1",ylab="LDA Dimension 2",
    main="Classfication of Switches (Test)")
legend("topright",c("No Switch","Partial","Full"),pch=(y.test+2),col=c(5,6,7))

plot(z,col=(y.train+2),pch=5,xlab="LDA Dimension 1",ylab="LDA Dimension 2",
    main="Classfication of Switches (Error)")
points(z2,col=(y.test+5),pch=4)
points(z2[which(ifelse(test.lda$class == y.test,0,1) == 1),],col="black",lwd=2)
legend("topleft",c("Train","Test","Missed"),pch=c(5,4,1))

# Testing Misclass Rate: 19.82%
mistakes <- sum(ifelse(test.lda$class == y.test,0,1))
misclass <- mistakes/nrow(x.test)
misclass
```

```
# QDA Misclass: 15.4%
head(x.train)
qda.train <- qda(y.train~x.train[,4]+x.train[,5]+x.train[,6]+x.train[,7])
predict(qda.train)$class
qda.mistakes <- sum(ifelse(predict(qda.train)$class == y.train,0,1))
qda.misclass <- qda.mistakes/nrow(x.train)
qda.misclass

test.qda$class
test.qda <- predict(qda.train,newx = data.frame(x.test[,c(4,5,6,7)]))
qda.test.mistakes <- sum(ifelse(test.qda$class == y.test,0,1))
misclass <- mistakes/nrow(x.test)
misclass

qda(switched~s.train$Gender+s.train$US.Citizen+s.train$Greek+train$First.Dept
   +train$Grad.College+train$When.Declare)

# Logistic Regression
head(train)
Add.Yes <- ifelse(train$Add.College == "None",0,1)
switched <- ifelse(train$Switch == 0,0,1)
log.train <- with(train,cbind(Gender,US.Citizen,Greek,First.Dept
                  ,First.College,When.Declare,Add.Yes))
log.response <- switched

Add.Yes.Test <- ifelse(test$Add.College == "None",0,1)
switched.test <- ifelse(test$Switch == 0,0,1)
log.test <- with(test,cbind(Gender,US.Citizen,Greek,First.Dept
                  ,First.College,When.Declare,Add.Yes=Add.Yes.Test))
log.test.response <- switched.test

# Misclass Logistic Regression: 25.5%
library(glmnet)
log.fxn <- glm(log.response~log.train,family="binomial")
summary(log.fxn)
log.pred <- predict(log.fxn,type="response")
pred.switches <- rbinom(nrow(train),1,log.pred)
misclass.log.table <- xtabs(~pred.switches+log.response)
(sum(misclass.log.table) - sum(diag(misclass.log.table)))/sum(misclass.log.table)

# Logistic Regression with Transformed When.Declare Variable
dec.time <- with(train,1/(1+When.Declare^2))
log.train <- with(train,cbind(Gender,US.Citizen,Greek,First.Dept
                  ,First.College,dec.time,Add.Yes))
log.response <- switched
```

```
Add.Yes.Test <- ifelse(test$Add.College == "None",0,1)
switched.test <- ifelse(test$Switch == 0,0,1)
dec.time.test <- with(test,1/(1+When.Declare^2))
log.test <- with(test,cbind(Gender,US.Citizen,Greek,First.Dept
                        ,First.College,dec.time.test,Add.Yes=Add.Yes.Test))
log.test.response <- switched.test

# Misclass Logistic Regression: 25.5%
library(glmnet)
log.fxn <- glm(log.response~log.train,family="binomial")
summary(log.fxn)
cv_log <- cv.glmnet(as.matrix(log.train),
                switched,type.measure='class',family='binomial')
cv_log$cvm[which(cv_log$lambda==cv_log$lambda.1se)]

log.pred <- predict(cv_log,newx=as.matrix(log.test),type="class")
pred.switches <- rbinom(nrow(train),1,log.pred)

(nrow(train)-sum(diag(misclass.table)))/nrow(train)

table(data$Switch)
185/(185+951)

# Random Forest OOB 9.02%

library(randomForest)
my_forest <- randomForest(x=log.train,y=as.factor(log.response))
predict.forest <- predict(my_forest,newdata=log.test,type='class')
forest.test.matrix <- xtabs(~log.test.response+predict.forest)
(sum(forest.test.matrix)-sum(diag(forest.test.matrix)))/sum(forest.test.matrix)

# Test Error Random Forest 0.1057 10.57% Error
?varImpPlot
varImpPlot(my_forest,col="darkblue",pch=16,main="Variable Importance")
partialPlot(my_forest,pred.data=log.train,x.var="When.Declare",
          lwd=2,col="darkgreen")



# Survey Data

survey <- as.data.frame(read.csv("SurveyData_Final.csv",header=T))
head(survey)
survey <- survey[,-3]
```

```
barplot(table(survey$Which.Major)/nrow(survey)*100,names=c("None","1st","2nd","3rd","4th")
,col="hotpink",xlab="Which Interest"
      ,ylab="Percent of Total",main="Plot of Which Initial Interest = Grad Major")
nrow(survey)
head(survey)
with(survey,hist(No.Interests,col="coral2",xlab="Number of Interests",main=""))
rand <- sample(1:nrow(survey))
n_test <- round(length(rand)/5)
survey.noid <- survey[,-1]
head(survey.noid)

s.mat <- survey.noid
s.mat$Gender <- ifelse(s.mat$Gender == "M",0,1)
s.mat$US.Citizen <- ifelse(s.mat$US.Citizen == "Yes",1,0)
s.mat$Greek <- ifelse(s.mat$Greek == "Yes",1,0)
s.mat$First.Dept <- as.numeric(s.mat$First.Dept)
s.mat$Grad.Dept <- as.numeric(s.mat$Grad.Dept)
s.mat$Grad.Major <- as.numeric(s.mat$Grad.Major)
s.mat$Switch <- ifelse(s.mat$Switch=="No Switch",0,1)

s.test <- s.mat[rand[1:n_test],]
s.train <- s.mat[rand[(n_test+1):length(rand)],]
head(s.train)
sx.train <- s.train[,c(1,2,3,4,7,9,10)]
sx.train <- as.matrix(sx.train)
head(sx.train)
sy.train <- s.train[,8]
s.train.lda <- lda(sx.train,sy.train)


dim(sx.train)
dim(s.train.lda$scaling)
s.z <- sx.train %*% s.train.lda$scaling
plot(s.z,col=(sy.train+2))

s.train.mistakes <- sum(ifelse(predict(s.train.lda)$class ==sy.train,0,1))
s.train.misclass <- s.train.mistakes/nrow(sx.train)
s.train.misclass

sx.test <- s.test[,c(1,2,3,4,7,9,10)]
sx.test <- as.matrix(sx.test)
sy.test <- s.test[,8]

s.test.lda <- predict(s.train.lda,newdata=sx.test)
s.mistakes <- sum(ifelse(s.test.lda$class == sy.test,0,1))
s.misclass <- s.mistakes/nrow(sx.test)
```

```
s.misclass

dim(x.test)
z2 <- x.test %*% train.lda$scaling
plot(z2,col=(sy.train+3))

table(survey$Switch)/nrow(survey)
40/167

# Random Forest OOB 9.02%

library(randomForest)
my_forest_survey <- randomForest(x=sx.train,y=as.factor(sy.train))
predict.s.forest <- predict(my_forest_survey,newdata=sx.test,type='class')
forest.s.matrix <- xtabs(~sy.test+predict.s.forest)
(sum(forest.s.matrix)-sum(diag(forest.s.matrix)))/sum(forest.s.matrix)

# Test Error Random Forest 0.1057 10.57% Error
par(mfrow=c(1,2))
varImpPlot(my_forest_survey,col="darkblue",pch=16,main="Variable Importance")
partialPlot(my_forest_survey,pred.data=sx.train,x.var="Which.Major",
        lwd=2,col="darkgreen")



# Logistic Survey

head(survey.noid)
survey.switch <- ifelse(survey.noid$Switch == "No Switch",0,1)
survey.log.fxn <- glm(survey.switch~US.Citizen+Greek+Gender+Which.Major+First.Dept
          +No.Interests+When.Declare,data=survey.noid,family="binomial")

# Misclass Logistic Regression: 25.15%
summary(survey.log.fxn)
hist(predict.glm(survey.log.fxn,type="response")) #probabilities
s.log.pred <- predict(survey.log.fxn,type="response")
s.pred.switches <- rbinom(nrow(survey.noid),1,s.log.pred)
s.misclass.table <- xtabs(~s.pred.switches+survey.switch)
(nrow(survey.noid)-sum(diag(s.misclass.table)))/nrow(survey.noid)
```