March 2017

Enhancing Career Opportunities: Using Supervised Learning to Analyze Career Outcomes Data

Emily Catherine Weber
Worcester Polytechnic Institute

Sadie Grace Gauthier
Worcester Polytechnic Institute

Zachary Michael Peters
Worcester Polytechnic Institute

Enhancing Career Opportunities: Using Supervised Learning to Analyze Career Outcomes Data

A Major Qualifying Project

by

Sadie G. Gauthier
Robert C. Vigeant
Zachary M. Peters
Emily C. Weber

Submitted in partial fulfillment
of the requirements for the degree of:

Bachelor of Science

Mathematical Sciences
Worcester Polytechnic Institute
March 17, 2017

Approved by:

Professor Randy C. Paffenroth

Abstract

The Career Development Center (CDC) at Worcester Polytechnic Institute (WPI) is a key organization in helping students develop the skills necessary for finding a career path after graduation. Our team analyzed graduation, CDC usage, and internship data in order to better understand factors which predict student outcomes after graduation. Through machine learning, we pinpointed the most relevant predictors for post-graduate success. The CDC can use this information to better use their resources and provide insight for future analysis.

Acknowledgements

We would like to thank our sponsor, Dave Ortendahl, Director of Coroprate Relations at the Worcester Polytechnic Institute Career Development Center (CDC). Dave has been extremely helpful during this project, helping us define our scope, providing us with continuous feedback during key junctures in the project, and offering insights into nuanced aspects of the CDC.

We would also like to thank Maggie Becker, Director of the CDC, who worked with us to clean the data file and particularly to redefine the subcategories of the finely-defined workshops the CDC offers.

We would also like to thank Allyson Bernard, Senior Recruiting Coordinator at the CDC, for her work in anonymizing and initially preparing the data to be handed over to the group. Allyson was receptive to several of our early questions at the start of the project as well. In addition, we would like to thank all other staff at the CDC who helped collect, prepare, and anonymize the data, without which this project would not be possible.

Matthew Weiss offered his experience with Python, scikit-learn, machine learning, and the general math and programming concepts underlying the project throughout its course. He was a constant source of information both conceptually and practically. We thank him for his assistance throughout the project.

Finally, we would like to thank Professor Randy Paffenroth, our project advisor, who motivated us, kept us on task, and was monumental in helping us overcome several hurdles from beginning to end. We cannot thank him enough for his passionate, energetic, and supportive role in leading our team to a successful project outcome.

Contents

List of Figures

List of Tables

Executive Summary

One of the most important outcomes of a college career is a student's post-graduate job prospects, graduate school attendance, or other markers of educational success. Like most universities, Worcester Polytechnic Institute (WPI) has a Career Development Center (CDC), which offers students a wide range of resources, programs, and networking opportunities to enhance their prospective job opportunities and other post-graduate ambitions. But while the CDC annually reports on alumni status, there has not yet been a dedicated effort to analyze career outcomes data in-depth using mathematically rigorous methods. Although the CDC can summarize student outcomes for a given year, deeper questions about what aspects of college life predict those outcomes demand more inspection.

The goal of this project was to apply supervised machine learning methods to analyze the CDC's class 2015 student outcomes data. This data was merged with two additional datasets which tracked information on student internships and CDC usage for the class of 2015. These datasets are referred to as Full Class of 2015, Internship, and CDC Usage, respectively. After combining these datasets, pre-processing the data, and devising an encoding scheme, we studied several classification methods. After some consideration, we decided that decision trees were the optimal approach for our dataset since they do not require separators and were more efficient than the alternative methods.

We researched and then modified an over-sampling algorithm called Synthetic Memory Over-Sampling Technique (SMOTE) and then implemented a more accurate Borderline-SMOTE variant to resolve the issue of imbalanced predictors within our datasets. We then applied k-fold cross-validation, split the data into training and testing sets, and ran our algorithms through several runs on the three different datasets: Full Class of 2015, Internship, and CDC Usage.

During these runs, we determined the top four predictors using SMOTE and Borderline-SMOTE for each of the three datasets, finding that the Borderline-SMOTE runs were either as accurate or more accurate than the regular SMOTE runs in all three cases. For the Full Class of 2015 data, we found that Walk-Ins were the most important predictor, followed by Degree Date (when a student received his or her degree), Career Fair, and Degree Type (what type of degree a student received).

Run	Most Important	2nd Most	3rd Most	4th Most
1	Walk-In	Degree Date	Career Fair	Science Degree
2	Walk-In	Degree Date	Career Fair	Workshop
3	Walk-In	Degree Date	Workshop	Career Fair
4	Walk-In	Degree Date	Career Fair	Resumazing
5	Walk-In	Degree Date	Career Fair	Degree Type
6	Walk-In	Degree Date	Career Fair	Degree Type
7	Walk-In	Degree Date	Degree Type	Career Fair
8	Walk-In	Degree Date	Career Fair	Resumazing
9	Walk-In	Degree Date	Workshop	Career Fair
10	Walk-In	Degree Date	Degree Type	Career Fair

Table 1: Top Four Predictors (in order of importance) for each of the Ten Borderline-SMOTE Runs for the Full Class of 2015 dataset.

As a note, the accuracy scores in Table 2 are examples of the scores we got from testing. These scores will differ from run to run.

Category Added	Accuracy Post-Category Addition
Walk-In	58.550 %
Degree Date	68.891 %
Career Fair	71.328 %
Degree Type	72.934 %

Table 2: Example Run Results of Category importance for Borderline-SMOTE found by Forward Subset Selection for the Full Class of 2015 dataset.

For the Internship data, we found a less clear-cut result. Degree Date was the most significant predictor, followed by Resumazing. However, there were only marginal increases in accuracy with added categories from the following predictors: Nationality, Summer Internship, and so on up to IMGD Degree. Here "marginal increases in accuracy" refers to increases in prediction accuracy of $< 5\%$ from adding a new predictor. Finally, for CDC Usage, we found Degree Date again was the most significant predictor, followed by Walk-In and then with marginal increases in accuracy from added categories from the following predictors: Degree Type, Business Engineering, Summer Internship, and so on up to Mathematics Degree.

Category Added	Accuracy Post-Category Addition
Degree Date	62.58 %
Resumazing	74.41 %
Nationality	76.69 %
Summer Internship	80.05 %
Degree Type	82.77 %
Science Degree	83.90 %
Workshop	84.90 %
Co-op	85.39 %
IMGD Degree	85.60 %

Table 3: Example Run Results of Category importance for Borderline-SMOTE found by Forward Subset Selection for the Internship dataset.

Category Added	Accuracy Post-Category Addition
Degree Date	58.41 %
Walk-In	72.60 %
Degree Type	74.22 %
Business Engineering	74.90 %
Summer Internship	75.23 %
Company Info Session	75.69 %
Civil Engineering	76.04 %
Mathematics Degree	76.05 %

Table 4: Example Run Results of Category importance for Borderline-SMOTE found by Forward Subset Selection for the CDC Usage dataset.

We concluded that the most relevant predictors across the datasets are Walk-In and Degree Date. It should be noted that by claiming, for instance, that the four predictors in the Full Class of 2015 dataset have a total accuracy of nearly 73% when they are all added to the model is *not* equivalent to saying that going to a certain number of Walk-Ins, graduating at a certain time, going to a certain amount of Career Fairs, and having a certain level of degree will guarantee a student an approximately 73% chance of any particular outcome. Rather, this means that when they are all applied, the top four predictors will be able to predict a generic student's outcomes with a 73% accuracy, whether they received a job, went to graduate school, volunteered, are still seeking employment, or did not get a job at all. Further analysis is required to make stronger assumptions about why those predic-

tors lead to those outcomes. The best way to get these results is to analyze the decision tree pathways and see which conditions led to which predictions of the response variables.

For future works, we suggest that the next steps are as follows. First, improving CDC survey-gathering techniques and in-house data cleanup to produce more easy to process datasets without the need for manual and programmatic modifications. Second, attempting a more in-depth run of the class of 2015 data using our methods applied to a dataset including a Walk-In subcategory breakdown in line with the CDC's recommendations due to its significance as a predictor. And finally, applying these methods to both a new (2016 or later) dataset for comparison and to determine trends. Older datasets (2014 or earlier) may be useful for analyzing such trends as well. Other potential additions might include using CDC usage data to determine ranges of optimally beneficial use of certain programs and to add other intersecting datasets beyond the ones used in this project to investigate correlations among other aspects of undergraduate life.

1 Introduction

College is one of the most costly investments one can make, so achieving a stable and well-paying job after graduation is necessary to justify the expenses. A student's activities and decisions during college greatly affects his or her career path and other post-graduate opportunities. In this project, we used cutting-edge data analysis to discover trends that lead to successful outcomes for college students and to guide career development programs and services to reflect this information.

The Career Development Center (CDC) at Worcester Polytechnic Institute (WPI) routinely collects self-reporting survey data from alumni. These forms include information on employment, graduate school attendance, and other markers of post-graduation progress. Additionally, the CDC tracks records of attendance for its programs, such as Career Fairs, Workshops, and employer-led information panels. They also ask students to report on internships and cooperative opportunities they have participated in while at attending WPI. We used all of this information to compile the three sources into one large dataset for analysis.

Combined, these three datasets offer a substantial framework for analyzing the success of alumni, relative to their career paths and additional schooling. Using the thousands of data points provided by the CDC, we were able to apply machine learning methods, including decision trees, to analyze what parts of our data were most indicative of certain outcomes. Through these analytic approaches, we were able to provide the Career Development Center with unique information that will help them program future events and help steer students in the right directions. They can use this analysis to determine which events are most beneficial to helping a student get a job and what other factors might contribute to better opportunities.

1.1 Background

This section provides the background of our sponsor, the CDC at WPI. It also reviews the origin and nature of the datasets used, as well as key concepts in machine learning which were used throughout the project. More specific information about the datasets, including how they were cleaned and modified for analysis, are discussed in later sections.

1.2 Measuring Outcomes

The CDC serves as the primary career-oriented resource for students at WPI, offering a large selection of programs and services to help students with networking, resume writing, interviewing, and finding careers that suit their interests. These events are presented in several ways depending on the audience and its needs. The CDC hosts workshops, employer-led discussion panels, mock interviews, and Career Fairs throughout the year. The CDC also collects information on the general usage of their services and attendance at their events, as well as self-reported data from full time students who participated in an internship, a co-op, or some form of summer research opportunities.

A clear marker of student success for the CDC is reflected in positive post-graduate standings in employment, continued education, volunteer work or military involvement. Determining the effectiveness of their programs motivates the CDC to compile an annual report of outcomes for alumni. It is from the 2015 report that we have obtained the core data for our project. According to the 2015 report, the university acquired data from 1420 of the 1671 graduates that year - a nearly 85% coverage rate for the student body [7]. The introduction further claims,

> Again this year, WPI can boast a very strong success rate: 91.7% of graduates for whom we have outcomes data were either employed, entering graduate school, serving in the military, or participating in volunteer service (such as City Year, Teach for America, etc.). [7]

At least for the previous year, this was true as well [8]. For the class of 2014, WPIs CDC confidently concluded that the previous years graduates had a post-graduate success rate of over 90% [8]. This was further broken down in the introduction of the 2014 report showing an increase in success rates from the bachelor level (89.4%) to masters (91.7%) and doctoral (96.0%) [8].

These reports and their high-level outcomes give us a significant reference point for outlining our project goals. The converse of the reports success story, for instance, is that 8.3% of students surveyed did not have outcomes which were successful by the CDCs measure. These graduates are either unemployed or are actively seeking employment without a secondary plan of action.

Knowing this leads us with our first major question to try and answer, using information from all of these available databases: What can the CDC

do to improve outcomes for these 8.3% of alumni? And the reverse of this question is worth keeping in mind while addressing this concern: What is the CDC doing right (or best) to produce optimal outcomes for the majority of postgraduate students? Once these results are identified, not only should we ask how they can further enhance the quality of outcomes for those already benefiting from the CDCs programs, but also how the most effective services be tailored and translated to the minority of students with negative outcomes.

2 Literature Review

Here we provide a brief overview of the existing literature on the subject of using data science and machine learning methods in the context of post-secondary career development. A review of existing literature did not reveal any major published works using the methods applied in this project for the purposes of enhancing career outcomes through university career development organizations. In fact, the use of data mining and machine learning data analysis techniques in higher education in general is a new frontier which deviates from traditional applications in business decision-making [9]. Pal describes the emerging field of Educational Data Mining, which deals with data mining to discover knowledge from data originating from educational environments [10], but this scope does not necessarily include career development pathways. Rather, it is more exclusively focused on answering questions about educational outcomes within the classroom.

2.1 Academic Outcomes with Machine Learning

Researchers have conducted several studies which focus on student academic outcomes. Delen applied a six-step approach and analyzed model performance using 10-fold cross-validation. Notably in close alignment with our own project, Delen remarks that approximately 80% of project time is devoted to the methods first three steps, which include determining the scope and goals of the project, identifying and accessing the necessary datasets, data cleanup, and data transformation [9]. The study was based on data from 16,066 students enrolled in an unnamed public university from 2004 to 2008 and relevant variables from the student datasets included degree, major, various measures of GPA, sex, SAT scores, age, and transfer hours among others [9].

The major outcomes of Delen's study are threefold. First, he confirmed that machine learning prediction models outperform statistical ones [9]. Sec-

ond, that there was unacceptably low accuracy of 50% for the imbalanced No class, a predictor which corresponded to students who dropped out after a semester [9]. A balanced dataset was unequivocally better at producing results with a higher accuracy than on which was imbalanced. Finally, the imbalanced data problem was addressed to produce higher accuracy predictions [9]. Delen separated the minority class samples from the set and randomly selected an equal number of samples from the majority class, repeated ten times [9]. Support vector matrices produced the best results in this study, but all of the machine learning methods employed had prediction rates around 80% both before and after the imbalanced data was addressed [9].

Addressing the differences between data analysis in education and in the private sector for business decision-making, Luan identifies critical questions a data scientist might ask in either case [2]. For instance, from a business perspective, one might want to get to the bottom of which customers are the most profitable or what makes them likely to use a competitors services [2]. But within an educational context, those questions might be concerned with which students exhibit the most academic prowess and what the university can do to get more students to enroll [2]. The common theme is that both of these domains ask questions mainly about the behavior of their major constituents (consumers and students) and about what the organization in question (a business or university) can do to appeal to them based on their behaviors and trends.

For our purposes, the major constituents are students and the organization in question is the universitys Career Development Center. The types of questions we will attempt to answer are closer to academic success, but contain some elements of the business world, as they are career-oriented and contingent upon the individual having a job, pursuing the foundations for employment, or serving in some other capacity. The information on salaries from our dataset was anonymized and encoded in the form of tiers, rather than specific dollar amounts, yet is still an integral aspect of measuring success. In this way, the question of outcomes appears more closely related to that of business-oriented analysis than academic.

Luan introduced a case study similar to the one found in Delen. In this instance, the SPSS software Clementine was used to make predictions about students who transfer into a four-year college from a community college [2]. The reader is invited to read the case study for a further look at the specific outcomes, but the basic process is one which is closely paralleled to our own. Those steps are determining the most accurate model (neural networks, for this case study), testing accuracy using confusion matrices,

adjusting parameters as needed to produce a maximal prediction accuracy rate, and analyzing results of the prediction to draw conclusions about the data [2]. Luan's case study yielded a prediction accuracy rate of 76.5% with neural networks and sorted features by importance, ranging from number of liberal arts classes taken (0.315) as the most important, to number of degree applicable courses taken (0.074) as the least important [2]. Features were measured as a range from 0 (lowest) to 1 (highest) in this study.

Combinations of the approaches and questions from these two studies have been applied elsewhere to produce similar outcomes in similar contexts. Beck and Woolf [11] developed a machine learning algorithm to measure student responses in terms of time taken to answer and whether or not their answer was correct. Kotsiantis *et al.* used machine learning approaches to predict student performance while participating in a distance learning system at Hellenic Open University, where the student and instructor are in physically different locations and communication is done remotely [12]. The Naive Bayes algorithm was determined to be the most accurate, and was used to help construct a prediction model to assist tutors to help them understand which of their students were most likely to complete their course in a timely fashion [12]. Naive Bayes is a Bayesian network method which assumes feature independence. We did not consider it for our study, so the curious reader is encouraged to seek further reading, such as the brief explanation offered in [12].

3 Project Outline

From the start, the goal of the project was defined broadly as follows: Use supervised machine learning methods to analyze the outcomes report and other related 2015 alumni datasets to suggest data-driven improvements to CDC programs and outreach. To accomplish this goal, we took three major steps. First, to practice with methods that would become useful later on using a test dataset. To do this we used a dataset provided by the University of California- Irvine to learn how the methods described above apply to an actual dataset. Second, we needed to apply what was learned in that process to the datasets themselves, which included an active process of outlining goals based on what could be derived from the data. And finally, to structure a detailed plan of action for the CDC based on an interpretation of the results. In many ways, this is our most crucial step as it will help them plan future events and work with future project teams.

3.1 Test Dataset

Before acquiring an anonymized and analysis-ready initial dataset, we first planned to work on an openly-available dataset on a Portuguese secondary school performance. During this time, we familiarized ourselves with the concepts of data science, programming in Python, and using other software (Atom, Bitbucket, etc.). The process applied to the test dataset would become a model for how we would approach the actual datasets for our project. Our general outline was to familiarize ourselves with the data and the given features; determine the scope of what we wanted to analyze; clean and transform the data, encoding it so it could be properly processed; run a variety of algorithms on the dataset; determine the most accurate model to use through cross-validation; perform feature selection to determine the most relevant predictors; and determine outcomes using the chosen method. These were all important lessons to learn before we got our actual dataset. This way we could quickly begin executing our own data clean up and analysis, armed with the knowledge of challenges we might face and ways to possibly fix them.

3.2 Data Pre-Processing

With the data in our hands, it was evident before starting the process that cleanup, manipulation, encoding and transformation would be required. In addition, since the datasets of interest were not combined, we had to manually merge them by intersecting them based on a common key between them both: a student identifier, which is a random alphanumeric code generated for each student in both databases and is consistent across the databases so that the same alphanumeric code is used for a single student. Using Excel and basic filtering and manipulation, we combined the two datasets into one. Primarily taking advantage of the same built-in functionality in both Excel and Google Sheets, along with Python, we cleaned the data of duplicates and erroneous information, transformed it to meet our processing needs, and encoded non-categorical data using a one hot encoding scheme. We had to do this in order to make sure that our programs would not be thrown off by duplicates or hard to read data. This cut down on the number of issues we might face in the future with regards to things we could change and fix.

3.3 Selecting and Executing Methods

Considering the nature of the data in question while also applying standard tests to measure accuracy of prediction models, we settled on a single model

which resulted in the highest prediction accuracy and made the most sense given the structure of our data and its classes. This model was one of the four we tested on the initial dataset. It is called decision trees and allows us to graphically show the CDC what helps a student when they graduate. We also had to keep in mind that our data is imbalanced. This means it is heavily loaded in one direction and this was a major obstacle we had to overcome. We had to carefully choose our methods in order to make sure we were giving the CDC proper and not skewed results. To accomplish this, we applied k-fold cross-validation using $k = 10$ folds.

3.4 Crafting Recommendations CDC Programs

Based on our outcomes, we would finally analyze the full spread of data and devise recommendations for the CDC for how to use the results we found to improve their own operations. Using our conclusions, the CDC can help students achieve better career outcomes and attempt to lower the percentage of graduates who fall into the Seeking/Serving status category. Also, based on our methods we would be able to provide code and other materials for future students who work with the CDC on projects of this nature. We hope to provide the CDC with a means to help themselves for years to come.

3.5 Crafting Data Recommendations for the CDC

We were very lucky to be provided with so much data from the CDC so that we would have enough to draw some conclusions. While the large amounts of data were greatly appreciated, there are still a few things that could have been done to make the data processing easier. This really has nothing to do with the data being imbalanced, this is real life and that happens, but the data received was at times both all over the place and inconsistent. Besides giving the CDC recommendations for their future programs we hope to provide them with some guidance in how to collect the data so that this process goes smoother in the future. Things like keeping surveys consistent that data can be compared from year to year and not repeating pieces of data within a spreadsheet will greatly speed up future work. It is our hope with these recommendations that the CDC will be able help future students who work on projects like this by making the data transition smoother and more efficient.

3.5.1 Data Issues

A significant hurdle we had to overcome during this project was manipulating and cleaning the data from its original form. This was distinct from the problems of set intersection, in which we had to create parity between the three major datasets, and encoding, in which we had to convert non-numeric data into binary and other numeric values. The major issues of this component of the project included general errors in the data, inconsistent recording across datasets, ungrouped data categories, and missing information.

General errors in the data were to be expected from the start. Essentially any situation in which an individual enters information which is not pre-defined will be subject to misspellings, multiple different ways to reference one thing, or other similar inconsistencies which have to manually be scrubbed. For instance, one of the first things we had to address was the problem of inconsistent city data. Before we even began deciding on column predictors, we had to consider the fact that there were multiple ways to reference one city because some had slight misspellings or capitalization errors. As an example, "Beijing" could have been recorded in the primary dataset as simply "Beijing," "Beijing" followed by an equivalent ZIP Code identifier, or "Beijing" preceded by a similar identifier. We ultimately decided against using cities in our process because in the highest level of categorizing we could group them into (countries), a skewed majority of them were in the US. Because we had another means of identifying if the student was a US citizen or an international student, we eschewed this category.

Still, similar issues had to be addressed in other categories. For instance, the CDC Usage dataset's kiosk categories had a multitude of different ways of recording the same information. As an example, the kiosk name "Interview Workshop" was recorded in this form. But it was also recorded with dates before it "2_25_15 Interview Workshop," and combined with a separate category from itself "10_28_14 Tour Guide Resume/Interview Workshop." There was also the category "Interviewing Workshop" which also had a date precede it. We discuss the issue of categorization in terms of grouping similar categories into larger but still distinct groups in a later section. This particular issue was in reconciling categories which were clearly the same but which were recorded differently due to differences in event naming and recording across various years for certain events. We recommend that the CDC standardize its recording of kiosk programs by distinguishing date identifiers and event name. The date is already present in a separate category designated "Kiosk Time Recorded," timestamped in mm/dd/yyy hh:mm format, so having it included in any form in the kiosk name category

is redundant. Combining the two also makes it extremely difficult to deal with the data beyond the highest level of consideration by further making entries unnecessarily unique. We also suggest that the CDC begin including higher-level categorization of its programs into its dataset for future entries. Attached in the appendix is a version of the dataset which was manipulated and re-categorized by adding a new column for certain higher-level categories.

There were some elements of information in these datasets which we had to incorporate manually from other sources, or which limited our ability to deal explicitly with certain data types. The 'special request' variant of kiosks–in which a program is offered on-request for a class, fraternity, or other organization/individual on campus–was not well-understood until late into the project. Since we did not have adequate time to venture into answering questions relating to special request program usage and its relation to future program usage, we did not pursue re-categorizing our data into special request versus regular programs. However, if the CDC continues to have an interest in answering this question, it would be useful to create a breakdown category inherent to the dataset which designates whether the course was special request or a regular group program. This would make data processing simpler instead of having special request tied into the kiosk name similar to the date as mentioned above. As an example, instead of "2_11_SR: AEI - Job Finder," the CDC would use the existing date category; a separate column having "SR," "Yes," "1" or some other indicator that the event was special request, and "Job Finder" as the kiosk name. The CDC could also consider designating where or for whom these events were held, since "AEI" in this case is such a modifier on "Job Finder" as a program.

4 Machine Learning Techniques

For the sake of reference and providing background knowledge, we offer a brief overview of the basic methods in supervised machine learning which were considered and either used or rejected for the purposes of this project. These include cross-validation methods, classification methods, and decision trees. We discuss the specific rationale behind decisions to use or reject given methods in Section 5.

4.1 Training and Testing Data using Cross-Validation

For machine learning to work on a dataset, the data must be broken up into a training set and a testing set. A training set is a set of data that is used

to teach the statistical method how to find the target value [1]. A test set is a set of data in which the model is not given what the target variable is, this set allows the user to see how well their model predicts other data. Training and Testing datasets are used so that the model is more robust to other datasets that are used with it.

For our datasets, we used k-fold cross-validation to make our training and testing sets. Cross-validation is used to ensure that the algorithm is giving the best result possible without overfitting or data snooping. In k-fold cross-validation, the algorithm randomly divides the data in k folds, or groups [1]. The first group is used as the test dataset and the remaining $k-1$ groups are used as the training set [1]. This process is done for k times each with a different fold as the test set [1]. At the end of the process the average of all the models is taken and that gives the accuracy. For our k-fold cross-validation we used $k = 10$. We chose this number because it is a common practice in data science to use $k = 10$ for this method.

Figure 1: Example of K-Fold Cross-Validation with $k = 5$. The blue boxes are testing data and the white are training data.

The above figure shows an example of k-fold cross-validation. The blue boxes represent how k-fold picks out the testing data during each fold. Each box represents a percentage of the data that is put in to be the testing data. After that fold those points may not be used for testing data again.

4.2 Linear Discriminant Analysis

Also known as LDA, linear discriminant analysis is a classification method used as an alternative to linear regression. Some insights into this method are given in [13], describing it as a way to project a dataset onto a lower-dimensional space with good class-separability in order to avoid overfitting," as well to speed up computation [13]. For this reason, it is much like principal component analysis (PCA), which also is used to map from a higher-dimensional space to a lower-dimensional one.

James, *et al.* describe how LDA works in *An Introduction to Statistical Learning* [1]. First, we define conditional distributions according to Hildebrand, as "the distributions obtained d by fixing a row or column in the matrix and rescaling the entries in that row or column so that they again add up to 1" [14]. Say we want to model the conditional distribution of the response $Y = k$, given that the predictor(s) $X = x$. This is expressed in the form: $P(Y = k|X = x)$.

In an alternative approach to logistic regression, the distribution of X is modeled separately for each response class Y [1]. Bayes theorem is then used to generate estimates of the conditional distribution for the response given predictor(s). Bayes' Theorem is used for classifying observation into one of K classes, $K \geq 2$ where $f_k(x)$ is the density function defined by the given probability function, $f_k(X) = Pr(X = x|Y = k)$. We let π_k denote the prior probability "that a randomly chosen observation comes from the kth category of the response variable," in this case Y [1].

Bayes' theorem is then given as:

$$p_k(X) = Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{1}$$

An estimate of π_k is easily computable by computing the fraction of training data in the kth class, and so it can be substituted into Equation 2 with an estimate of $f_k(X)$.

The cut-point decision boundary estimated by LDA depends largely on the assumption that $f_k(x)$ is Gaussian. If there is only one predictor, such that $p = 1$, then the following is the one-dimensional normal density if $f_k(x)$ is assumed to be Gaussian (normal):

$$f_k(x) = \frac{1}{\sqrt{2\pi\delta_k}}\ exp(-\frac{1}{2\delta_k^2}(x - \mu_k)^2) \tag{2}$$

Where μ_k and δ_k^2 are the mean and variance parameters for the kth class [1]. Using this estimate of $f_k(X)$ and plugging it into Equation 2 yields

$$p_k(X) = \frac{\frac{1}{\sqrt{2\pi\delta_k}}\ exp(-\frac{1}{2\delta^2}(x-\mu_k)^2)}{\sum_{l=1}^{K}\pi_l\frac{1}{\sqrt{2\pi\delta_k}}\ exp(-\frac{1}{2\delta^2}(x-\mu_l)^2)} \tag{3}$$

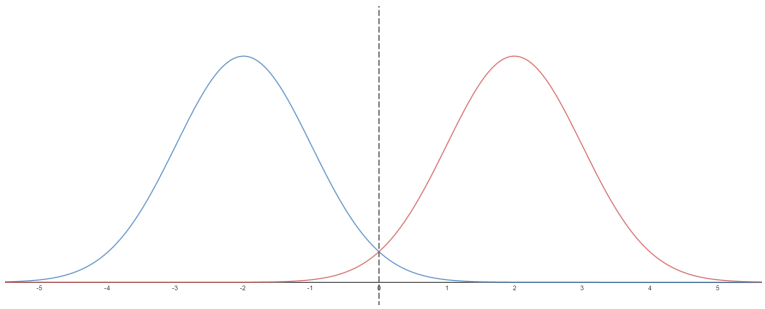

Figure 2: Two normal-density functions with a Bayes decision boundary indicated by a dashed line along the y-axis. LDA estimates this boundary from sample data assuming a normal distribution. Adapted from [1].

By taking the log of Equation 3, it can be shown that assigning an observation $X = x$ to the class for which the equation is largest is equivalent to assigning the observation for the class for which the following expression is also the largest:

$$\delta_k(x) = x * \frac{\mu_k}{\delta^2} - \frac{\mu_k^2}{2\delta^2} + log(\pi_k) \tag{4}$$

The Bayes decision boundary then can be found to correspond to the point where:

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2} \tag{5}$$

In essence, LDA approximates the Bayes classifier by taking estimates of π_k, μ_k, and δ^2 and substituting them into Equation 4 [1]. Two estimates are used:

$$\hat{\mu_k} = \frac{1}{n_k}\sum_{i:y_i=k} x_i \tag{6}$$

$$\hat{\delta^2} = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \hat{\mu_k})^2 \tag{7}$$

$$\hat{\pi_k} = \frac{n_k}{n} \tag{8}$$

Where n is the total number of training observations from the dataset and n_k is the number of training observations belonging to the kth class [1]. The value for $\hat{\mu_k}$ is the average of all training observations from the kth class, and $\hat{\delta^2}$ is the weighted average of the sample variances for each of the K classes [1].

For cases where there are two or more predictors such that $p > 1$, LDA assumes that the observations in the kth class come from a multivariate Gaussian distribution. Parameter estimation similar to the case where $p = 1$ is done.

More specifically, the LDA procedure can be broken down into the following steps when being implemented through Python using scikit-learn:

1. Compute the d-dimensional mean vectors for the classes in the given dataset
2. Compute two scatter matrices: in-between-class and within-class
3. Compute eigenvectors and eigenvalues for the above scatter matrices
4. Choose keigenvectors with the largest eigenvalues from a sorted list of eigenvectors sorted in decreasing order based on their eigenvalues. Then form a $d * k$ matrix W.
5. Use the matrix of eigenvectors to transform samples onto a new subspace.

A more in-depth resource for LDA can be found using Raschka's resource [13], or by reading James, *et al.* [1] Additionally, James, *et al.* give three reasons why LDA would be preferred over logistic regression:

1. It is not hampered by instability when classes are well-separated
2. It is more stable for small values of n with an approximately normal distribution of the predictors.
3. It is more commonly used when there are two or more response classes.

4.3 Support Vector Machines

A method for data analysis, also known as SVM, support vector machines are best considered when binary classification is concerned [15]. The machine is a given set of vectors in a space and their labels. The SVM in its simplest

form is a hyperplane that separates the data by a maximal margin. The hyperplane can be considered up to a p-dimensional setting and its equation has the form:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0 \tag{9}$$

In the above equation β_1 through β_p are the parameters and $X = (X_1, X_2, ...X_p)^T$ is a point in p-dimensional space. If a point put into that equation ends up greater than 0 then it lies on one side of the hyperplane and if it is less than 0 then it lies on the other. Tong and Koller [15] tell us that if this is the case then of all vectors lying on one side of the hyperplane are then labeled as -1, while those on the other side of the hyperplane are labeled as 1. Now we have two separate classes of points [15]. Once a hyperplane that can separate the data into classes is determined one may find that there are infinite hyperplanes that divide the data. In this case a maximal margin hyperplane is chosen so that the hyperplane is farthest from the training observations [1]. All of this can be seen in the linear SVM example below which was created a basic set of points.

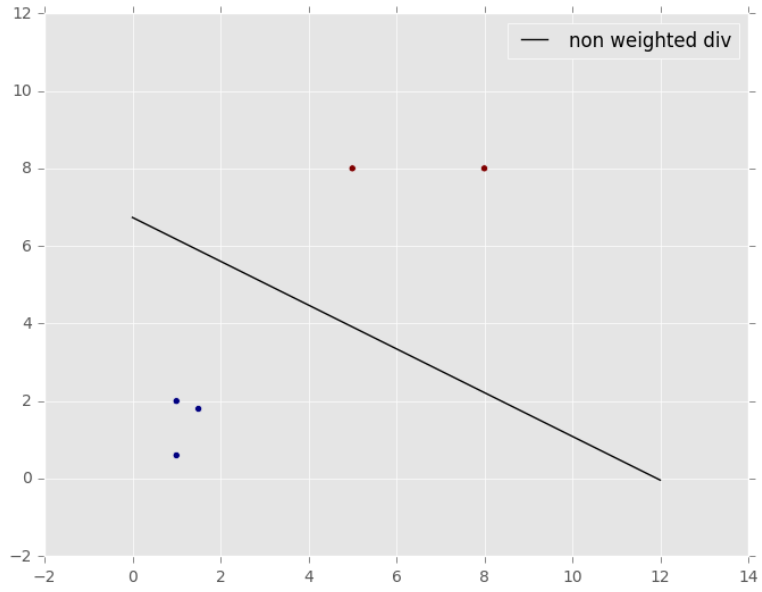

Figure 3: Example of a visual representation of a Support Vector Machine in which a small number of points are classified based on their attributes

The training instances that lie closest to the hyperplane are then called support vectors which a classified to help build the SVM. The SVM itself is an extension of the support vectors that results from enlarging the feature space in a specific way using something called kernels [1]. Kernels are similarity functions that a programmer provides to the algorithm to determine how similar two pieces of data are. Kernels are usally linear, but can be polynomials creating different shaped boundaries instead of just a straight line. An example of a kernel equation can be seen below.

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{10}$$

The above kernel equation is one of a linear nature, but they can be polynomial as well. By adding 1 and raising the expression to a power d, then the result is a polynomial kernel of degree d. When a non-linear kernel is combined with the support vector classifier the resulting classifier is a support vector machine. This non-linear case has the form

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \tag{11}$$

Generally speaking, SVM works best in a two-class or binary setting. The concept of separating hyperplanes upon which SVMs are based does not lend itself to more than two classes [1]. There are a myriad of proposals on how to extend SVM to a k-class case, but no one proposal has been chosen as definitive at this point. One of the most popular approaches is called one-versus-one, which constructs $\binom{k}{2}$ SVMs which compares pairs of classes [1]. The other is called one-versus-all which fits k SVMs, each time comparing one of the k classes to the remaining $k-1$ classes [1]. Neither of these are considered the universal way to approach multiple classes.

4.4 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a classification method used in data analysis, in which given an integer k, the algorithm looks at the K points around a specified point to determine what class that specified point would fall into [1]. Using the equation below,

$$Pr(Y = j | X = x_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j) \tag{12}$$

The algorithm estimates the conditional probability for each class in the target variable and then picks which class the specified point is based on the classes from its neighbors [1]. The algorithm chooses the class for the specified point based on which class has the highest probability. Choosing k is extremely important because if the k-value is too large, then it can increase the bias toward the training dataset in the model and cause overfitting. However, an extremely small value of k will not predict well on either dataset [1]. KNN is extremely good at predicting models correctly, it is however extremely expensive when it comes to computing power. Thus, in a large dataset it, it is not the most ideal method to use.

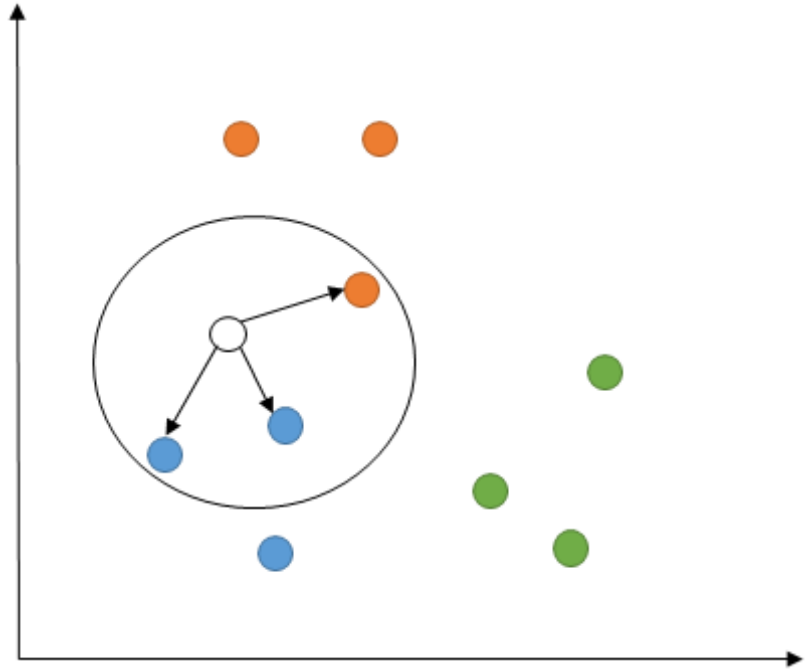

Figure 4: Example of a K-Nearest Neighbors, the blue, orange and green points represent different classes and the black outlined point represents the point trying to be classified with $k = 3$. Adapted from [1].

The above figure shows what the KNN algorithm is design to do. In this image, there are three classes, blue, orange and green. The black outlined point is the point that this algorithm is trying to predict. Here $k = 3$, so the algorithm looks for the three closest points to the outlined point to determine which class that point would be in. The three closest points are

circled in Figure 4. Two of the circled points are from the blue class and one point is from the orange class. This is where the algorithm votes and blue wins this vote two to one against orange. Therefore the algorithm will decide that the outlined point is from the blue class and will predict "blue."

4.5 Decision Trees

Decision trees are a statistical method that segment the predictor space into many regions and then use those regions to make predictions on the data [1]. The regions are called nodes or leaves and the segments connecting the nodes are called branches [1]. There are two kinds of decision trees: regression trees and classification trees. A regression tree is used for regression problems, where the tree will guess a number at the end of the tree. A classification tree is used for a classification problem, where the tree will guess a class at the end. An example is provided below in Figure 5.

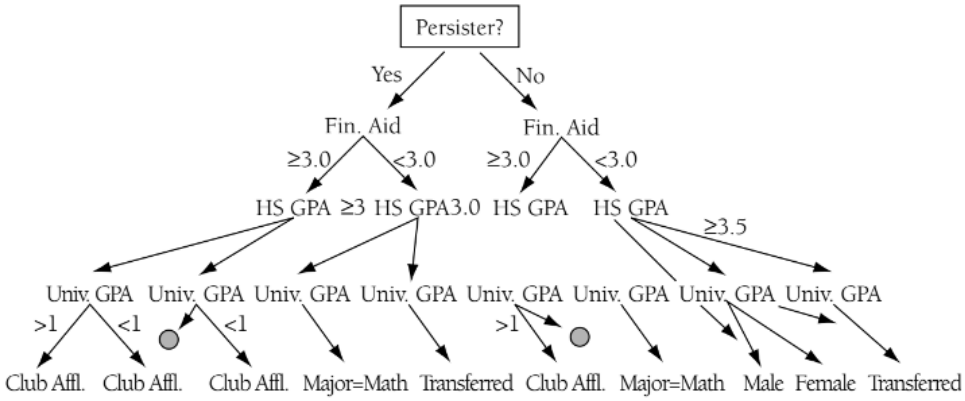

Figure 5: Example of a decision tree, in which a set of rules determines whether or not a student is a "persister, and then runs through possible choices for each subcategory based on other features to make predictions about major, gender, and club affiliation [2].

For purposes with both the test dataset and the CDC datasets, we used classification trees. A classification tree is grown by creating binary splits in the data in order to make a prediction. Each split is made by observing the gini index to determine the best split per region [1]. The gini index is a measure of node purity, so a small value indicates that the leaf contains observations from mostly one class. The gini index equation is given as:

$$G = \sum_{k=1}^{K} p_{mk}(1 - p_{mk}) \tag{13}$$

The following figure shows an example of how these binary cuts are made. This figure shows four cuts that were made in the CDC dataset. The first cut (red line) is how many Walk-Ins the student attended and the cut is made at five Walk-Ins. The next cut (blue line) is when a student received their degree, this cut was made at two where two is spring and zero or one is fall or winter. The third cut (green line) is how many Career Fair the student attended, this cut was made at two Career Fairs. The final cut (purple line) is what type of degree a student received. This cut was made at six where four is Bachelors Degree and six and ten are Masters and PhD. A decision tree can make more than just these cuts, this is just an example of one of the pathways a tree can make.

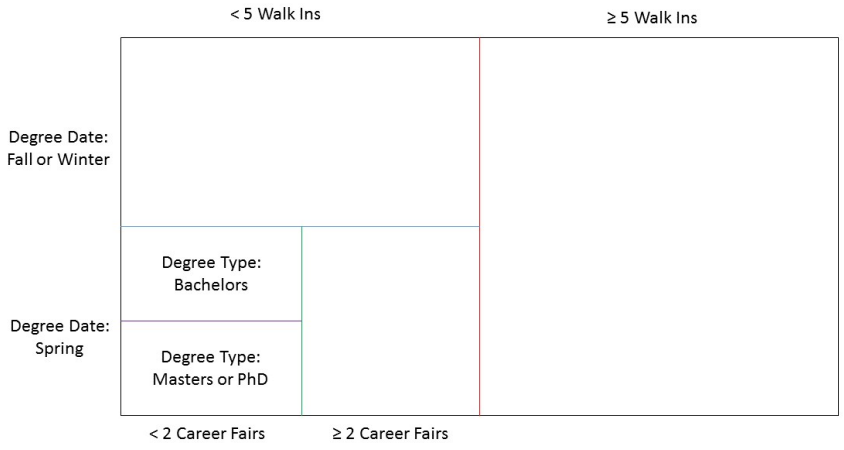

Figure 6: Example of how a tree makes binary splits using the CDC Usage dataset top four categories

5 Methodology

The following is a summary of our methodology for this study, following the planned project outline as detailed in the previous Chapter.

5.1 Practice Data Testing

Before working with specific self-reporting data from the CDC, we used a practice dataset from the UCI Machine Learning Repository [16]. The dataset we used measured student performance in secondary education at two schools in Portugal. There was two sets, one that focused on performance in math and the other in learning the Portuguese language. We mainly focused on the students performance in math. The dataset had thirty three categories of information on all types of topics that was gathered through school reports and questionnaires.

Once we became familiar with the data we had to use our understanding of the dataset to clean it up for analysis. We looked for unusual patterns and found one involving students with a final grade of zero. After some research we figured out a reason for the zero final grade which was that it meant the student dropped out or did not finish the school year. We determined it was safe to remove those from the set for analysis purposes. We had decided that we would be predicting the final grade using all of the other categories.

From there we had to determine what predictors we wanted to use. We ended up removing fourteen of the thirty three categories for several reasons. One reason for removing categories was based on subjective ratings. Some predictors such as health, alcohol consumption, and free time were rated on a scale of one to five with no standardization so that could skew the data. Other categories were removed because there was not enough variation in the data and that would imbalance the data for our calculations. The desire to continue on to higher education, for example, was near unanimous so it was removed. Finally, other categories were removed because they seemed irrelevant to the analysis. For example, since we were trying to predict the final grade, we determined that whether or not studetns were in a romantic relationship or not was irrelevant.

The next step involved encoding some of our predictors so that we could analyze them. Quantitative data is easier to process than qualitative data. Categories such as which school a student attended, what their gender was, and their parents cohabitation status were transformed from qualitative answers to numerical values. Other categories that were binary Yes or No answers were converted into a 1 for Yes and a 0 for No. Once all of this was completed we could perform some analysis.

We were able to use this practice dataset to try out different methods of analysis. There were four different machine learning techniques we used: k-nearest neighbors, decision trees, linear discriminant analysis, and support vector machines. We coded these methods in Python and then ran our

data through them. From there we looked at the error rates and what we could learn from these methods. We decided that for the dataset received from the CDC, decision trees would be the best method to use. This set the groundwork for our decision-making process for the actual datasets we received from the CDC.

5.2 Career Development Center Datasets

We received access to the datasets from the CDC and applied the following methodology to begin our analysis. This methodology involves data clean up, intersecting multiple sets of data, training and testing the data in our programs, perfecting feature selection, creating decision trees and pruning them so that they provide the best information.

5.2.1 Data Clean Up

Few datasets are naturally well-formatted and "clean" to begin wtih. In order to process a dataset, it must be reviewed to ensure that information is not inaccurate, mischaracterized, missing altogether, or otherwise erroneous. One of the first things we noticed in the dataset was that there doubles of some student identifiers. This was because some students had double majors, so they were listed twice with each major listed in both of the categories labeled Major 1 and Major 2. To prevent redundancy we removed the second instance of all of these cases making sure to leave the the instance where their primary major was listed in the Major 1 column.

We then had make sure that for categories that had a written response that the answers were all spelled the same. We then renamed and combined some answers within categories so that we did not have categories that were too small for analysis. For example, because there are such a large amount of individual majors at the undergraduate and graduate level offered at WPI, it was necessary for us to group similar majors together into larger categories. Mathematics, for instance, was designed to include Mathematical Sciences, but also Actuarial Mathematics, Financial Mathematics, and other similar majors.

Next we had to pull out only the students from the class of 2015 from the CDC usage and Internships data sets. The data we got was a massive file of every student who used the CDC in a certain time period and we had pull all relevant data. The same goes for the internship data. Finally, we had to decide which categories from all of these datasets that we wanted to use for our analysis. This was not a final list, but a starting point because

we had other ways, as described below in great detail, to make sure we were selecting the right pieces of data.

5.2.2 Set Intersection

The first step before analysis could begin was to create a single, easy to process data package. Using the common key identifiers in each dataset, corresponding to unique students, a simple intersection scheme was designed in Excel to create a single database from which data could be encoded and processed easily, with relevant information from all datasets present in one location.

	ACADEMIC			CDC USAGE			INTERNSHIPS	
ID	DEGR	Major 1	Unique Identifier	Career Fair	Company Info	Unique Identifier	Co-Op (Traditional)	Summer Co-op
nESjF	BS	Mechanical Engineering	bT1TA	1	0	W61EI	0	0
nESjF	MS	Management	bT1TA	1	0	aXJwu	0	0
oLhry	MS	Materials Science & Engineering	bvz/Y	8	1	B55un	0	1
oLhry	BS	Mechanical Engineering	ul7IG	8	1	grRGw	0	1
dFjfO	BS	Robotics Engineering	TjhiD	2	0	I1hJc	0	0
dFjfO	MS	Robotics Engineering	qcrwr	2	0	aXJwu	0	0
k5wJW	BS	Civil Engineering	Dw7px	2	0	ZxDXO	0	0
k5wJW	MS	Civil Engineering	IzbPV	2	0	WVtiJ	0	0
k3dxP	BS	Industrial Engineering	R0j26	3	0	Tr1af	0	0
k3dxP	MBA	MBA	9mffW	3	0	s+ZhT	0	0
17td8	BS	Aerospace Engineering	Fr+QB	1	0	4g9w5	0	0
17td8	MS	Manufacturing Engineering	nzieU	1	0	YS81g	0	0
42g6R	MS	Electrical & Computer Engineering	AD9Hx	5	0	Pn9cp	0	0
42g6R	BS	Electrical & Computer Engineering	GFYId	5	0	1ee9i	0	0
sRPwn	MS	Mechanical Engineering	Fr+QB	0	0	B55un	0	0
sRPwn	BS	Mechanical Engineering	9mffW	0	0	1yu95	0	0
5lAUl	BS	Mechanical Engineering	ayYPW	3	0	eFGBw	0	0
5lAUl	MS	Materials Science & Engineering	EvBNB	3	0	JQip8	0	0
XHfKK	BS	Mechanical Engineering	xlq3S	4	0	IHJJy	0	0
XHfKK	MS	Mechanical Engineering	xmIlc	4	0	WEvx+	0	0
xVoDX	BS	Mechanical Engineering	/gVhm	6	1	SER1D	0	0
xVoDX	MS	Mechanical Engineering	3h3XI	6	1	bT1TA	0	0
lKOyr	MS	Systems Engineering	wEwEj	0	0	g+ZT/	0	0
nB+H+	MENGR	Power Systems Engineering	+FzJI	0	0	wTz8x	0	0
DHkGA	MBA	MBA	yV48t	0	0	qh9LK	0	0
mMURN	MS	Construction Project Management	YZ2Yd	0	0	Hapfr	0	0
xhLTW	BS	Psychological Science	UoL6s	0	0	wo8ym	0	0
I1hJc	BS	Mechanical Engineering	IjXw2	7	0	8X3vD	1	1
Ir/ax	MS	Interactive Media & Game Development	n8Rip	0	0	aXJwu	0	0

Figure 7: Selection from our intersection dataset combining information from three datasets: the class of 2015 set, CDC usage, and internship/co-op data.

The system works as follows. First, each dataset (CDC Usage and Internships) was filtered in its own file. Identifiers from the 2015 outcomes file were imported. These were then compared against the identifiers in the original set using a simple MATCH() function. The matches were called out and sorted so that a file was produced containing only data whose members were common to both sets.

When entered into the common merged dataset, the two intersection sets of identifiers from the CDC Usage and Internship files were imported along with relevant categorical information from each file. A COUNTIFS() function in Google Sheets was implemented to create a frequency matrix to encode relevant features against the IDs in the primary set. In the above

table, the column highlighted green are these primary IDs, against which all other sets were compared. The blue column is from the CDC usage file and the golden column is the identifiers from the Internship data.

None of the latter two columns required sorting to align with the first column of IDs, nor did they actually need to be included and processed during the analysis. They only exist in the file to be used as references to generate the frequency matrices. For example, take the identifier nESjF in the first entry. The COUNTIFS() function on nESjF checks if it appears at all in the CDC Usage column Unique Identifiers. If it does not, a 0 is produced. If it does, it then counts how many instances of the category are affiliated with that ID. In the case of Career Fair, it checks to see how many nESjF identifiers have a Career Fair associated with them. That is, it checks how many times that student used attended a Career Fair. Since the count is specific to the row including the first columns IDs, the remaining list of identifiers doesnt need to match them. The COUNTIFS() are specific to the ordered list in the first column. In this way, data can be easily and dynamically added or removed from the intersected files to allow us to quickly encode and visualize data common to all three major datasets.

5.2.3 Further Categorization

Initially, we used a set of high level categories in order to classify CDC program usage. There were six in total. The list included: Career Fair, Company Info Session, Pre-Career Fair Event, Resumazing, Walk-In, and Workshop. Below is an example of how we used a matrix using the system described in the previous section based on these sections.

However, this was not going to be the final state of our analysis. The CDC would not benefit from merely knowing information about their services at this high of a level. For instance, Workshop at the time had well over 100 subcategories which were specific to event type, but also major class and location (for example, an event targeting chemical engineering majors or one held at a Greek house). And the Career Fair is held several times during the year and there are related Career Fairs which target specific major domains at times. To capture this level of detail we had to do two things. First, we needed to re-classify the data to the best of our abilities. This was done using Excel and transforming the column of existing Workshop categories. For the other columns of data, a check on feature relevance for each sub-category was done by filtering to find the occurrences of all other sub-categories in the CDC Usage dataset and then using formulas to calculate the percent that each occurred under its meta category. See Figure 9

		CDC USAGE MATRIX						
Unique Identifier	Kiosk Category	Career Fair	Company Info	Pre-Career Fair	Resumazing	Walk In	Workshop	Total
bT1TA	Walk In	1	0	0	0	0	0	1
bT1TA	Walk In	1	0	0	0	0	0	1
bvz/Y	Walk In	8	1	0	1	3	3	16
ul7IG	Walk In	8	1	0	1	3	3	16
TjhiD	Walk In	2	0	0	2	1	2	7
qcrwr	Walk In	2	0	0	2	1	2	7
Dw7px	Walk In	2	0	0	0	6	10	18
lzbPV	Walk In	2	0	0	0	6	10	18
R0j26	Walk In	3	0	0	2	0	2	7
9mffW	Walk In	3	0	0	2	0	2	7
Fr+QB	Walk In	1	0	0	0	0	1	2
nzieU	Walk In	1	0	0	0	0	1	2
AD9Hx	Walk In	5	0	0	0	5	2	12
GFYId	Walk In	5	0	0	0	5	2	12
Fr+QB	Walk In	0	0	1	0	1	1	3
9mffW	Walk In	0	0	1	0	1	1	3
ayYPW	Walk In	3	0	1	0	3	2	9
EvBNB	Walk In	3	0	1	0	3	2	9
xlq3S	Walk In	4	0	0	1	1	2	8
xmllc	Walk In	4	0	0	1	1	2	8
/gVhm	Workshop	6	1	0	0	3	7	17
3h3XI	Walk In	6	1	0	0	3	7	17
wEwEj	Walk In	0	0	0	0	0	0	0
+FzJI	Walk In	0	0	0	0	0	0	0
yV48t	Walk In	0	0	0	0	0	0	0
YZ2Yd	Walk In	0	0	0	0	1	3	4
UoL6s	Workshop	0	0	0	0	0	0	0
IjXw2	Workshop	7	0	1	4	4	13	29
n8Rip	Workshop	0	0	0	0	0	0	0

Figure 8: Frequency matrix in our original intersecting dataset using the six basic high-level categories for CDC usage.

below for examples of Walk-In and Career Fair categories.

Second, we needed to discuss this one-on-one with the CDC to determine their perspective. Our sub-category re-classification, mainly for the Workshop category, was the one left primarily to interpretation because there were so many sub-categories with minor significance, similar names to other categories, subtle differences in details, and so on. We met in person and then exchanged our original categorization scheme for them to consider. Figure 10 is the outcome of that meeting and exchange, and was finalized by the team as the Workshop sub-category spread that we would use to expand our CDC Usage features. Time constraints meant that for the sake of this project, we did not investigate to this level of depth for Workshops. Walk-Ins tended to be more accurate predictors than Workshops, but Workshops did have minor relevance in the SMOTE and Borderline-SMOTE runs of the Internship dataset, for which the two versions of the algorithm performed about equally well. Therefore, it is worth including these sub-categories in future models. We discuss this in more detail in Chapter 10, Limitations and Future Works.

5.2.4 Predictors and Target Variables

The datasets had 28 predictors and one target variable that we decided to use. In total, 4 of those predictors were categorical, 8 of them were made

Category	Sub-Category	Frequency	Percent
WALK IN	Walk In	1582	92%
	Drop-Box Resume Critiques	38	2%
	CDC Hits The Road	50	3%
	Email Resume Critiques	55	3%
	TOTAL	1725	100%

Category	Sub-Category	Frequency	Percent
CAREER FAIR	2011 Fall Career Fair	172	5%
	2012 Fall Career Fair	297	8%
	2013 Fall Career Fair	403	11%
	2014 Fall Career Fair	633	18%
	2015 Fall Career Fair	3	0%
	2011 Spring Career Fair	25	1%
	2014 Spring Career Fair	492	14%
	2015 Spring Career Fair	451	13%
	2012 Life Sciences & Bioengineering Career Fair	27	1%
	2013 Life Sciences & Bioengineering Career Fair	111	3%
	2014 Life Sciences & Bioengineering Career Fair	94	3%
	2015 Life Sciences & Bioengineering Career Fair	146	4%
	Biotech, Health & Life Sciences Virtual Career Fair	57	2%
	Bio Fair	2	0%
	Summer Internship & Job Fair	663	19%
	TOTAL	3576	100%

Figure 9: Example of subcategory breakdown from the original CDC Usage dataset, with Walk-In and Career Fair as the first two major categories considered.

from the Major 1 and Major 2 variables and 8 of them were nominal. The categorical variables had to be encoded so that the model could understand what the categories were. The variables made from Major 1 and Major 2 had to be one-hot-encoded, using ones and zeros to describe whether a student is a major or not. The following list describes what each of the variables mean, how they were categorized and how they were encoded.

Nationality: This predictor showed whether a student is and International Student or Not.

Encoding: 0 meant they were not International and 1 meant they were international. This was chosen because it needed a simple yes or no on whether the student was International or not.

Degree Type: This predictor showed what type of degree a student received.

Encoding: 4 meant the student received a bachelors degree, 6 meant the student received a masters degree and 10 meant the student received a PhD. This encoding was based on the average amount of time it takes to earn that degree starting from freshman year. So, it takes four years to earn a bachelors degree, 2 more to earn a masters and 4 more to earn a PhD.

Proposed Category List by MQP	CDC Recommended Categories
Intro to CDC	CDC Overview
Careers in...	Careers
Cover letter	Cover Letter
Job search	Job Searching
Interview	Interviewing
Graduate school	Graduate School Advising
Job offer	Job Offer
Start smart	Job Offer
Group/organization/department titled things that include other career topics	I don't believe there were anything in here-delete
Resume	Resume/CV
Marketing your..... experience (athletic, eng without boarders, greek life,	Marketing Yourself
CV	Resume/CV
CL Writing (CL is our abbreviation for cover letter)	Cover Letter
Careers and Job Search....	Job Searching
Job Finder	Job Searching
LinkedIn	Networking & LinkedIn
Sophomore TLC	Job Searching
Americanizing....	Resume/CV
Developing Your Elevator Pitch	Career Fair Preparation
How to Sell your Overseas....	Marketing Yourself
How to work a career fair	Career Fair Preparation
Finding a federal government job...	Job Searching
Company interview panel	Interviewing
Any type of panel	Careers
Greek Career Night	Careers
Why should 1st year students go to career fair	Career Fair Preparation
Kick off event	CDC Overview
CDC Game Night	CDC Overview
Any SOB, or other department, Senior Class Meeting	Job Searching or Networking - & LinkedIn based on the program titles
Other	
Working with Your Image	Marketing Yourself
Transition Workshop for Seniors	Job Searching
Time to Leverage Those Leadership Skills	Marketing Yourself
Skills that Will Get You Hired	Marketing Yourself
Networking Workshop for ESL	Networking & LinkedIn
Networking Workshop (Special Request, MBAWI)	Networking & LinkedIn
Networking Workshop	Networking & LinkedIn
JSS, SR: Fin Math	Job Searching
JSS Workshop	Job Searching
How to Sell Your Skills as a Graduate Student	Marketing Yourself
How to Introduce Yourself to Employers	Career Fair Preparation
How to Find a Job	Job Searching

Figure 10: Our breakdown of kiosk Workshop categories (left) and the CDC's recommendation based off this list (right).

Degree Date: This predictor showed when a student received their degree.

Encoding: The dates started out as October, December and May which were then translated to Fall, Winter and Spring respectively. These words were then encoding by 0 being fall, 1 being winter and 2 being spring. This encoding was chosen because there needed to be a way to distinguish between the times when students graduated but the time between each did not need to be considered for our purposes as we just wanted to see if the degree date determined primary status.

Gender: This predictor shows the gender of each student.

Encoding: 0 meant the students was male and 1 meant the student was female. We just needed a simple encoding for gender to identify which students were male and which were female.

The following variables are the nominal variables and do not require encoding

Career Fair: This predictor indicates how many Career Fairs each student attended. This includes the Fall Career Fair, Spring Career Fair and Life Sciences Career Fair.

Company Info Session: This predictor indicates how many company info sessions a student attended. Examples of subcategories of this predictor include Info Sessions hosted by Analog Devices, Cisco Systems, Google, Juniper Networks, NASA, Sonos, and Towers Watson.

Pre-Career Fair: This predictor indicates how many pre-Career Fair events a student attend. The three subcategories of this predictor are Employer Resume Critiques, How to Work a Career Fair with Pegasytems, and How to Work a Career Fair with GE.

Resumazing: This predictor indicates how many resumazing events a student attended. Resumazing is a resume critique that is done by local employers instead of the CDC.

Walk-In: This predictor indicates how many Walk-Ins a person went to. A Walk-In includes resume critiques done by the CDC, cover letter critiques and mock interviews.

Workshop: This predictor indicates how many workshops a student attended. Examples of subcategories include Job Searching, Career Fair Preparation, Marketing Yourself, Resume/CV, and Cover Letter (based on CDC input on categorization).

Summer Internship: This predictor indicates how many summer internships a student reported having.

Co-op: This predictor indicates how many traditional co-ops or summer co-ops a student reported having.

The following predictors are the one-hot encoded majors, so if there is a one in a category the student was this major and if there was a 0 the student was not this major. Some students had two ones in their row which indicates that the student had a double major. Each of these were made by categorizing every major in the set into more general majors.

Mathematics: This predictor included all the people who indicated their first or second major was mathematics. The majors that went into this category were Actuarial Math, Applied Math, Applied Statistics, Financial Math and Mathematical Sciences.

Aerospace Engineering: This predictor included all the people who indicated their first or second major was Aerospace Engineering. Aerospace Engineering was the only major that went into this category.

Civil Engineering: This predictor included all the people who indicated their first or second major was Civil, Architectural, Environmental Engineering or any major related to them. The majors that went into this category included Architectural Engineering, Civil Engineering, Construction Project Management, Environmental and Sustainability Studies and Environmental Engineering. We chose to put these majors together because they are in the same department at WPI.

Science: This predictor included all the people who indicated their first or second major was Biology, Chemistry or Physics. The majors included in this category were Bioformatics and Computational Biology, Biology and Biotechnology, Chemistry, Biochemistry, Physics, Bioscience Administration and Engineering Physics. We decided to put these majors together because they are all general sciences.

Biomedical Engineering: This predictor included all the people who indicated their first or second major was Biomedical Engineering (BME). BME was the only major that went into this category.

Chemical Engineering:This predictor included all the people who indicated their first or second major was Chemical Engineer. Chemical Engineering was the only major that went into this category.

Robotics Engineering: This predictor included all the people who indicated their first or second major was Robotics Engineering (RBE). RBE was the only major that went into this category.

Computer Science:This predictor included all the people who indicted their first or second major was Computer Science or a major related to it. The majors included in this category were Computer Science, Information Technology and Management Information Systems.

Liberal Arts: This predictor included all the people who indicated their first or second major was related to Liberal Arts. The majors included in this category were Humanities and Arts, Professional Writing, Economic Science, Psychological Science, Society, Technology and Policy, International Studies, Masters for Physics Educators, Master of Mathematics for Educators and Learning Sciences and Technologies. These majors were grouped together because they were all humanities and social sciences.

Electrical and Computer Engineering: This predictor included all the people who indicated their first or second major was Electrical and Computer Engineering (ECE) and those major related to it. The majors included in this category were ECE, Power Systems Engineering, Systems Engineering and System Dynamics.

Materials: This predictor included all the people who indicated their first or second major was related Materials Science. The majors included

in this category were Materials Process Engineering and Materials Science and Engineering.

Business Engineering: This predictor included all the people who indicated their first or second major was related to Business and Engineering. The majors included in this category were Industrial Engineering, Management Engineering, Manufacturing Management and Manufacturing Engineering. These majors were placed together because they are a part of the business school at WPI as well as an "engineering" at WPI.

Mechanical Engineering: This predictor included all the people indicated their first and second major was Mechanical Engineering. Mechanical Engineering was the only major that went into this category.

Interactive Media and Game Development:This predictor included all the people who indicated their first and second major was Interactive Media and Game Development (IMGD). IMGD was the only major that went into this category.

Fire Protection Engineering: This predictor included all the people who indicated their first or second major was Fire Protection Engineering (FPE). FPE was the only major that went into this category.

Business: This predictor included all the people who indicated their first or second major was related to Business. The majors included in this category were MBA, Marketing and Technological Innovation, Management and Operation, Design and Leadership. These majors were placed together because they are all major at WPI's school of business.

The final variable is the target variable. This is the variable that we are trying to predict. This variable is called Primary Status. The data is taken from a survey that students fill out after graduation. The primary status is what the student is doing after graduation from WPI. A student can pick answers ranging from full time job, going to or thinking of pursuing grad school, serving in the military, volunteering, and still seeking or not seeking employment. There is also students who have blank responses, these are the students who did not fill out the survey that the CDC sent out. These students were then placed in the Unknown category.

5.2.5 PCA

Principal Component Analysis (PCA) is a process in which linear combinations of predictors (principal components) are computed and then used to make sense of the data. These principal components are made using p features and then p principal components are made. [1] We wanted to use PCA to visualize our data and see if there were any principal components

that could help further explain it. The first principal component is made by making a linear combination of features with the largest variance. The second component is made using features with the largest variance that is uncorrelated with the first component. [1] These linear combinations followthe logic of the following equation.

$$Z_n = \phi_{1n}X_1 + \phi_{2n}X_2 + ... + \phi_{pn}X_p \tag{14}$$

In Equation 14, n is the component and p is the number of features. $\phi_{1n}, ..., \phi_{pn}$ are known as the loadings of the principal component. These loadings are used to help make the variance as large as possible among the predictors.[1]

To visualize our data, we ran PCA to find the first principal component. In figure 11, 3 is Full Time Job, 2 is Grad School, 1 is Not Seeking or Serving and 0 is Unknown or Seeking. This figure shows that our data does not have a clear relationship between the classes. It also shows the problem that our data is imbalanced because Full Time Job clearly has more points then any of the other classes.

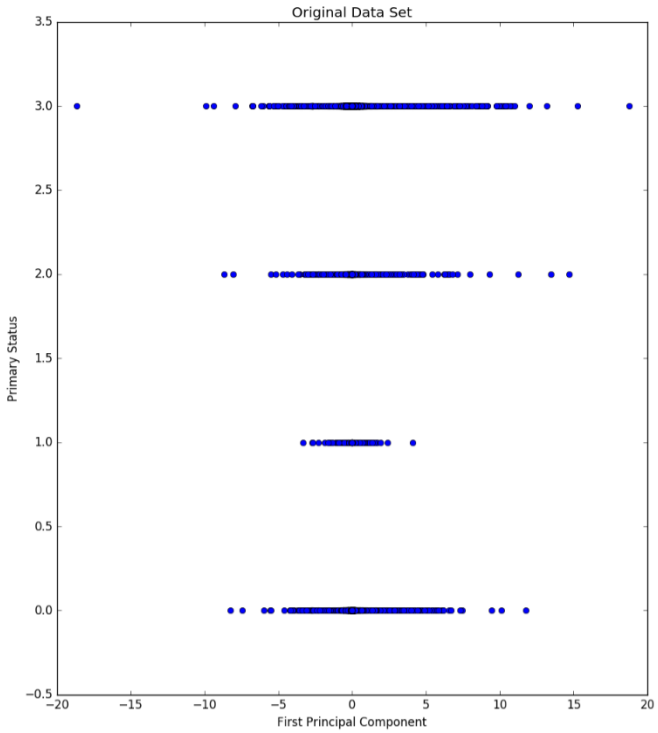

Figure 11: Plot of the First Principal Component versus Primary Status using the CDC Usage dataset

We also investigated relationships between predictors by plotting one component against another. As seen in figure 12, there are only two dots. All the comparisons looked similar to this. Since there were only two dots, this means that there is not a clear relationship among the predictors. Figure 13 shows a comparison of principal components that show a clear relationship.

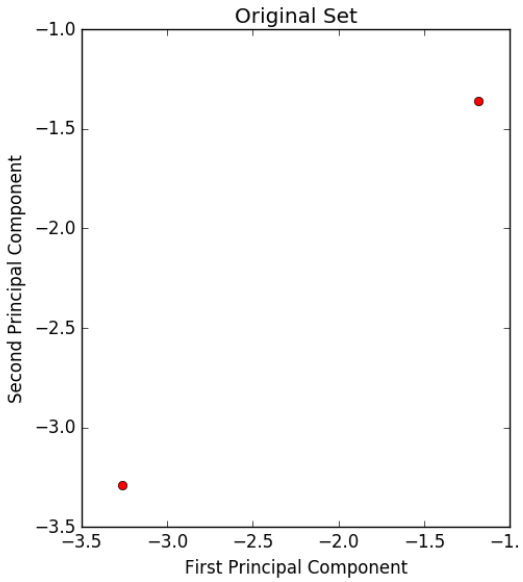

Figure 12: Plot of the First Principal Component versus Second Principal Component using the CDC Data

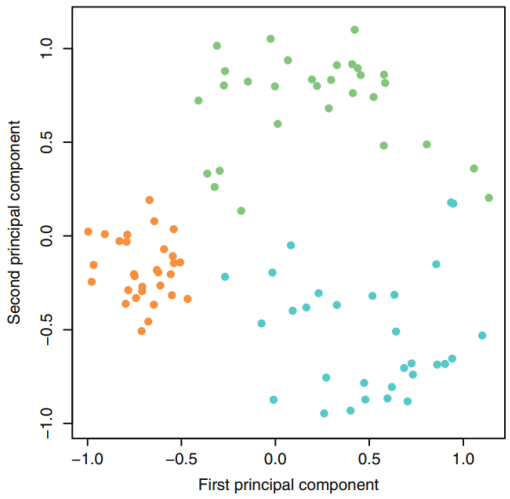

Figure 13: Plot of the First Principal Component versus Second Principal Component from [1]

5.2.6 Feature Selection

Making a model using all the predictors is not a good choice. In our dataset, there are a total of 28 predictors, which makes the model be in 28-dimensional space. A way to make this model smaller is to reduce the number of predictors used but to do this predictors must be chosen in a certain way so that the accuracy or error rate is still good. The way we did this was using Forward Subset Selection. Forward Subset Selection is a method for selecting a subset of predictors by finding the best model using only one predictor, then the best model using that predictor and another predictor and so on until all the predictors are used. [1] The way to make a subset of predictors is to choose the accuracy rate that is the highest for each predictor added and keep going until a predictor is added which makes the accuracy decrease.

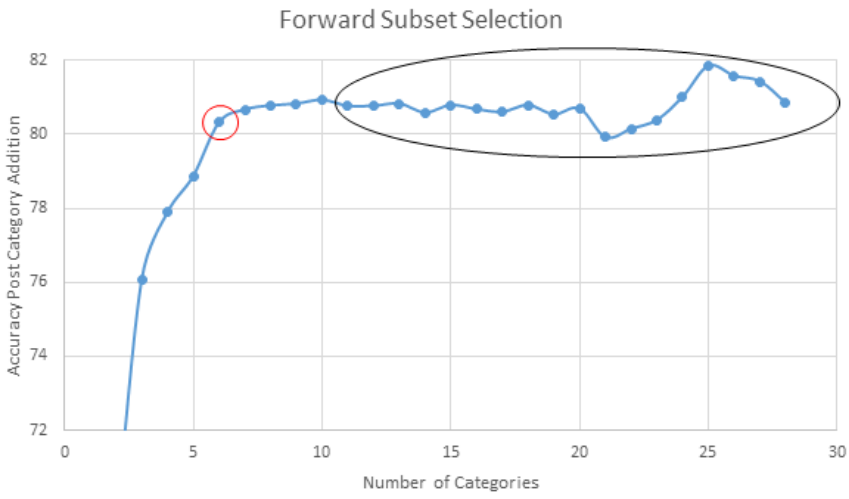

Figure 14: An Example graph plotting the accuarcy post category addition. The red circle indicates where to stop Forward Subset Selection and the black circle indicates noise in the accuarcy

The above graph shows an example of how Forward Subset Selection works. The red circle indicates where to stop considering predictors for the subset of predictors. This is chosen because in the graph this is the first peak encountered. The black circled area indicates noise in the selection. So, while there are points that have a better accuracy then the point chosen, it is not necessarily true that the greater number of predictors will always

have a better accuracy. We chose Forward Subset Selection because it was computationally efficient to do for how many points and predictors we had. We could have chosen Best Subset Selection, which is similar to Forward Subset Selection except it chooses that best model for every number of predictors, however that would be too computationally expensive for us to do in the time we had.

5.2.7 Classification Trees and Bagging

At the end of the analysis of the test data, we decided classification trees were the best way to get the results we needed for the CDC data. This is because trees do not need a separator like LDA or SVM do and the original data did not seem to have a clear separator. Trees also are computationally efficient unlike KNN. Since trees are computationally efficient, we are able to make a lot of them in a small amount of time to get a more accurate result on our data. This allows us to perform bagging, in which many trees are built and the result is averaged [1]. As an added feature of using trees, they provide a simple and straightforward process to follow when they are made, which gives a nice visual for the CDC to use in the future.

We used three different datasets when making these trees: the full dataset, the Internship dataset, and the CDC Usage dataset. The full dataset contained everybody that was included in the self-reporting data that was given to us by the CDC. The Internship dataset contained only those people who reported having internships or co-ops. The CDC dataset contained only those people who had gone to at least one event presented by the CDC. When we would run our bagging program, we would end up with an average accuracy and a summed-average confusion matrix. This summed confusion matrix would be all the matrices added up from k-fold cross-validation and then averaged from all the trees.

5.2.8 Tree Pruning

Tree pruning is the process of making a smaller tree to reduce the chances of overfitting and to improve the performance of a test set[1]. There are two ways to prune a tree. One way is to grow a large tree and then prune it back to make a subtree[1]. The other way is to build a tree that is only so big and then to use cross-validation to decide how big to continue growing the tree. The first way is considered a more desired method of pruning because its alternative might not generate the right splits to produce the best possible subtree[1].

Although the first approach is generally cited as being more effective, scikit learn does not have a way of pruning this way so we have to use the second way. To build small trees, we will be using the categories found from feature selection and then we will use attributes such as min_samples and max_leaf_nodes. Min_samples is an attribute in decision trees for scikit learn that will only make a node if that node has more than the specified number of samples and max_leaf_nodes is an attribute that will grow a tree with a specified number of leaf nodes.

5.3 Imbalanced Data

One major problem we had with the dataset was that the target variable (Primary Status) was imbalanced. This is a common issue wherein a single class is so prevalent that the algorithm guesses the dominant result, without accurately capturing minority classes. In this dataset, there are four major classes based on the students outcome after graduating. They are: full Time, Graduate School, Not Seeking a job or Serving, and Unknown or Seeking a job. Full Time makes up 64% of our data, with Graduate School consisting of 13% and the remaining classes comprising 23%. When decision trees are created without balancing the data most, if not all the decisions predict Full Time job. This way has an accuracy of around 55% which reasonably succeeds at its goal of having a decent accuracy while being entirely misleading and useless for data analysis.

True/ Predicted	**Full Time**	**Grad School**	**Not Seeking/ Serving**	**Unknown/Seeking**
Full Time	94	13	2	11
Grad School	7	8	0	5
Not Seeking/ Serving	3	1	0	0
Unknown/ Seeking	10	4	0	6

Table 5: An example of an early confusion matrix from the same test set as the tree above.

The above tree illustrates how this problem arose and how it impacted our early stage results. Using our original methods, the algorithm only

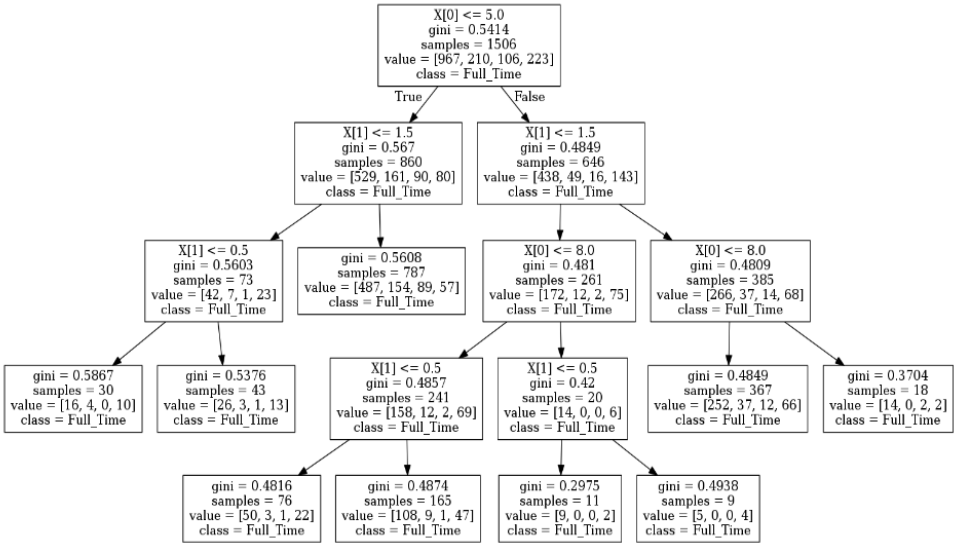

Figure 15: An example of an early tree that we ran on the Full Class of 2015 dataset. Notice how the majority class "Full_Time" is predicted in every instance.

predicted Full_Time, the dominant Full Time majority class. While this is correct most of the time, by the above logic, it still resulted in a massive error rate. The confusion matrix further helps show this. While it predicts accurately for the Full Time employed, predicting 97 of the 117 (82.91%), it does not come close to this accuracy for any of the other predictors. The next best one is graduate school, predicting 8 out of 26 (30.77%). For one that was guessed correctly, one was misidentified as Full Time. The final two categories, Not Seeking/Serving and those who are either Unknown or Seeking employment had even wider discrepancies at this stage. The algorithm predicted Full Time instead of Not Seeking/Serving for both instances of that class, meaning the actual class was not predicted at all (0%). It also did so for 11 out of the 22 Unknown or Seeking cases as well, resulting in an extremely low (27.27%) accuracy. To fix this issue we researched many ways to balance and weight the data to see if we get a better outcome.

5.3.1 Weighting the Data

Weighting the data is one way to make data seem more balanced. To do this, "weights" are assigned to each class to make it seem as though each class has the same amount of weight in the data. Weighting can be done

two ways. The decision tree algorithm has two built in weighting functions, class_weight and sample_weight. Class_weight uses a keyword "balanced to balance the data used for the tree. If the key word is used the algorithm uses an equation to adjust weights that are inversely proportional to the class frequency. The equation used is $n\_samples/(n\_classes*np.bicount(y))$. This means it takes the number of samples in the data divided by the number of classes in the data times the number of occurrences of the class being weighted. Sample_weight is used whenfitting the tree to the training data. Sample_weight needs a weight specified for each sample in the data. The approach we took was to make the weight a ratio of majority class to one of the minority classes. Sample_weight augments the gini measure so that the tree will take a less pure cut of the majority class in favor of having a more pure cut of one of the minority classes.

Type of Weighting	**Accuracy**
No Weights	55.9 %
Class_weight	44.0 %
Sample_weight	43.9 %

Table 6: Chart of Accuracies using weighting functions

Table 6 shows that weighting actually performed worse than the data imbalanced data. Because of this we discovered we had to find a different way to balance the data to improve accuracy.

5.3.2 SMOTE

Through our literature review, we came across an algorithm that aids in balancing data using synthetic minority over-sampling technique, also known as SMOTE. By design, SMOTE avoids over-sampling with replacement by generating synthetic examples of the minority class which are then over-sampled[3]. The overall SMOTE process is illustrated in Figure 16. We found a Python package online called imbalance-learn that could do SMOTE. However, the package did not work well on datasets with extremely few minority instances and the packaged code ran too slow on our large dataset and thus we coded the algorithm ourselves.

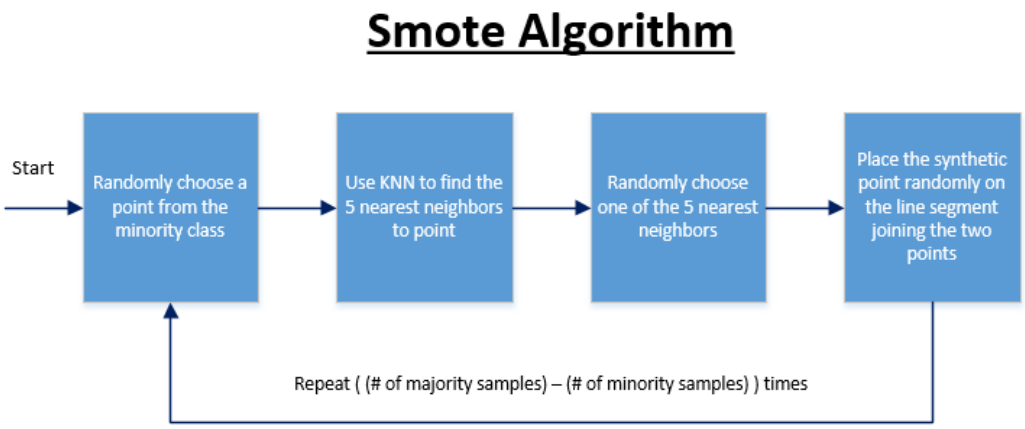

Figure 16: Diagram of the SMOTE algorithm adapted from [3].

The SMOTE algorithm begins by randomly selecting one of the points in the minority class to base the new point off of. In our project, all random selection of points was done by using a built-in Python function called random.randrange(A) which randomly selects an element from the numbers 0 to A (not including A). Almost all of the functions in the Random Library of Python use the Mersenne Twister as the core pseudo-random number generator (PRNG) [17]. The Mersenne Twister algorithm, which the generator is based off, generates random numbers with nearly uniform distribution and a large period length meaning the sequence of generated numbers wont be repeated for a longer time [18]. Along with its speed, this makes the Mersenne Twister pseudo-random number generator popular in many applications such as Python [19].

Next, we find the five nearest neighbors of the chosen minority point. Each data "point in the dataset is a single-row vector where each entry in the vector is predictor. To find the 5 nearest neighbors to the chosen minority point, we must first calculate how far each data point is from the chosen minority point. Since our predictors have different ranges of possible values, we must normalize each column of all the data points (vectors) so that a predictor that has a larger value doesnt dominate the distance calculation [20]. To standardize the points we used the scikit-learn function StandardScaler(). The function automatically transform the data to look like a normal distribution. StandardScaler() is finding the mean and the standard deviation (std) of the data points and then scales the data using

the equation: $DataScaled = (data - mean)/std$. [21] StandardScaler() then stores this transform so that it can later reapply it to denormalize the data. [22]

Once the columns of all the data points are normalized, we now calculate the Euclidean Distance (15) between each point and the chosen minority point:

$$EuclideanDistance = \sqrt[2]{\sum_{i=1}^{27}(\vec{u_i} - \vec{v_i})^2} \tag{15}$$

where $\vec{u_i}$ is the i^{th} entry of the chosen minority point and $\vec{v_i}$ is the i^{th} entry of the data point we are currently looking at. Note: the range of i was determined by the length of the vectors in our big CDC dataset. Once all of the distances are calculated, we sort the distances and choose the five data points that correspond to the five distances with the lowest value. These will be our five nearest neighbors for the chosen minority point. We decided to choose $k = 5$ since this was the standard in the original design coded in [3] and this is also a typical choice for general KNN classification. SMOTE next randomly chooses one of the five nearest neighbors to work with, again using the same built-in PRNG that Python implements. The final step, illustrated by figure 17 is done by placing the new synthetic point on the line segment joining the chosen nearest neighbor and the chosen minority sample [3]. Each new synthetic point is generated as follows:

$$\alpha * |\vec{V}_{knn} - \vec{V}_{minority}| + \vec{V}_{minority} \tag{16}$$

where, α is a random number between 0 and 1, $\vec{V}_{knn}$ is the vector of the randomly chosen nearest neighbor and $\vec{V}_{minority}$ is the vector of the randomly chosen minority point. Notice that Equation 16 has absolute values around the difference of vectors [3]. Originally, we started with an equation that looked like:

$$(1 - \alpha) * \vec{V}_{minority} + \alpha * \vec{V}_{knn} \tag{17}$$

with all of the variables being the same as the variables in equation 16. We quickly noticed that when one expands the equation out:

$$\vec{V}_{minority} + \alpha * (\vec{V}_{knn} - \vec{V}_{minority}) \tag{18}$$

After looking at the equations geometrically, we noticed that both equations 17 and 18 create a point that is not between the two starting points if

$\vec{V}_{minority} > \vec{V}_{knn}$. Thus, the absolute value was placed into the equation to avoid this problem in the future.

After the synthetic point has been given a position, the point is then denormalized using the mean and std from normalizing. From here, the point is then reviewed by a series of if statements to ensure our encoding stayed intact. Once the synthetic point has been generated, the algorithm repeats the process until the number of synthetic points made is equal to (number of majority samples) − (number of minority samples).

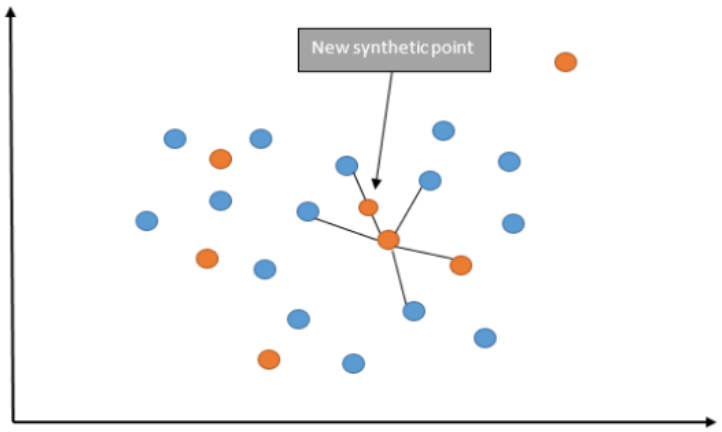

Figure 17: Illustration of the SMOTE algorithm adapted from [4] using $k = 5$.

We decided to compare the results from using SMOTE to balance the data and simply using the imbalanced data. We decided to compute confusion matrices for each case. Table 7 represents the confusion matrix showing the results of prediction without using SMOTE, in which the majority class is over-predicted for all classes. Table 8, representing the confusion matrix produced by using the over-sampling procedure with SMOTE, yields results are much closer to their targets from the training set overall. For example, Table 7 tells us that the model didn't guess any of the Not Seeking/Serving students correctly while Table 8 tells us that once SMOTE was applied to the dataset, our model was able to predict some of the Not Seeking/Serving students correctly. This provided evidence that SMOTE is balancing the data which will stop our model from blindly guessing Full Time as the correct answer all the time.

True/ Predicted	Full Time	Grad School	Not Seeking/ Serving	Unknown/ Seeking
Full Time	97	13	2	11
Grad School	7	8	0	5
Not Seeking/ Serving	3	1	0	0
Unknown/ Seeking	10	4	0	6

Table 7: Confusion Matrix on Full Class of 2015 dataset without using SMOTE.

True/ Predicted	Full Time	Grad School	Not Seeking/ Serving	Unknown/ Seeking
Full Time	75	12	2	18
Grad School	14	58	13	23
Not Seeking/ Serving	1	7	90	10
Unknown/ Seeking	24	21	13	49

Table 8: Confusion Matrix on Full Class of 2015 dataset using SMOTE.

We also modified the value of k in equation 12 to see its effects on the SMOTE algorithm and to verify the approach taken by [3]. We ran SMOTE using two standard values for k, namely $k = 5$ and $k = 10$. Figure 18 shows that when using a value of ten for k in equation 12, we seem to generate synthetic points whose overall distribution is tighter and "closer" to the original minority points that were over-sampled. However, using a value of five for k seems to give us less extreme outliers in the newly generated, synthetic minority samples meaning these points would more likely be misclassified as there are no minority points around that area. Simply comparing the two distributions in Figure 18, it's hard to determine which value of k is superior, so for most of the project we decided to go with

the literature [3] and use a value of five for k to provide consistency.

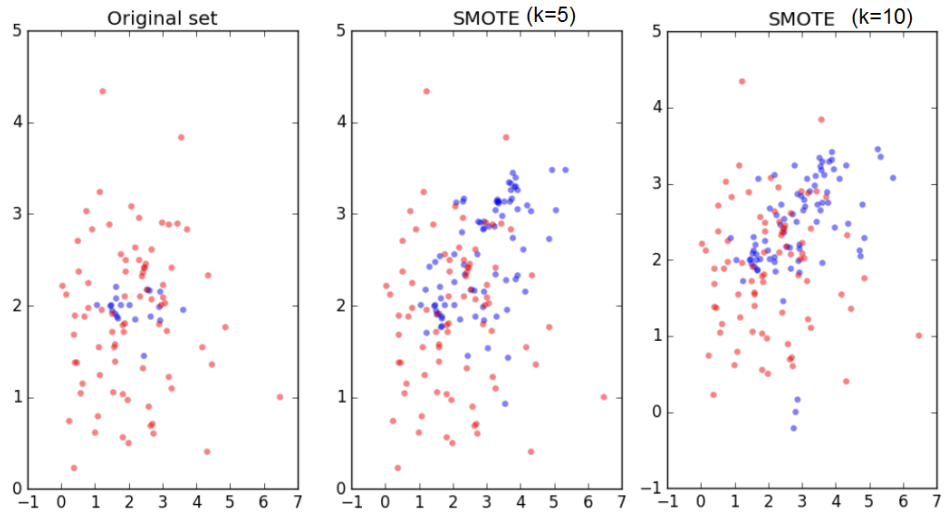

Figure 18: Difference in synthetic point generation using $k = 5$ and $k = 10$ in the SMOTE algorithm. Adapted from [5].

5.3.3 Borderline-SMOTE

While the application of the SMOTE algorithm has improved results, the algorithm is not perfect. The SMOTE algorithm focuses on using the nearest neighbors of one of the minority points to generate the new synthetic point, but the algorithm does not consider what those neighbors look like. This lack of consideration can cause problems such as overlapping between majority and minority classes [4]. Through further research, we were able to find an improvement on the SMOTE algorithm called Borderline-SMOTE. The overall process of Borderline-SMOTE is illustrated in Figure 19. Similarly to SMOTE, Borderline-SMOTE randomly chooses one of the five nearest neighbors to use to generate the new synthetic point.

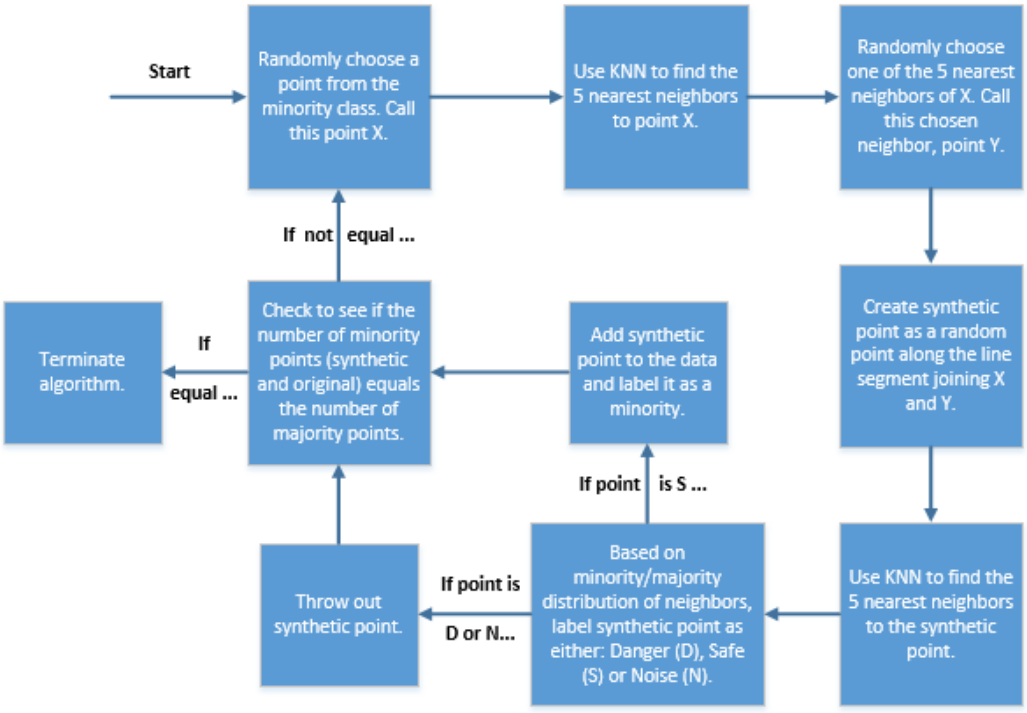

Figure 19: Illustration of the Borderline-SMOTE algorithm.

Once the synthetic point is generated, the Borderline-SMOTE algorithm then places the point into one of three different categories: "Danger, "Noise and "Safe. To determine these categories, the algorithm finds the Euclidean distance of all the neighbors of the synthetic point, makes a sorted list of these neighbors and takes note of what class the neighbors are a part of. If a synthetic point has almost all (or all) of its nearest neighbors of the majority class, then it is put into the Danger class as it has a higher chance of being misclassified. A synthetic point whose nearest neighbors are all, or almost all, of the minority class will be put into the Noise category as they wouldnt contribute to the data since we already know there is a plethora of minority points in that area [4]. The Borderline-SMOTE algorithm only accepts synthetic points which are put into the Safe category meaning they satisfy the inequality:

$$\frac{m}{2} \leq |S_{i:m-NN} \cap S_{maj}| < m \tag{19}$$

Where $S_{i:m-NN}$ is the set of nearest neighbors for each point in the

minority class, S_{maj} is the set of majority points, and m is the number of minority samples in the data [4].

Borderline-SMOTE is designed to only keep the synthetic points that meet a specific criteria causing issues for minority classes of small sample size. Our team ran into this issue while implementing Borderline-SMOTE on the big CDC Dataset. Our original categorization of "Primary Status" consisted of: Full Time Job, Graduate School, Not Seeking/Serving, and Unknown/Seeking Employment. While this categorization worked for SMOTE, once we ran Borderline-SMOTE, our program hung for hourstrying to synthetically generate points for the minority class "Not Seeking/Serving". After debugging the code, we observed that Borderline-SMOTE was throwing away every single synthetic point being generated and therefore was never making any "acceptable" synthetic points. We believe this is due to the small sample size of this minority class and could also have to do with how these minority class points are distributed because if all of these points are clumped together away from majority points, Borderline-SMOTE may keep classifying the new generated points as "Noise" and choose not to keep the point.

To get around this issue, we had to come up with a new categorization. First, we decided that since the students at the university who are enrolled in the ROTC program sign a contract to serve their country for a minimum of five years after graduation, we saw this as a set in stone "Primary Status" that couldn't be changed by the predictors we were working with. Since their status was already chosen for them by signing this contract, we felt comfortable taking these few military students out of our dataset. Our new, and final, categorization for "Primary Status" is: Full Time Job, Graduate School and Unknown/Not Paid. The Full Time Job and Graduate School groups had the same criteria as before but the Unknown/Not-Paid group consisted of students whose status was: unknown, not seeking employment, seeking employment and volunteering. After running both SMOTE and Borderline-SMOTE on this new categorization, we achieved more consistent results and were able to implement Borderline-SMOTE. Chawla *et al.* provided pseudocode for the Borderline-SMOTE algorithm which we coded using Python. Once SMOTE and Borderline-SMOTE were coded, we tested our code on small, sample datasets. Figure 20 displays the visual differences.

Borderline-SMOTE created a tighter distribution of synthetic points and steered away from creating points around minority points that have the potential of being outliers. SMOTE seemed to create more synthetic minority samples in more sparse areas which could be an artifact from randomly choosing a nearest neighbor and generating the synthetic point without

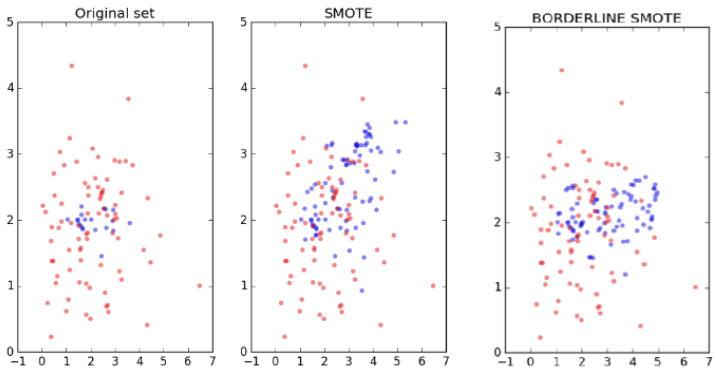

Figure 20: Difference in synthetic point generation between SMOTE and Borderline-SMOTE using $k = 5$. Adapted from [5].

consideration to the neighborhood of the chosen point. Just as we did for SMOTE, we also decided to run our code on the small, sample dataset using $k = 10$ for the KNN to see if the choice of $k = 5$ from [3] would be sufficient for our project. Figure 21 shows that the increase in k seemed to cause the distribution for Borderline-SMOTE to become more loose. This looseness could cause problems in the overlapping between classes. After visualizing the changes that varying the k value made in both Borderline-SMOTE and SMOTE, we felt most comfortable going with a k value that optimized Borderline-SMOTE since Borderline-SMOTE has been proven to be superior to SMOTE. Thus we moved forward using a k value of 5 just as [3] has suggested.

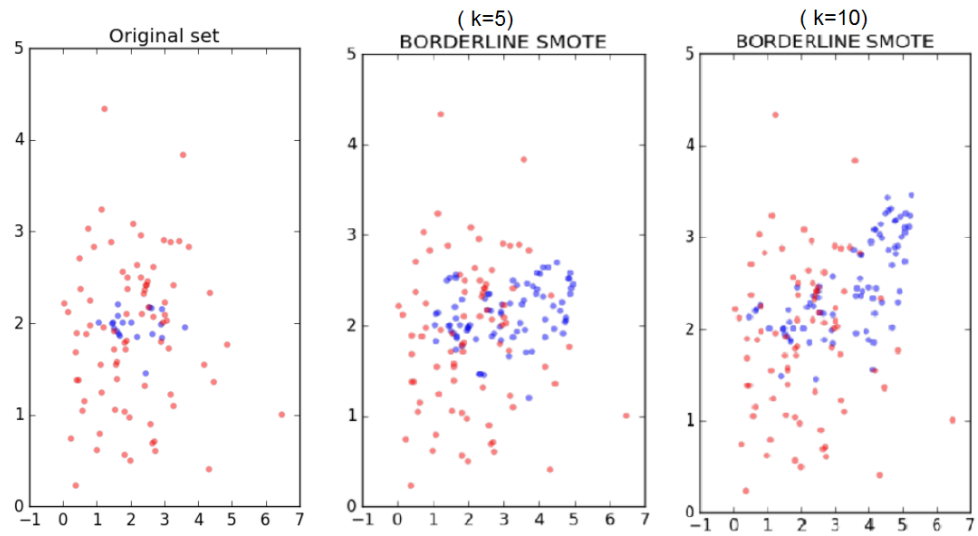

Figure 21: Difference in synthetic point generation using $k = 5$ and $k = 10$ in the Borderline-SMOTE algorithm. Adapted from [5].

6 Results

6.1 Overview and Parameters Used

To maintain consistency across different datasets, we standardized our code with permanently defined parameters. For all k-fold calculations, we used a standard of 10 folds, meaning at the end we had to divide our accuracy by 10 to get the average accuracy per fold. For both SMOTE and Borderline-SMOTE, we used a standard of $k = 5$ for the K-nearest Neighbors calculations and created synthetic points for all categories that were non-majority. To find the most important predictors, we used random forests containing 50 trees. After making all 50 trees, each by using 10 folds in k-folds, we divided our accuracy by 500 to get the appropriate average accuracy score.

When identifying the most important predictors, we ran this forward subselection 10 times per synthetic point algorithm performing each subselection on the exact same set of data and synthetic points. We noticed that when we reran the forward subselection process for both SMOTE and Borderline-SMOTE, the order of important predictors changed often (as well as the accuracy). This fluctuation is an artifact from the way the dataset

was split up into training and testing data before applying SMOTE and Borderline-SMOTE. To work with this issue and reveal meaningful results, we looked into how often the predictors showed up in the importance and where they arose in the ten runs. We also realized that the order of the important features used to make the trees did not matter since we use all of the predictors at once when creating a tree. Lastly, when producing confusion matrices for the new categorization, we observed the average confusion matrix over all runs, since these matrices seemed to be the most consistent. Similar to the important features, we ran into the issue of complete consistency with the confusion matrices as well due to the nature of the partitioning of the dataset.

In terms of making trees, we chose to only consider trees from Borderline-SMOTE runs, as this method has proved time and time again to outperform SMOTE. Each tree utilized attributes from the decision tree classifier function found in the scikit-learn Python package. This function used tree pruning parameters such as minimizing the number of samples per leaf and setting the maximum number of nodes per leaf. Lastly, cross-validation was used to determine the tree parameters for any given tree.

Each section has three confusion matrices. The first matrix is the aforementioned average confusion matrix. The next matrix is a comparison of two summed confusion matrices. This comparison is between a confusion matrix run without any SMOTE or Borderline SMOTE and a confusion matrix run with either SMOTE or Borderline SMOTE. We choose to do this comparison to see how well our algorithm did at predicting only the original points and not the "fake" points. All these confusion matrices were run using the same random seed so that the non-SMOTE and SMOTE runs would use the same training and testing data. A notable observation about this comparison is that the confusion matrices do not seem to change. We attributed this to the fact that our testing datasets were only the original points so that set is much smaller than the set with the SMOTE points. This means that our tree is not receiving enough training points to train our algorithm so when we tested the tree was not as accurate as it could be. We discuss in Section 8.2 that is it possible to fix this issue by using SMOTE only on the training points and not all points so that the tree can be trained on as many points as possible. However, because of time constraints we were unable to try this ourselves. Although it appears that our algorithm did not work, it does work in the respect that if one wants to use our trees on people not the class of 2015, it should predict their primary status pretty accurately.

6.2 Full Dataset Results

6.2.1 Four Category Categorization

The first scheme we tried for categorizing "Primary Status" left us with four categories: Full Time, Grad School, Not Seeking/Serving and Unknown/Seeking. As we discovered in Section 5.3.3, this categorization did not lend well to the Borderline-SMOTE algorithm because we had so few data points in one of the categories. Focusing solely on SMOTE, using this categorization, and using all predictors, our model predicted a given student's "Primary Status" with an average accuracy of approximately 68%. Our model also showed the most important categories needed to predict "Primary Status" via Forward Subset Selection. These categories, as well as the average accuracy of the prediction taken after the category was added, is displayed in Table 9 below.

Category Added	**Accuracy Post-Category Addition**
Degree Date	45.3779 %
Career Fair	51.4696 %
Workshop	58.1217 %
Summer Internship	60.0670 %
Walk-In	62.2717 %
Degree	62.6899 %
Science All	64.5445 %

Table 9: Category importance found by performing Forward Subset Selection on the Full Class of 2015 dataset where SMOTE was applied and with the four category categorization.

6.2.2 Final Categorization

From Section 5.3.3, we introduced the final version of the categorization of "Primary Status". Since both SMOTE and Borderline-SMOTE worked on this categorization, we have results for each of the algorithms. In terms of accuracy, Borderline-SMOTE outperformed SMOTE by having an average accuracy of 73-76% while SMOTE had an accuracy of 63-65%. As described in Section 6.1 , the order of important features varied from run to run. Table 10 shows the top five most important features for each of the ten runs of the Forward Subset Selection for SMOTE.

Run	Most Important	2nd Most	3rd Most	4th Most	5th Most
1	Degree Date	ECE	Business Eng.	US/ International	Summer Internship
2	Degree Date	ECE	Business Eng.	Company Info Session	Summer Internship
3	Degree Date	ECE	Business Eng.	Company Info Session	Summer Internship
4	Degree Date	ECE	Business Eng.	Company Info Session	Summer Internship
5	Degree Date	ECE	Business Eng.	Company Info Session	Summer Internship
6	Degree Date	ECE	Business Eng.	Company Info Session	Summer Internship
7	Degree Date	ECE	Business Eng.	US/ International	Summer Internship
8	Degree Date	Career Fair	ECE	Summer Internship	Business Eng.
9	Degree Date	ECE	Business Eng.	US/ International	Summer Internship
10	Degree Date	ECE	Business Eng.	US/ International	Summer Internship

Table 10: Top Five Predictors (in order of importance) for each of the Ten SMOTE Runs using the Class of 2015 dataset.

We notice that Degree Date is the most important category in every run. We can see that whether a student has an ECE major proved to be the second most important feature 9/10 times and on the 1/10 time that it wasn't the second most important, ECE was the third most important. We continued this path of analysis and also kept in mind that when we used the most important features to create our tree (model), the order of the predictors doesn't matter to the tree. Overall, Table 10 shows us that Degree Date, ECE, Business Engineering and Summer Internship appeared in the top five most important predictors for all of the ten runs. Since we are

looking at making a statement about the top five most important feature, we needed five features so we also chose the next most prominent predictor, Company Information Session (in the top five for 5/10 runs) to round out the top five most important features of SMOTE. The accuracy scores in Table 11 are just an example of the accuracy scores we got from testing and will differ from run to run. Table 12 gives an example of an average confusion matrix of our model using only the five important predictors. Again, this average confusion matrix will vary from run to run. Table 13 shows confusion matrices summed over all 10 folds for the original data points when using no SMOTE and SMOTE in conjunction with the top predictors.

Category Added	**Accuracy Post-Category Addition**
Degree Date	54.684 %
ECE	58.761 %
Business Eng.	60.420 %
Company Info Session	61.253 %
Summer Internship	61.914 %

Table 11: Example Run Results of Category importance for SMOTE found by Forward Subset Selection using the Class of 2015 dataset.

True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	96.0	4.1	7.0
Grad School	22.6	78.4	6.1
Unknown/ Not Paid	32.2	52.8	22.1

Table 12: Example of a Class of 2015 Confusion Matrix using SMOTE and all of the "Important Features" as predictors.

Table 13 shows the comparison of the confusions matrices of no SMOTE Applied and SMOTE Applied. Both confusion matrices show the same numbers in the cells. This means that SMOTE did not affect the orginal points and their predictions.

	No SMOTE Applied			SMOTE Applied		
True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	1065	3	3	1065	3	3
Grad School	226	3	0	226	3	0
Unknown/ Not Paid	360	0	0	360	0	0

Table 13: Comparing the confusion matrices (summing over all 10 folds for the original data points) when using no SMOTE versus SMOTE. Both runs used the same top predictors using the Class of 2015 dataset.

Table 14 shows the top four most important features in each of the ten runs of the Forward Subset Selection for Borderline SMOTE. Whether or not a student attended a CDC Walk-In event is the most important category in every run and the degree date of the student is always the second most important feature. CDC Career Fair Event attendance was third most important, occuring only 6/10 times.Career Fair is in the top four most important features for 10/10 runs. Our goal then is to make a statement about the top four most important features, so we need a fourth top feature so we also chose the next most prominent predictor, Degree Type (BA, MA, etc.) (in the top five for 5/10 runs).

Run	**Most Important**	**2nd Most**	**3rd Most**	**4th Most**
1	Walk-In	Degree Date	Career Fair	Science Degree
2	Walk-In	Degree Date	Career Fair	Workshop
3	Walk-In	Degree Date	Workshop	Career Fair
4	Walk-In	Degree Date	Career Fair	Resumazing
5	Walk-In	Degree Date	Career Fair	Degree Type
6	Walk-In	Degree Date	Career Fair	Degree Type
7	Walk-In	Degree Date	Degree Type	Career Fair
8	Walk-In	Degree Date	Career Fair	Resumazing
9	Walk-In	Degree Date	Workshop	Career Fair
10	Walk-In	Degree Date	Degree Type	Career Fair

Table 14: Top Four Predictors (in order of importance) for each of the Ten Borderline-SMOTE Runs using the Class of 2015 dataset.

The accuracy scores in Table 15 are just an example of the accuracy

scores we got from testing and will differ from run to run. Table 16 gives an example of an average confusion matrix of our model using only the four most important predictors. Again, this average confusion matrix will vary from run to run. Table 17 shows confusion matrices summed over all 10 folds for the original data points when using no SMOTE and SMOTE in conjunction with the top predictors.

Category Added	**Accuracy Post-Category Addition**
Walk-In	58.550 %
Degree Date	68.891 %
Career Fair	71.328 %
Degree Type	72.934 %

Table 15: Example Run Results of Category importance for Borderline-SMOTE found by Forward Subset Selection using the Class of 2015 dataset.

True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	96.5	5.3	5.3
Grad School	16.4	82.0	8.7
Unknown/ Not Paid	32.9	17.4	56.8

Table 16: Example of a Class of 2015 dataset Confusion Matrix using Borderline-SMOTE and all of the "Important Features" as predictors.

Table 17 shows the comparison of the confusions matrices of no Borderline-SMOTE Applied and Borderline-SMOTE Applied. The Borderline-SMOTE applied matrix shows that some of the points are classified differently when Borderline is used. For example, five more points were misclassified as Grad School when they were actually Full Time. This means that borderline SMOTE placed too many Grad School points in that area of Full Time points that those points got misclassified. This table also shows that the number of Grad School points classified correctly did not change and 2 more of the Unknown/Not Paid points got misclassified after Borderline. From this we can conclude that while Borderline classified points differently it did not classify them correctly.

	No Borderline-SMOTE Applied			Borderline-SMOTE Applied		
True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	991	50	30	988	55	28
Grad School	164	60	5	163	60	6
Unknown/ Not Paid	311	34	15	312	35	13

Table 17: Comparing the confusion matrices (summing over all 10 folds for the original data points) when using no Borderline-SMOTE versus Borderline-SMOTE. Both runs used the same top predictors using the Class of 2015 dataset.

The final step is to make a tree using the four predictors mentioned above from Borderline SMOTE. From cross-validation, we determined that that having a maximum node leaf number of 15 and minimum samples number of 50, gave the best original to pruned tree accuracy. The original accuracy of the unpruned tree was 78.77% and the accuracy of the pruned tree was 72.89%. The accuracy increased from original to pruned tree which means that the pruned tree's predictions are just as accurate as the originals. In the tree shown below, X[0] is a student's degree type, X[1] is a student's degree date, X[2] is the number of Career Fair Events a student attended and X[3] is the number of CDC Walk-Ins that a student has gone to.

Figure 34 in Appendix A above shows the decisions made by the tree when making predictors for the three classes. An example of the decisions goes as followed, if a student went to less than 5 Walk-Ins, received their degree in the spring, attended less than 2 Career Fairs and received a degree of Masters or PhD, they are likely to have a primary status of Full Time Job. Figure 22 shows this pathway from the tree.

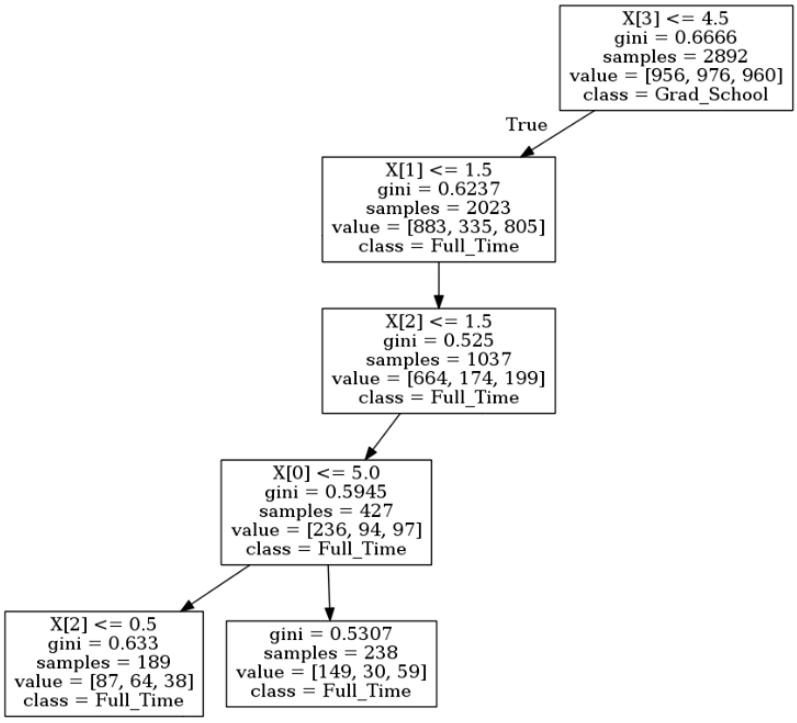

Figure 22: Zoomed in pathway from the Class of 2015 dataset Tree found in Appendix A

6.3 Internship Dataset Results

In this section, we looked into the Internship Dataset, this set included only those students who had reported having at least one internship during their time at WPI. SMOTE and Borderline-SMOTE worked for the categorization from 5.3.3, so we have results for both of these algorithms. SMOTE had an average accuracy of 76% - 80% and Borderline SMOTE had an average accuracy of 76% - 80%. Notice that SMOTE and Borderline SMOTE have around the same accuracy for this dataset which is untrue for the other two sets. However, since we look at Borderline SMOTE for the trees for the other two sets, we will also be look at Borderline for the intern dataset trees as well. As described in Section 6.1, the order of important features varied from run to run. Table 18 shows the top five most important features for each of the ten runs of Forward Subset Selection for SMOTE.

Run	Most Important	2nd Most	3rd Most	4th Most	5th Most
1	Degree Date	Career Fair	Walk-In	Workshop	Science Degree
2	Degree Date	Resumazing	Company Info Session	Nationality	Workshop
3	Degree Date	Workshop	Career Fair	Nationality	Science Degree
4	Degree Date	Company Info Session	Nationality	Business Engineer-ing	ECE
5	Degree Date	Resumazing	Gender	Science Degree	Company Info Session
6	Degree Date	Company Info Session	Career Fair	Workshop	Degree Type
7	Degree Date	Company Info Session	Walk-In	Career Fair	Workshop
8	Degree Date	Resumazing	Science Degree	Workshop	Nationality
9	Degree Date	Career Fair	Workshop	Nationality	Science Degree
10	Degree Date	Company Info Session	Nationality	Workshop	Career Fair

Table 18: Top Five Predictors (in order of importance) for each of the Ten SMOTE Runs using the Internship dataset.

This dataset most important predictors tended to differ more than any other dataset so we had to choose the top five predictors because 3 of these predictors were in the top five 6/10 times. We notice that Degree Date is the most predictive category in every run and that Workshop appears in the top five 8/10 times. Company Info Session, Career Fair and Nationality each appear 6/10 times making there be five predictors. From this table we can see that Degree Date, Career Fair, Company Info Session, Nationality and Workshop are the top five predictors. The accuracy scores in Table 27

are just an example of the accuracy scores we got from testing and will differ from run to run. Table 28 gives an example of an average confusion matrix of our model using only the four important predictors.

Category Added	**Accuracy Post-Category Addition**
Degree Date	62.55 %
Workshop	70.60 %
Career Fair	77.27 %
Nationality	81.06 %
Science Degree	81.99 %
Business Engineering	82.23 %
IMGD Degree	82.48 %

Table 19: Example Run Results of Category importance for SMOTE found by Forward Subset Selection using the Internship dataset.

True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	17.8	2.5	0.9
Grad School	2.7	16.4	2.1
Unknown/ Not Paid	1.3	4.2	15.7

Table 20: Example of a Internship dataset Confusion Matrix using SMOTE and all of the "Important Features" as predictors.

Table 21 shows the comparison of the confusions matrices of no SMOTE Applied and SMOTE Applied. The SMOTE applied matrix shows that some of the points are classified differently. This table shows that the number of grad school points classified correctly increased by seven points but 11 more points were misclassified as Unknown/Not Paid. Unknown/Not Paid classified one more point correctly without misclassifying anymore points than before. and 2 more of the Unknown/Not Paid points got misclassified after Borderline. Two Full Time points were misclassified as Unknown/Not Paid and two points that were misclassified as Grad School were now misclassified as Unknown/Not Paid. From this we can conclude that SMOTE did a slightly better job at classifying the internship data but points are still being greatly misclassified.

	No SMOTE Applied			SMOTE Applied		
True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	183	22	7	181	20	11
Grad School	32	17	0	23	24	11
Unknown/ Not Paid	9	3	1	9	2	2

Table 21: Comparing the confusion matrices (summing over all 10 folds for the original data points) when using no SMOTE versus SMOTE. Both runs used the same top predictors using the Internship dataset.

Table 22 shows the top four most predictive features in each of the ten runs of the Forward Subset Selection for SMOTE. This set of predictors was chosen because they appeared in the top five the most. The reason the top five predictors was not chosen was because when running the model repeatedly the accuracy very a lot so using the top gave a more stable model. Degree Date appears in the most important spot for all 10 runs. Resumazing appears in the top four 7/10 and Nationality appears in the top four 8/10 times. Workshop and Career Fair only appeared in the top four 5/10 times and 4/10 times respectively. We felt we needed at least a top 4 features so we looked at the 5th most important and saw that Workshop appeared in the top five 8/10 times where Career Fair only appeared 6/10 times so we choose Workshop for the final feature. From this table we can see that Degree Date, Resumazing, Nationality and Workshop are the top four predictors.

Run	Most Important	2nd Most	3rd Most	4th Most	5th Most
1	Degree Date	Resumazing	Workshop	Nationality	Career Fair
2	Degree Date	Walk-In	Career Fair	Workshop	Degree Type
3	Degree Date	Resumazing	Workshop	Nationality	Company Info Session
4	Degree Date	Resumazing	Nationality	Degree Type	Career Fair
5	Degree Date	Resumazing	Workshop	Nationality	Company Info Session
6	Degree Daten	Resumazing	Summer Internship	Nationality	Workshop
7	Degree Date	Nationality	Company Info Session	Resumazing	Workshop
8	Degree Date	Walk-In	Career Fair	Nationality	Degree Type
9	Degree Date	Nationality	Resumazing	Career Fair	Workshop
10	Degree Date	Walk-In	Workshop	Career Fair	Degree Type

Table 22: Top Four Predictors (in order of importance) for each of the Ten Borderline-SMOTE Runs using the Internship dataset.

The accuracy scores in Table 23 are just an example of the accuracy scores we got from testing and will differ from run to run. Table 24 gives an example of an average confusion matrix of our model using only the four most important predictors.

Category Added	Accuracy Post-Category Addition
Degree Date	62.58 %
Resumazing	74.41 %
Nationality	76.69 %
Summer Internship	80.05 %
Degree Type	82.77 %
Science Degree	83.90 %
Workshop	84.90 %
Co-op	85.39 %
IMGD Degree	85.60 %

Table 23: Example Run Results of Category importance for Borderline-SMOTE found by Forward Subset Selection using the Internship dataset.

True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	20	0.8	0.4
Grad School	4.5	14.5	2.2
Unknown/ Not Paid	1.6	3.6	16

Table 24: Example of a Internship dataset Confusion Matrix using Borderline-SMOTE and all of the "Important Features" as predictors.

Table 25 shows the comparison of the confusions matrices of no Borderline-SMOTE Applied and Borderline-SMOTE Applied. The Borderline-SMOTE applied matrix shows that some of the points are classified differently. This table shows that the most the points were classified the same and some misclassifications increased by one or two. The only correct classification that changed was Unknown/Not Paid which increased its correct classification by two. From this we can conclude that Borderline-SMOTE did not make much of an affect on the classifying the points better but as was stated before Borderline and SMOTE work very similarly with the intern dataset so it is not surprsing that SMOTE worked better in this case.

	No Borderline-SMOTE Applied			Borderline-SMOTE Applied		
True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	201	7	4	199	8	5
Grad School	48	1	0	47	1	1
Unknown/ Not Paid	13	0	0	11	0	2

Table 25: Comparing the confusion matrices (summing over all 10 folds for the original data points) when using no Borderline-SMOTE versus Borderline-SMOTE. Both runs used the same top predictors using the Internship dataset.

The final step is to make a tree using the four predictors mentioned above from Borderline SMOTE. From cross-validation, we determined that that having a maximum node leaf number of 15 and minimum samples number of 20, gave the best original to pruned tree accuracy. The original accuracy of the unpruned tree was 79.39% and the accuracy of the pruned tree was 79.39%. The accuracy is the same from original to pruned tree so this means that the pruned tree's predictions are just as accurate as the original's. In the tree shown below, X[0] is a whether a student is International or American, X[1] is a student's degree date, X[2] is the number of Resumazing events that a student attended and X[3] is the number of CDC Workshops that a student attended.

Figure 35 in Appendix B above shows the decisions made by the tree when making predictors for the three classes. An example of the desions goes as followed, if a student recieved their degree in the fall or winter, the student was American, attended no Resumazing events and attended at least one workshop, they are likely to have a primary status of Grad School. Figure 23 shows this pathway from the tree.

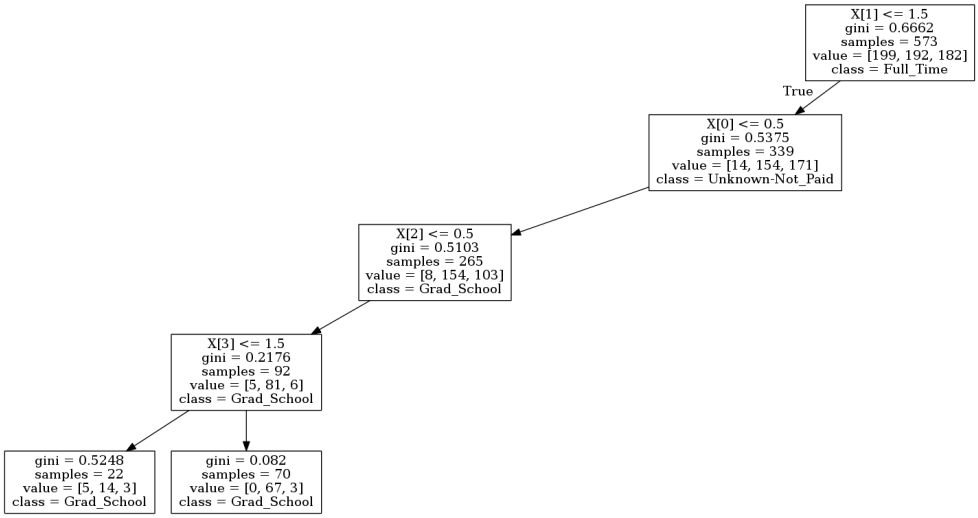

Figure 23: Zoomed in pathway from the Internship dataset Tree found in Appendix B

6.4 CDC Dataset Results

In this section we looked into the CDC Dataset, this set included only those students who had attended a CDC sponsored event at least once in their time at WPI. SMOTE and Borderline SMOTE worked for the categorization from 5.3.3, so we have results for both of these algorithms. SMOTE had an average accuracy of 61% - 63% and Borderline SMOTE had an average accuracy of 75% - 77%. As described in Section 6.1 , the order of important features varied from run to run. Table 26 shows the top four most important features for each of the ten runs of Forward Subset Selection for SMOTE.

Run	Most Important	2nd Most	3rd Most	4th Most
1	Degree Date	Summer Internship	Science Degree	Company Info Session
2	Degree Date	ECE	Company Info Session	Civil Engineering
3	Degree Date	Workshop	Company Info Session	Summer Internship
4	Degree Date	Summer Internship	Workshop	Pre-Career Fair
5	Degree Date	Summer Internship	Science Degree	Workshop
6	Degree Date	Workshop	Liberal Arts Degree	Company Info Session
7	Degree Date	Workshop	Company Info Session	Liberal Arts Degree
8	Degree Date	Summer Internship	Science Degree	Pre-Career Fair
9	Degree Date	Workshop	Mathematics Degree	Summer Internship
10	Degree Date	Workshop	Company Info Session	Summer Internship

Table 26: Top Four Predictors (in order of importance) for each of the Ten SMOTE Runs using the CDC Usage dataset.

We notice that Degree Date is the most predictive category in every run. We then notice that the next three features seem to change around a lot. However, since when we run a tree to get a final accuracy that order does not matter we can look at how often a predictor comes up in the top four. For Instance, Summer Internship appears in 2nd most important 4/10 times but overall it appears in the top four 7/10 times. A similar instance happens with Workshop and Company Info Session, Workshop appears in the top four 7/10 and Company Info Session appears 6/10. In this chart, there a few pink boxes however those indicate predictors that appeared randomly in that spot. So from this analysis, we can say that Degree Date, Workshop, Summer Internship and Company Info Session appeared in the top four categories the most. The accuracy scores in Table 27 are just an example of the accuracy scores we got from testing and will differ from run to run. Table 28 gives an example of an average confusion matrix of our model using

only the four important predictors.

Category Added	Accuracy Post-Category Addition
Degree Date	58.32 %
ECE	61.03 %
Company Info Session	62.84 %
Civil Engineering	63.82 %
Materials Degree	64.38 %
Business Engineering	64.38 %
Mathematics Degree	65.59 %
Summer Internship	65.59 %
Co-op	65.72 %
Liberal Arts Degree	65.77 %

Table 27: Example Run Results of Category importance for SMOTE found by Forward Subset Selection using the CDC Usage dataset.

True/ Predicted	Full Time	Grad School	Unknown/ Not Paid
Full Time	63.1	1.7	3.1
Grad School	13.6	39.7	14.6
Unknown/ Not Paid	17.2	26.7	24

Table 28: Example of a CDC Usage dataset Confusion Matrix using SMOTE and all of the "Important Features" as predictors.

Table 29 shows the comparison of the confusions matrices of no SMOTE Applied and SMOTE Applied. The SMOTE applied matrix shows that some of the points are classified differently. This table shows that more Full Time points were misclassified as Grad School and Unknown/Not Paid but although many Grad School and Unknown/Not Paid points were still misclassified as Full Time, the number of points correctly classified increased for both classes. From this we can conclude that SMOTE did help a little bit in classifying points correctly. It however did not help classify Full Time better because more points from other classes were placed next to Full Time points making it seem as though the Full Time points were from another class.

	No SMOTE Applied			SMOTE Applied		
True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	656	7	16	618	21	40
Grad School	133	0	2	128	6	1
Unknown/ Not Paid	174	1	1	152	3	21

Table 29: Comparing the confusion matrices (summing over all 10 folds for the original data points) when using no SMOTE versus SMOTE. Both runs used the same top predictors using the CDC Usage dataset.

Table 30 shows the top six most predictive features in each of the ten runs of the Forward Subset Selection for SMOTE. This set of data has six most predictive features for Borderline because, as shown by the table, the top 6 predictors appear almost always in the top 6. Degree Date and Walk-In appear in the first two spots respectivley for all 10 runs. Degree Type appears in the top six 9/10 times, Company Info Session appears 10/10 times, Summer Internship appears 8/10 and Business Engineering appears 8/10 times. This table shows that Degree Date, Walk-In, Degree Type, Company Info Session, Summer Internship and Business Engineering are the top six predictors that should be used when doing borderline SMOTE.

Run	Most Important	2nd Most	3rd Most	4th Most	5th Most	6th Most
1	Degree Date	Walk-In	Degree Type	Company Info Session	Summer Internship	Business Engineering
2	Degree Date	Walk-In	Degree Type	Company Info Session	Summer Internship	Business Engineering
3	Degree Date	Walk-In	Degree Type	Business Engineering	Summer Internship	Company Info Session
4	Degree Date	Walk-In	Degree Type	Company Info Session	Summer Internship	Math Degree
5	Degree Date	Walk-In	Degree Type	Civil Engineering	Company Info Session	Business Engineering
6	Degree Daten	Walk-In	Workshop	Company Info Session	Civil Engineering	Science Degree
7	Degree Date	Walk-In	Summer Internship	Company Info Session	Degree Type	Business Engineering
8	Degree Date	Walk-In	Company Info Session	Summer Internship	Degree Type	Business Engineering
9	Degree Date	Walk-In	Summer Internship	Company Info Session	Degree Type	Business Engineering
10	Degree Date	Walk-In	Degree Type	Company Info Session	Summer Internship	Business Engineering

Table 30: Top Four Predictors (in order of importance) for each of the Ten Borderline-SMOTE Runs using the CDC Usage dataset.

The accuracy scores in Table 31 are just an example of the accuracy scores we got from testing and will differ from run to run. Table 32 gives an example of an average confusion matrix of our model using only the four

most important predictors.

Category Added	Accuracy Post-Category Addition
Degree Date	58.41 %
Walk-In	72.60 %
Degree Type	74.22 %
Business Engineering	74.90 %
Summer Internship	75.23 %
Company Info Session	75.69 %
Civil Engineering	76.04 %
Mathematics Degree	76.05 %

Table 31: Example Run Results of Category importance for Borderline-SMOTE found by Forward Subset Selection using the CDC Usage dataset.

True/ Predicted	Full Time	Grad School	Unknown/ Not Paid
Full Time	63.1	2.4	2.4
Grad School	13.1	49.8	5
Unknown/ Not Paid	16.4	8.2	43.3

Table 32: Example of a CDC Usage dataset Confusion Matrix using Borderline-SMOTE and all of the "Important Features" as predictors.

Table 33 shows the comparison of the confusions matrices of no Borderline-SMOTE Applied and Borderline-SMOTE Applied. The Borderline-SMOTE applied matrix shows that some of the points are classified differently. This table shows that only a few more Full Time points were misclassified as Unknown/Not Paid. However, this table shows greatest improvement in correct classification of Grad School and Unknown/Not Paid. Grad School had an increase in 10 points correctly classified and only two more points misclassified as Unknown/Not Paid and Unknown/Not Paid had an increase in eight points correctly classified but six more points were misclassified as Grad School. From this we can conclude that Borderline-SMOTE helped a great deal with better classifying some of the CDC dataset points and that this dataset had the greatest improve among all the SMOTE and Borderline confusion matrix comparisons across all three datasets.

	No SMOTE Applied			SMOTE Applied		
True/ Predicted	**Full Time**	**Grad School**	**Unknown/ Not Paid**	**Full Time**	**Grad School**	**Unknown/ Not Paid**
Full Time	639	18	22	636	18	25
Grad School	131	2	2	119	12	4
Unknown/ Not Paid	165	3	8	151	9	16

Table 33: Comparing the confusion matrices (summing over all 10 folds for the original data points) when using no Borderline-SMOTE versus Borderline-SMOTE. Both runs used the same top predictors using the CDC Usage dataset.

The final step is to make a tree using the six predictors mentioned above from Borderline SMOTE. From cross-validation, we determined that that having a maximum node leaf number of 15 and minimum samples number of 25, gave the best original to pruned tree accuracy. The original accuracy of the unpruned tree was 76.09% and the accuracy of the pruned tree was 75.11%. The accuracy decreased from original to pruned tree but only by 1% so this means that the pruned tree's predictions are almost as accurate as the original's. In the tree shown below, X[0] is a student's degree type, X[1] is a student's degree date, X[2] is the number of Company Info Sessions a student attended and X[3] is the number of CDC Walk-Ins that a student has gone to, X[4] is how many summer internships a student had and X[5] is whether or not a student had Business Engineering as one of their majors.

Figure 36 in Appendix C shows the decisions made by the tree when making predictors for the three classes. An example of the decisions goes as followed, if a student recieved their degree in the fall or winter, went to more than 5 Walk-Ins at the CDC, attended at least one company info session and attended at least 3 company info sessions, they are likely to have a primary status of Grad School. Figure 24 shows this pathway from the tree.

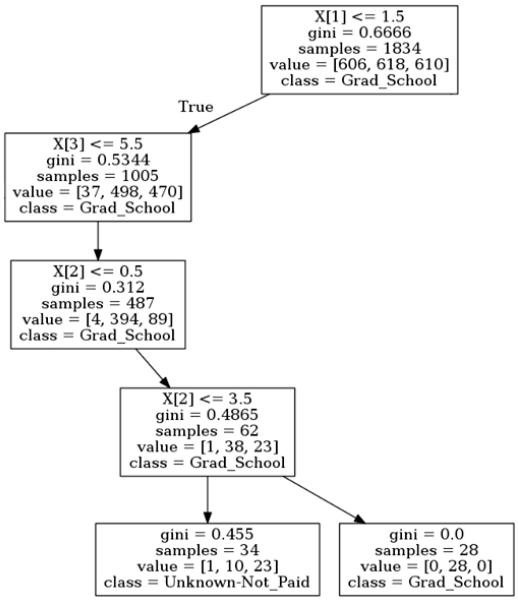

Figure 24: Zoomed in pathway from the CDC Usage dataset Tree found in Appendix C

7 Discussion

7.1 Algorithm Accuracy

Borderline-SMOTE outperformed SMOTE on the final version of our categorization of Primary Status for the first dataset of Final Class of 2015 data. The accuracy of Borderline-SMOTE averaged to about 73-76%, whereas regular SMOTE had a lower accuracy of 63-65%. Accuracy rates were similar for the second dataset of internship (76-80%) for both SMOTE and Borderline-SMOTE. As with the first case, in the final dataset for CDC usage, Borderline-SMOTE again outperformed SMOTE by an even greater margin. For this final dataset, Borderline-SMOTE had an average accuracy of 75-77% and SMOTE had an accuracy of only 61-63%.

It is clear that Borderline-SMOTE with our modifications for low-populated, multi-class predictors is as accurate or more accurate than the original SMOTE algorithm for selecting relevant features and making predictions that overcome data imbalance. This is significant because it is not likely that future versions of these datasets are likely to lack the imbalance inherent in the 2015 selections. With regards to primary status, for instance, the Back-

ground Chapter details how past reports have consistently demonstrated positive outcomes for more than 90% of graduates. This is an achievement on the part of both the university and the CDC as well. This method will be crucial for future studies dealing data of this type and for synthesizing results across datasets from different years.

7.2 Predictors of Success

From the Borderline-SMOTE runs on the first dataset, Walk-In was the most important predictor for each of the ten runs, followed by Degree Date and then Career Fair. It is important to note what this feature selection means relative to CDC operations. To say that Walk-Ins are the most important predictor means that it is the first decision-point in the tree used to reach a conclusion about the response. It is not entirely accurate to say, for example, that because Walk-In was the top predictor for the first dataset's ten runs, that using Walk-Ins at the CDC will result in a student getting a full time job. More attention is needed to see how the decision pathways describe likelihoods based on whether certain conditions of that predictor are met.

Using the four important predictors from these runs (Walk-In, Degree Date, Career Fair, and Degree Type), our decision tree shows some outcomes worth considering. For students who attended more than 5 Walk-Ins during their undergraduate careers and graduated in the spring, for instance, the tree predicts full time employment.However, if the student did not graduate in the Spring, then the model predicts graduate school attendance which is not necessarily correct for every given student as our model is not 100% accurate.This suggests that of the class of 2015 alumni, early graduates who used more Walk-Ins at the CDC went on to graduate school, while those who graduated later and were proactive with Walk-In use were more likely to be predicted as receiving full time jobs. This makes sense if one considers the idea that early graduates may be doing a five-year Master's program in which they might get their undergraduate credit early. Or if students are not enrolled in such a program, they might want to get their credits done earlier to allow them more time to pursue a higher-level degree in a shorter period of time.

We can compare this to another decision pathway, too. If a student attended less than 5 Walk-Ins and graduated in the spring, they almost invariably are predicted to receive a full time job regardless of any other condition. The only pathway that results in a different response is if the same student attends less than 2 Career Fairs and has a Masters or PhD, then they are predicted as being in graduate school.

The results from the other two datasets were less clean-cut. For Internships, Degree Date and Resumazing were the top predictors, followed by Nationality and Summer Internship. But these latter two categories and all subsequent ones only added marginal (5% or less) accuracy to the model. Degree Date and Walk-In were the top predictors in the CDC Usage dataset, and similarly, all subsequent categories including the next top predictors (Degree Type and Business Engineering) added only marginal (5% or less) accuracy. The consistency of Degree Date as a top predictor in all three datasets suggests that it is one of the strongest overall predictors of outcomes, which is in line with an assumption about degree levels corresponding to a greater likelihood of getting hired, as well as a demonstration of individual drive to succeed.

The fact that Resumazing, Walk-Ins, Career Fairs, and so on are also top predictors should come with a word of elaboration. That is, there is a need for student motivation to attend these sessions. It is not entirely clear whether it is the events themselves or if it is the fact that a student makes the decision to attend them and be proactive, or not, which ultimately decides outcomes. Our analysis provides highly-accurate categorical predictors of career outcomes, but what it cannot speak in depth to is how much these results are confounded by other factors not included in our model. The primary ones to consider are student drive and academic success.

8 Limitations and Future Work

Our study focused on analyzing outcomes relative to postgraduate career success as well as other opportunities such as graduate school and volunteer work. We worked directly with information provided to us from the CDC, used established methods in machine learning, and adapted algorithms from scikit learn and SMOTE. There were inherent limitations given our setup and goals, as well as assumptions we had to make which might need reconsideration. An overview of some of these limitations and possible assumptions worth reevaluation are offered in order to inform not only the CDC, but also researchers conducting future studies.

8.1 Limitations

The limiting factors we were competing against during this project were time, available data, computing power, and an imbalanced dataset. The time constraint is a given for any project, and forced us to make certain decisions toward the end of the project about which unfinished goals were

going to be prioritized. For instance, an investigation into the relationship between special request CDC Workshops and future use by students, as well as the determination of an optimally beneficial frequency of use, were not pursued because they would require pre-processing of a new dataset as well as additional analysis which would extend past the end of the term. This example is described in more detail in the Future Work section, and might be pursued by other teams in ensuing years.

Our available datasets covered a wide array that was satisfactory for the CDCs primary interests and for us to make robust predictions about career opportunities based on kiosk usage, major, internship history, and other categories. But this does not necessarily depict a full picture of undergraduate conditions for positive outcomes. Whatever our data says about the accuracy of predictions based on features in the data we used, critical information which sheds light on other dimensions of the problem are omitted simply because they were not included in the datasets we had access to. Academic information is the main missing piece in our case. The only information tracked in this domain was major, which did not have a significant impact to begin with. Beyond this, student GPA, extracurricular activities, and other indicators of academic and social status were not available for us to integrate into our prediction models.

The lack of academically-focused information is outside of the scope of the CDC and therefore the project, since their ability to act on our recommendations is limited to their organization, staff, programs, and events. Nonetheless, a more complete profile of students lives during their four year undergraduate careers would be useful for making stronger cases for correlations between actions or the lack thereof at the undergraduate level and career opportunities out of college.

Computing power was a real limitation when we began to apply SMOTE, Borderline-SMOTE, and our modified algorithm to the large primary dataset. The Jetstream cloud environment using the supercomputer was helpful, but runtimes were still incredibly slow for running the necessary code. Additionally, since we were using a remote supercomputer and not one on a local machine, there were further difficulties when problems arose since they could not always be solved immediately. It would have been better to have this ability natively, and to have had access to the supercomputer earlier in the project than we did.

As described extensively above, the major issue inherent to the data itself was that it was highly imbalanced. This was addressed by using a modified SMOTE algorithm, but remains a difficult problem in general. Unlike almost all the other limiting factors, imbalanced data is not something that

can be resolved simply by extending time or increasing resources. It still undoubtedly limited our ability to quickly and accurately make predictions using basic methods. A primary obstacle in this final category of limitations was the fact that SMOTE by its nature does not handle multi-class data. Chawla *et al.* notes that two separate versions of SMOTE, SMOTE-N and SMOTE-NC, are theoretically capable of handling nominal data whereas the original version was only meant to handle a mixture of nominal and continuous data [3]. For example, in the example confusion matrix of Table 12 we see that "Grad School was wrongly predicted as "Unknown/Not Paid 52.8 times. This was a starting point for us to consider working outside the normal parameters of SMOTE. Namely, the idea was to calibrate Borderline-SMOTE to include information about the second most likely misclassification in the scoring of the potential, new synthetic point. Figure 25 describes the process that we hope will be looked into in the future.

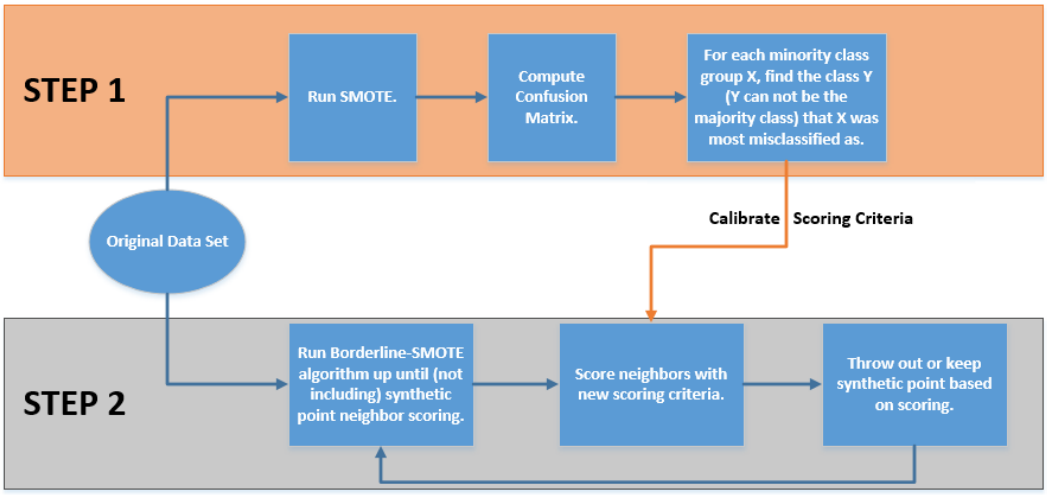

Figure 25: Proposed outline of using confusion matrices to better calibrate SMOTE/Borderline-SMOTE to avoid minority group misclassifications.

At the end of the project we noticed that the original points were not being classified any more correctly then they were in the beginning. This is due to the fact that when we tested only the original points we had a smaller set of points then the points we had for training so we could only use a subset of points to train the model. In order to remedy this problem, we could do SMOTE or Borderline-SMOTE on only the training data instead of the who dataset and then make a tree using those points. The test data would not

have seen SMOTE and would be used to make predictions. This method would hopefully make the original points be correctly classified more but since there was a time constraint we were unable to see if this would work. This method is shown in the figure below.

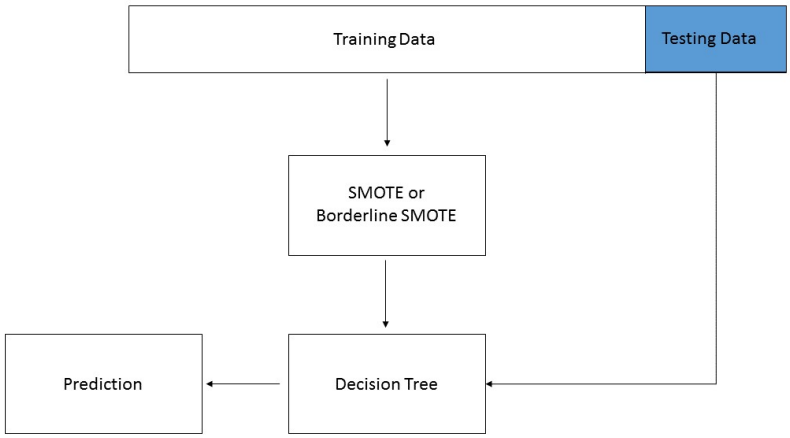

Figure 26: Method of using SMOTE and Borderline in the training data in hopes of classifying the orginal points better

8.2 Future Work

This project covered data from the class of 2015. Future studies could benefit from replication for future years, as well as conducting similarly rigorous analysis on past datasets such as that from the classes of 2013 or 2014. Comparative studies could then be conceived between different years to establish or confirm patterns of interest. Even larger combined datasets covering multiple class years would offer more insights than a single year can. At the very least, it is in the CDCs interest to continue doing this level of analysis on incoming datasets for the class of 2016 and beyond.

The CDC would benefit from further improving its survey-gathering techniques and in-house data cleanup. If they wish to use our process and algorithms for future data analysis, one of the most time-consuming and non-programmatic steps they will have to take will consist of data cleanup, transformation, and encoding. Making data capture more consistent from year to year and standardizing procedures would reduce the need for extensive data cleanup and transformation during pre-processing. Along with the CDCs own survey standards being altered, future studies could dedicate

some time to developing a more programmatic approach than what was used in this project. For instance, data cleanup was mostly done using a combination of filtering and basic formulas for transforming large quantities of column data in Excel and Google Sheets. Using regular expression operations, macro scripts, or Python, there are several ways one could conceivably begin to develop a system that relies less on hands-on data massaging.

Because our data showed that Walk-Ins were a top predictor for the Full Class of 2015 dataset, we recommend that future projects use the subcategory categorization scheme devised by the MQP team and modified with CDC input to run analysis on a dataset with these subcategories rather than simply "Walk-Ins." With the categories grouped as they are presented in an earlier section, there should be few enough each with enough data to generate a clean set of runs. Additionally, while the predictor was not as significant in any of the datasets as Walk-Ins or other predictors, Workshops were still significant, albeit marginally, in the Internship dataset and constitute a large percentage of CDC service capacity. We recommend that future teams consider incorporating the sub-categorization scheme found in 10.

Additional points of interest such as determining optimal CDC usage ranges for maximizing successful outcomes and integrating further academic and extracurricular information into the main dataset may yield more complete results beyond the framework of this project.

9 Conclusion

When we started out with this project our goal was to help the Career Development Center by identifying which factors help students get jobs after they graduate from WPI. In order to do that we had to come up with methods that would easily allow us to reach any conclusions. Due to imbalanced and messy data, we had to first find a way to balance it so that our conclusions would be accurate and usable. We were able to use an algorithm called SMOTE, and more specifically Borderline-SMOTE, to create data points that would balance out our data. We were able to modify the Borderline-SMOTE algorithm in order to account for low population and multi class predictors which the algorithm did not originally allow. The accuracy as described in previous sections was good enough so that we were confident with presenting our results to the Career Development Center.

Through this algorithm our team was able to produce some very intriguing results for the Career Development Center. We were able to tell them that one of the most important things a student can do to more accurately

predict their status after college is to have a Walk-In meeting with them at the CDC building on campus. Other CDC events such as Career Fairs and Workshops also showed a degree of importance in determining the student's primary status after graduation. After looking at the graduation class of 2015 as a whole we looked into smaller subsets to see if we could find any other interesting patterns. The muddiest smaller set involved students who had at least one internship during their college career. In this set, there were fewer patterns clear from the Borderline-SMOTE algorithm with date they received their degrees being the most important factor in determining their post-graduate primary statuses. We also looked at the set of students who attended at least one CDC sponsored event in their time at WPI. Due to the larger dataset these results were more clear than the internship data. Once again Walk-Ins had a big influence on primary status as well as summer internships and company information sessions. These are quite useful pieces for the CDC as they plan for the future.

These results have an immediate impact for the people at the Career Development Center. The first and most obvious impact is the fact that they can use what we have learned in order to better promote certain events in the future. This information also helps them understand what they should be focusing their resources on at the moment. Another less obvious impact is the possibilities for future projects. A lot of work went into developing the modified Borderline-SMOTE algorithm so that it could be used by future researchers working with the CDC when working with similar data. This could mean working with future data or even looking further into the past beyond just 2015. In many ways this is the biggest legacy of this project and the part that provides the most potential for future use by the Career Development Center.

Afterword: Data Science as an Emerging Profession

A Major Qualifying Project

by

Robert C. Vigeant

Submitted in partial fulfillment
of the requirements for the degree of:

Bachelor of Arts

Professional Writing
Worcester Polytechnic Institute
March 17, 2017

Approved by:

Professor Brenton Faber

10 Afterword

This Afterword was authored independently as a part of the Professional Writing component of the Major Qualifying Project. The following sections offer further considerations about data science and considers this particular project through the lens of rhetorical theory and its application in this case study of data science project design, implementation, and analysis. The Afterword is divided into two chapters. First, a chapter investigating data science as a proto-profession. And second, a chapter relating proto-professional rhetoric, as well as notions of heuristic and invention in data science, to the entire project. The first chapter consists of sections 11-15 in the Table of Contents. The second chapter consists of sections 16-20. The final two sections, 21 and 22, pertain to both chapters and conclude the Afterword as a whole.

Acknowledgements

Thank you to Professor Faber, who was instrumental in making this component of the MQP a reality. He helped guide the project, supplied research materials, and stimulated interesting and thought-provoking questions throughout the project. I can't thank him enough for his support.

Chapter 1: Proto-Professionalism

11 Introduction

With massive amounts of data being collected across industry and academia, data science has emerged from the traditional fields of mathematics and computer science as a powerful way to interpret this information. Data science employs statistics, machine learning, and graphical analysis to transform large sets of ambiguous data into more concrete and applicable results.

Data science largely emerged from within statistical circles. Influential statistician John W. Tukey wrote in 1962 about his growing interest in data analysis from his statistical background in The Future of Data Analysis [23]. Early applications of data science were largely concentrated within industry. Several early data science studies were published in business journals, targeting subjects such as supply chain management. However, it is only recently that data science has become an academically distinct force within the university setting. The discipline is threading a line between an occupational element and a profession in its own right, like the areas of mathematics and computer science from which it emerged.

This evolution should produce a shift in the rhetorical framework of writing within data science, specifically at the professional, academic level. If data science is an emerging profession, a 'proto-profession,' then this should be observable in the nature of significant texts within the area of study across journals and other publication sources in comparison to existing, related professions. This project will use grammatical analysis of existing data science articles to determine if there has been a significant change in the way that their authors have rhetorically constructed their field. Additionally, it will treat a data science Major Qualifying Project (MQP) at Worcester Polytechnic Institute (WPI) as a source of example and commentary on predominant rhetorical issues in data science practice and writing.

12 Background

There is an important distinction to make between occupational writing by an employee in a workplace and writing done by a professional in a professional environment. Within the private sector, workers are by definition forced to negotiate with their employer due to the fact that they are salaried [24]. Additionally, they neither have politically organized entities operating in their spaces nor do they control their spaces generally on account of the fact that "their projects, employment, and quality control are subject to capitalist oversight" [24].

Anne Beaufort provides a picture of how writing is shaped by the modern workplace [25]. For example, she cites a 1986 study that described how the hierarchical nature of an organization influenced decision-making about the style of the writing that employees produced [25]. In this study, conducted by Brown and Herndl, a connection was found between who the writing was directed toward and the use of nominalizations [25]. Specifically, the use of nominalizations and verbosity decreased the further down the hierarchy one was writing, and increased if the writing was targeted toward higher-ranking people in or out of the company [25]. This is of key interest to defining a distinction between the two domains of writing because it clearly illustrates how composition is dependent upon an environment over which the employee has limited to no control. Even the benefits and drawbacks of technological tools in the workplace [25], such as computers and nowadays tablets and smartphones sheds light on this. Apart from personal devices which are used, these tools are all provided for by the employer with the expectation that they will be used to complete tasks assigned to employees using them.

12.1 The Professions

As a contrast to occupational writing, the professions are characterized by three defining characteristics. First, professional writers have a well-defined and specific audience to which they write [24]. Although certainly a non-professional employee may write to specific people within the hierarchy of his or her company, the professional writes as an "audience advocate" who is not merely advancing or representing the company but rather writing for the well-being of this audience [24]. Second, a professional has a social responsibility to society as a whole. As such, a professional's work and writing doesn't solely serve a company and its interests, but rather "influences the life of a community" [24]. This is an extension of the idea that a professional is a source of objective knowledge. This "esoteric and complex" knowledge [24] is often a virtue to discover and disseminate by itself, especially for researchers. However, even more applied professionals like doctors or lawyers have a notion of doing good for society based on their own specialized knowledge. Finally, the professions are defined by an ethical awareness of their actions and a desire to maintain their credibility from this by excluding competition and easy access to their fields by outsiders [24].

13 Literature Review

The primary sources of literature on both occupational and professional writing have been covered in the above background section (Beaufort, Faber). Of greater interest is existing studies into how data science specifically has been evolving into a profession and how the field is redefining and reshaping discussions about rhetoric. The literature here covers both the emergence of data science as a professional field, as well as addressing or posing key questions about the rhetorical nature of writing or visual rhetoric within data science.

13.1 Impact within Information Systems

Agarwal and Dhar note that the modern inceptions of business analytics and data science have been driven by rapid socioeconomic development as well as the expedience of digital tools to make large-scale analysis possible with immense data sets [26]. Their analysis is within the area of business analytics known as information systems or IS. This area is described in terms of an occupational industry whose rhetorical framework is driven more by the needs and value determination of a company or partner, rather than by a higher pursuit of pure knowledge or a broader societal goal. In their piece, the authors describe the focus of their field in the following way: "We care deeply about business transformation and value creation through data, and less for algorithms or frameworks without a linkage to business value" [26].

Although the authors claim that many theories within areas such as economics and psychology have stemmed from research of the type they conduct [26], it is clear that the value of their operations and writing is largely dependent upon business outcomes. This is why the value of big data and data science to these individuals is largely utilitarian. However, from a research point of view, Agarwal and Dhar do value research aided by data science techniques that helps "pose an interesting and relevant question" [26] and in this way speaks to some pursuit of higher knowledge. It is interesting to note that one potential problem here with big data sets, which are the core of data science, is the "questionable at best or unknown at worst" credentials the source of such data sets might have [26]. This is one major issue that must be reconciled for data science to be considered a true profession, since there is a constant need for data sets, yet the open-endedness of open source data sets poses problems not inherent to the traditional professions.

13.2 Rhetorical Questions about Data Science

In a more rhetorical take, Boyd and Crawford ask key questions about data science and big data, which they define without simply relying on the notion of a large set of data to analyse [27]. Rather, they use a threefold approach to defining the term: maximizing computational power, drawing on large data sets to find patterns to make claims, and viewing these large data sets as inherently more objective or valid as sources of new knowledge and conclusions [27]. This sets the tone for big data having one of two potential types of rhetoric. The first is utopian, the notion that follows directly from a trusted interpretation of a data science approach to big data as revealing and useful for a society that values progress and new information [27]. The converse is a dystopian rhetoric in which privacy rights are further infringed upon and data is used as a vector for increasing control from the top [27].

The article defines six ways in which big data's influence demands certain thoughts and considerations from us a society. They are as follows:

1. *Changing the definition of knowledge.* This is the notion that "the numbers speak for themselves" [27]. The authors claim that although the destabilization that big data has brought to various fields and industries is challenging classical methodologies for discovery, it should not be taken purely on its own merit [27]. That is, we need to start considering exactly how robust these data science approaches are and what relying on them means if we want to reframe our understanding of knowledge in terms of their results.

2. *Subjectivity vs. objectivity.* The assumption that big data results are objective and inherently accurate is challenged here. The authors assert that just as with other branches of science, big data advocates can sometimes misconstrue what objectivity means in their practices and how subjectivity still is a presence to wrestle with [27]. Interpretation is still critical [27], and researchers have a critical role of making sense of their data [27]. To refer back to a prior sentiment, the numbers can't speak for themselves. That is the job of the researchers, and with human intervention comes biases and nuances. Additionally, the authors draw attention to the problem of apophenia, which is the phenomenon of "seeing patters where none actually exist, simply because enormous quantities of data offer connections that radiate in all directions" [27].

3. *Bigger is not necessarily better.* Twitter and Facebook are two familiar sources of extremely large amounts of people who provide exorbitant

amounts of data. And it has been a long sought after goal in the sciences to be able to work with data sets these large as opposed to the traditionally small sample sizes. But there are problems inherent to these immense population pools, such as problems with representation [27], compounding the error through intersecting large data sets [27], and valuing small data even to the anecdotal level that can offer insights beyond data mining [27].

4. *Context is key.* The authors note the distinction between 'personal networks' developed in traditional sociological research, 'articulated networks' generated technically or digitally, and 'behavioral networks' through interact through a communication channel [27]. Here the issue at hand is in the relevance of connections between individuals in these networks. Not every connection equal in its significance [27].

5. *Accessibility does not mean it is ethical.* Big data solutions most notably run into ethical problems by compromising privacy. Typical techniques include reverse-engineering anonymized data [27] or taking individuals' online content out of context [27].

6. *Limited access causes issues.* One major concern within data science is how one gets access to the critical data sets needed for analysis. Part of the issue is that for social network data sets, large social media companies control access, and a similar issue exists for universities which have a similarly restrictive means of access [27]. The additions of a barrier to entry for analysis based on computational knowledge and the pressure from the sponsor company on researchers means that there is a privileged divide based on who can access and use certain critical datasets and for what purposes [27].

13.3 Visual Rhetoric

Michael Salvo of Purdue University has examined the visual rhetoric that has emerged from big data analysis in recent years [28]. Salvo discusses at a high level how big data and visual rhetoric collide. He asserts that the critical "design of communication" of data-oriented results in any form is rhetorical [28]. That is, the way that data scientists visually construct results from their data analysis deal intimately with questions of rhetoric by their very nature. One important feature is targeting an audience. Salvo references accessible, audience-driven visual displays of data and information such as TED Talks and IBM's THINK exhibit [28] as examples of how individuals

are constructing methods for distributing ideas which are data-intensive and require special attention to get an audience to pay attention.

Interestingly, Salvo claims that an Aristotelian rhetoric is being revived through the visual rhetoric of big data. Specifically, this is being done in the form of re-establishing the "linkage between rhetoric and probability" [28]. If Salvo is correct, then this poses an interesting justification for studying the visual outcomes of big data in rhetorical terms.

13.4 Additional Literature

The idea of applying machine learning techniques and database analysis of the type found within the sphere of data science is "relatively new to higher education," according to Dursun Delen [9]. Studies in data science have not been heavily driven toward academia since the corporate world and social media networks have been the main sources for large data sets. However, there is an increased push for this type of research. Some researchers have taken to calling this area of study Educational Data Mining,' which uses standard data science techniques to analyse data sets acquired from educational sources, namely universities and secondary schools [10]. This is of particular relevance due to the case study of the ongoing WPI project bridging the gap between occupational, career-oriented data mining and Educational Data Mining, since the work is involved in the university's Career Development Center (CDC).

Provost and Fawcett reaffirm the idea data science is largely still a product of the private sector. They begin their piece by observing that "companies in almost every industry are focused on exploiting data for competitive advantage" [29]. They go on to assert that beyond using algorithms to get results, data scientists need to have the ability to "view business problems from a data perspective" [29]. Both these assertions fit into the authors' view that the primary goal of data science is to improve a company's decision-making process by informing relevant parties within the company using data as a source for their claims, as opposed to intuition alone [29]. This is a highly occupational interpretation of data science, and it is interesting to note that this article was published quite recently, in 2013. It was also published through *Mary Ann Liebert Inc.* in their *Big Data* journal, which claims to be a peer-reviewed journal specific to developments in data science and analytics according to the journal's description on the publisher's website.

14 Methodology

A collection of papers from traditional mathematics and data science will be collected and compared to illustrate similarities and differences in the structure of their writing. A grammatical analysis will be conducted on the abstract level. Key passages will be highlighted and examined in detail to illustrate differences among the samples of writing to offer insights about the rhetorical distinctions between them.

A grounding point for this research will be mathematical journal articles from professional sources. These include the *Journal of Applied Mathematics, Statistics, and Informatics* (JAMSI), *Probability and Mathematical Statistics*, and Cornell's *arXiv* journal archive on Statistical Theory. The areas of focus in the mathematical realm of discussion will be on recent (from the past five years) papers in statistical theory and probability. While mathematics is a widely diverse field encompassing numerous professional domains, for the sake of this study these fields of mathematics research will be considered as sufficient, long-established professional areas within the subject. Abstracts from two articles will be picked from the journals listed above.

Comparative analysis will be conducted between the grammatical structure and rhetorical moves in the abstracts from articles in the mathematics articles versus data science journal articles. There are two major types of these second articles which will be encountered. The first types are data science articles which pertain to a particular industry and are largely within the confines of occupational rhetoric. Classical data science articles tend to be published in journals pertaining to the field in question for which they conducted research. The example used in the pilot study (see below) was published in a business journal, for instance. The second types are articles of a more academic style, where the article was published in a specific journal pertaining to data science. These articles should be expected to share more similarities in their language construction with the mathematical articles. In this way, the evolution of data science out of a mathematical context into an applied proto-profession and finally into its own independent field of professional study can be identified and explored. To parallel the mathematical journal sources, two data science journal articles will be chosen.

Along with a grammatical analysis, some data will be generated based on two major points of interest: subject-verb pair proximity and subject word choice. In the former case, an average will be computed for all subjects and verbs in the abstract to show proximity as a measure of certain characteristics of the writing, such as emphasis on brevity, a consequence of

logical objectivity, lending time to descriptions, and so on. The basic formula will be to count the number of words between a subject and its verb, counting each word between an auxiliary verb pair and any words that split up compound subjects or verbs. As an example, in the sentence below:

> This **process** *has* good expansibility and adaptibility and can meet the needs of big data quality assessment.

The scheme would count the five highlighted words to produce a proximity score of 5, because the full phrase has two verbs which relate back to the same subject, even though one of those two pairs has no gaps between them ("This **process** *has*").

Second, word choice for the subjects of each clause will be broken down and compared between the journal articles. A more complete analysis would use a textual analysis program to create a word frequency count of all words in the journal article. Since the abstract is a limited framework and unlikely to yield substantial results given the limited sample size of words (generally less than 500 words at most), this will be considered only for subjects, where even small amounts of repetition and consistency can reveal important results.

15 Results

The following details the breakdown of analysis from each of the abstracts considered. A brief overview of the topic of each paper is offered, as well as the journal of publication and the general subject matter of the paper's thesis. Each abstract is then treated on its own and observed to determine relevant grammatical and rhetorical moves which distinctly characterize the professional environment in which they were made. These are then compared across the two disciplines to highlight key differences and similarities. A small sketch of the results is displayed through data visualization by counting occurrences of certain recurring instances of certain grammar usage and subject-verb proximity.

15.1 Statistical Theory Papers

Two papers were selected for an abstract analysis of topics relevant to the professional mathematical fields from which data science emerged, in this case statistics and probability. One paper was chosen from *Mathematical Methods of Statistics*, and the other from *Probability and Mathematical*

Statistics. Both papers have appeared in other mathematical journals and publication outlets as well.

15.1.1 Baxter's Inequality

The first case we consider is a paper by [30], published in *Mathematical Methods of Statistics* in April 2015 (with the original article dating to April 2014). This particular article by Meyer, et al. was chosen at random from a selection of recent additions to the article through the Springer webpage. The abstract appears in full below:

> A central problem in time series analysis is prediction of a future observation. The theory of optimal linear prediction has been well understood since the seminal work of A. Kolmogorov and N. Wiener during World War II. A simplifying assumption is to assume that one-step-ahead prediction is carried out based on observing the infinite past of the time series. In practice, however, only a finite stretch of the recent past is observed. In this context, Baxter's inequality is a fundamental tool for understanding how the coefficients in the finite-past predictor relate to those based on the infinite past. We prove a generalization of Baxter's inequality for triangular arrays of stationary random variables under the condition that the spectral density functions associated with the different rows converge. The motivating examples are statistical time series settings where the autoregressive coefficients are re-estimated as new data are acquired, producing new fitted processesand new predictors for each n. [30]

Unlike in the mathematical paper on algebraic geometry from the pilot study, the authors of this paper took the time in their abstract to discuss some background information regarding the topic. In fact, over half of the abstract is dedicated to framing the problem and giving information about it to the reader. Still, despite this being presented, the density of the subject matter is only slightly alleviated for a reader who is not privy to the technical terminology being used. For example, consider the sentences:

> A central problem in time series analysis is prediction of a future observation. The theory of optimal linear prediction has been well understood since the seminal work of A. Kolmogorov and N. Wiener during World War II. [30]

In the abstract's first two sentences, the main problem of the paper is explicitly outlined and some history is offered in brief. However, an understanding of the terminology is still required. Note the immediate introduction of terms like "time series analysis" and "optimal linear prediction," which are underlined in the passage above to distinguish them. The historically relevant background requires recognition of the work of Norbert Weiner and Andrey Komogorov, whose work on signal extraction and statistical filtering is relevant to this work [31]. Here, their work is deemed "seminal" and not described in any further detail in the paper. It is then left up to the reader to investigate.

The inherent brevity of an abstract needs to be taken into account, since only the most high-level and relevant information can be discussed in such a short passage. Compared to the pilot study's algebraic geometry example, these authors do a better job at giving readers of a varying knowledge background some context, and it develops the problem in more detail, while the pilot study's example. The statements about the authors' own work and added value to the field are reserved for the very last sentence in fact. Although the surrounding information needed to understand Baxter's inequality are provided here, they are done so in a way with an assumption about deep prerequisite knowledge in statistics, time series prediction, and linear algebra. Observe for example this first portion of the introduction from that same paper:

> Baxter's inequality (Baxter (1962), Baxter (1963)) provides a fundamental tool for understanding the behavior of linear predictors based on a finite observed history. In particular, let $(X_t)_{t\in Z}$ be a mean zero, weakly stationary time series. Under reasonably general conditions (see, for example, Kreiss, Paparoditis and Politis (2011), Pourahmadi (2001), Wiener and Masani (1958)), such a process admits an AR(∞) representation [30]

Where the authors go on to immediately set up a summation notation for the time series X_t in the final part of that paragraph. There is no time wasted in getting to the mathematics. Any lost information on the reader's part is directed to several papers from foundational authors of the subject, with two being more recent authors from 2001 and 2011.

In this example, the majority of the abstract is structured using to-be verbs (mainly "is") where in the descriptive background sections of the abstract, the distance between subjects and verbs is much greater than in the pilot study. Still, generally they are close in each clause. It is only when

we get to the sections relevant to the authors' own work that this proximity is reduced and remains consistent. Using a similar encoding scheme where subjects are **bolded** and verbs are *italicized*:

> A central **problem** in time series analysis *is* prediction of a future observation. The **theory** of optimal linear prediction *has been* well understood since the seminal work of A. Kolmogorov and N. Wiener during World War II. A simplifying **assumption** *is* to assume that one-step-ahead **prediction** *is carried* out based on observing the infinite past of the time series. In practice, however, only a finite **stretch** of the recent past *is observed*. In this context, **Baxter's inequality** *is* a fundamental tool for understanding how the **coefficients** in the finite-past predictor *relate* to those based on the infinite past. **We** *prove* a generalization of Baxter's inequality for triangular arrays of stationary random variables under the condition that the spectral density **functions** associated with the different rows *converge*. The motivating **examples** *are* statistical time series settings where the autoregressive **coefficients** *are re-estimated* as new data are acquired, producing new fitted processesand new predictors for each n. [30]

Notice that while the first half of the abstract contains many pairings which are not directly next to each other, it is mostly due to descriptive mathematical language in prepositional phrases or through long adjective phrases. Consider the following sentences, where the prepositional phrases are identified by parenthesis:

> The **theory** of optimal linear prediction *has been* well understood... [30]

> In this context, **Baxter's inequality** *is* a fundamental tool for understanding how the **coefficients** in the finite-past predictor *relate* to those based on the infinite past. [30]

These are necessary prepositional phrases to describe which types of "theories" and "coefficients" are being discussed by the authors. In the limited cases where strong subject-verb agreement is broken, the implication of logical objectivity and confident assertion is not so much replaced, but rather supplemented, by a technically descriptive break that is still only accessible to those within the professional space with the required knowledge to make sense of the topics.

15.1.2 Extremes of Chi-Square Processes

We now consider a paper by [32], published in *Probability and Mathematical Statistics* in 2016 (originally from February 2015) and archived through *arXiv*. In this piece, titled "Extremes of Chi-Square Processes with Trend," Liu and Ji investigate the chi-square process, which is a more theoretical and mathematically involved topic related to the familiar chi-square test which is often taught in elementary statistics courses [33]. The abstract is as follows:

> This paper studies the supremum of chi-square processes with trend over a threshold-dependent-time horizon. Under the assumptions that the chi-square process is generated from a centered self-similar Gaussian process and the trend function is modeled by a polynomial function, we obtain the exact tail asymptotics of the supremum of the chi-square process with trend. These results are of interest in applications in engineering insurance, queuing and statistics, etc. Some possible extensions of our results are also discussed. [32]

Clearer and more succinct subject-verb agreement dominates in each sentence. Unlike in the previous example, there is little attention given to contextual background information on the subjects being described. What is offered in terms of background information is described as a part of the formulation of the problem leading into the results, not offered separately in any historical or prerequisite context like in the "Baxter's Inequality" piece. The relevance of the results to various fields and potential applications are alluded to, but these do not lend any help to a reader outside the professional framework if they do not understand what precedes them.

We begin with a similar analysis as before, again having subjects in bold and verbs in italics to show the basic structure of each sentence:

> This **paper** *studies* the supremum of chi-square processes with trend over a threshold-dependent-time horizon. Under the assumptions that the chi-square **process** *is generated* from a centered self-similar Gaussian process and the trend **function** *is modeled* by a polynomial function, **we** *obtain* the exact tail asymptotics of the supremum of the chi-square process with trend. These **results** *are* of interest in applications in engineering insurance, queuing and statistics, etc. Some possible **extensions** of our results *are* also *discussed*.

Here the focus of subjects is largely upon the paper itself and the research that went into it. For example, the use of "paper" and "results" overwhelms any usage of terminology related to the field in question as the main subject of any particular clause, the exception being:

> Under the assumptions that the chi-square **process** *is generated* from a centered self-similar Gaussian process and the trend **function** *is modeled* by a polynomial function... [32]

This portion is where the mathematical content of the paper is given the most attention, segueing into the major result of the paper in the final clause. The result is described in terms of the "we" that represents the authors, who obtained it. This is set up as an assumption from which the results flow out of, which is a basic structure of a proof. That is, to establish a set of assumptions and from there construct a proof that produces a result. In the case of this paper, the body of the article is the middle step here, the proof itself, with the beginning and end summarized through the abstract in this portion.

Interestingly, this abstract embodies an element of the pilot study's mathematical case example, in that in the above two sentences, a complex noun phrase was not used as the subject. The difference being that in this usage, the passive voice was used to avoid this, whereas the opposite was true in the algebraic geometry example. The sentence could have been written to begin like so:

> Under the assumptions that a **centered self-similar Gaussian process** generates the chi-square process and a **polynomial function** models the trend function...

However, this variant buries the key point of interest in this description's first part: the chi-square process which is the main focus of the whole paper. While the second instance relating to the trend function is ambiguous as far as which variant is preferable, it is reasonable that the long noun phrase should not be used in the first part of this section.

Generally, we observe continued trends that stem from the nature of the mathematical framework in which the paper was written. A logical, objective flow of information at the high level emphasized by computational language such as the use of verbs like "generated" and "modeled." More significantly, this abstract shows the emphasis on the importance of veritable results. With the abstract's subjects and the paper itself being highly

featured, it demonstrates that the value of this paper lies in the results that the two authors, with their specialized and exclusive knowledge, were able to obtain. Applications and relevance to other fields is a secondary concern, coming at the end of the passage.

15.2 Data Science Papers

Two additional papers were selected for an abstract analysis of topics relevant to data science in the modern context. The first was chosen from a business journal titled the *Journal of Business Logistics*. The other was chosen from the *Data Science Journal*. The former is oriented toward a non-data science field, but one which has embraced the study as a way to enhance results across large swathes of the business sector. The latter is a journal specifically for publishing data science results from studies done about the subject in its own space.

15.2.1 Supply Chain Design and Management

The first data science article we consider is titled "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management" [34]. This piece was published in the *Journal of Business Logistics* in 2013 by Waller and Fawcett and pertains to areas of interest in the intersection of supply chain management and the new field of data science. Its abstract is as follows:

> We illuminate the myriad of opportunities for research where supply chain management (SCM) intersects with data science, predictive analytics, and big data, collectively referred to as DPB. We show that these terms are not only becoming popular but are also relevant to supply chain research and education. Data science requires both domain knowledge and a broad set of quantitative skills, but there is a dearth of literature on the topic and many questions. We call for research on skills that are needed by SCM data scientists and discuss how such skills and domain knowledge affect the effectiveness of an SCM data scientist. Such knowledge is crucial to develop future supply chain leaders. We propose definitions of data science and predictive analytics as applied to SCM. We examine possible applications of DPB in practice and provide examples of research questions from these applications, as well as examples of research questions employing DPB that stem from management theories. [34]

An initial reading shows a clear preference for a particular subject: we. If we treat the abstract as we have previously, coding subjects in bold and verbs in italics:

> **We** *illuminate* the myriad of opportunities for research where **supply chain management** (SCM) *intersects* with data science, predictive analytics, and big data, collectively referred to as DPB. **We** *show* that **these** terms *are* not only *becoming* popular but are also relevant to supply chain research and education. **Data science** *requires* both domain knowledge and a broad set of quantitative skills, but there is a dearth of literature on the topic and many questions. **We** *call* for research on skills that are needed by SCM data scientists and discuss how such **skills** and domain **knowledge** *affect* the effectiveness of an SCM data scientist. Such **knowledge** *is* crucial to develop future supply chain leaders. **We** *propose* definitions of data science and predictive analytics as applied to SCM. **We** *examine* possible applications of DPB in practice and *provide* examples of research questions from these applications, as well as examples of research questions employing DPB that stem from management theories. [34]

The "**We** *verb*" pairing is common throughout the whole abstract. This parallels the style of the second mathematical example in this paper more closely than the first, since in that paper the focus was heavily on the paper itself and its authors. But it is much more abundant here. This seems to suggest that the subject matter experts writing in this case have a claim to knowledge that they can assert directly. While this proximity and strong active voice is even stronger than in the mathematical examples, the overreliance on the actions of the researchers shows that the material itself does not suffice to stand alone to convey the necessary meaning.

When mathematicians write about their subject matter, they do assert a claim to their knowledge as researchers and mathematics scholars, but one crucial component of mathematical and all professional rhetoric is that while it takes a certain type of individual to acquire and expand the boundaries of knowledge, the knowledge itself is presumed to be immutable, standalone, and beyond any one individual's contributions. In mathematics, Pythagoras's Theorem is true, and Pythagoras discovered it in the Western world and is therefore the origin of its name. One could argue that Pythagoras had an esoteric level of knowledge that allowed him to draw that conclusion. But especially in the context of ancient Greek mathematics, this would be

rejected, because the fact that Pythagoras found was simply true by itself. Such mathematical realities, which in our time are given names by their discoverers, exist independently of those individuals. So in the case of this data science article calling for investigation into predictive analysis and other data science tools to be used in supply chain management, the authors are seeking out such knowledge, but make no claim to know it in the same way as a mathematician.

Notice the language that is used here as well. These authors "illuminate," rather than assert, facts about data science as it pertains to the business industry in question. They are "calling" for and "examining" these facts, in order that ways they may be useful in the occupational space of supply chain management can be determined and applied. Unlike in the mathematical examples, the word choice and language used is mostly non-technical, with some exceptions such as the few data science terms ("predictive analysis" and "big data"), which are not nearly as complex, narrow, or unknown to the public as terms in previous papers, such as "Gaussian processes" or stationary random variables." One could imagine almost any audience being able to read and understand the basic idea of this paper. This is the exact opposite case for most mathematical articles, applied or theoretical. A small, technically-advanced readership is expected by most researchers writing about areas like statistical theory or probability, let alone even more abstract fields like complex analysis, group theory, or algebraic geometry.

15.2.2 Data Quality and Quality Assessment

The second paper is a data science journal article titled "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era" by Cai and Zhu [35]. This was published in the *Data Science Journal* in 2015, which is from the more applied, industry results oriented articles data science articles might also be published in. Data science is treated as an emerging profession, or proto-profession, so there are different styles of writing which are written with different rhetorical frameworks. In this example, we look at an example where the writing is rhetorically closer to a professional piece of writing. Let us look at the abstract from the first page:

> High-quality data are the precondition for analyzing and using big data and for guaranteeing the value of the data. Currently, comprehensive analysis and research of quality standards and quality assessment methods for big data are lacking. First, this

> paper summarizes reviews of data quality research. Second, this paper analyzes the data characteristics of the big data environment, presents quality challenges faced by big data, and formulates a hierarchical data quality framework from the perspective of data users. This framework consists of big data quality dimensions, quality characteristics, and quality indexes. Finally, on the basis of this framework, this paper constructs a dynamic assessment process for data quality. This process has good expansibility and adaptability and can meet the needs of big data quality assessment. The research results enrich the theoretical scope of big data and lay a solid foundation for the future by establishing an assessment model and studying evaluation algorithms. [35]

We begin by similarly deconstructing the sentence structure by identifying subject-verb pairings and analyzing the rationale behind these choices, as well as how the surrounding language and word choice figures in. Identifying the subjects and verbs in this passage we find:

> High-quality **data** *are* the precondition for analyzing and using big data and for guaranteeing the value of the data. Currently, comprehensive **analysis** and **research** of quality standards and quality assessment methods for big data *are lacking*. First, this **paper** *summarizes* reviews of data quality research. Second, this **paper***analyzes* the data characteristics of the big data environment, *presents* quality challenges faced by big data, and *formulates* a hierarchical data quality framework from the perspective of data users. This **framework** *consists* of big data quality dimensions, quality characteristics, and quality indexes. Finally, on the basis of this framework, this **paper** *constructs* a dynamic assessment process for data quality. This **process** *has* good expansibility and adaptability and can meet the needs of big data quality assessment. The research **results** *enrich* the theoretical scope of big data and lay a solid foundation for the future by establishing an assessment model and studying evaluation algorithms. [35]

Several differences are notable between this and the mathematical passages. First, there is a significant abundance of surrounding context in this abstract. Before analyzing subject-verb pairings, it is important to note

that without even identifying them, the sentences written around them are largely explanatory. Consider the following excerpts:

> High-quality **data** *are* the precondition for analyzing and using big data and for guaranteeing the value of the data. [35]

Here the subject-verb pair is in close proximity, but it is not used to segue into a description of an action relevant to the research. Rather, it relates to a description of what "high-quality" data is in terms of data value. That is, this type of data is a precondition, a necessity, for analyzing big data in a meaningful way. There is no such explanatory clause in any sentence in the previous abstract for the mathematics example. It is evident that in this example, while some knowledge is still necessarily assumed on the part of the authors, there is a variety of information which is not assumed to be known by an audience. This is characteristic of a more general audience where the restrictions on readership are not as rigidly enforced in the rhetoric of describing new knowledge. Consider another example of such a clause in a sentence:

> This **framework** *consists* of big data quality dimensions, quality characteristics, and quality indexes. [35]

Here the "hierarchical data framework" mentioned in the sentence before this example is defined in terms of what it "consists" of. Again we see a case where a term is stated but then described in some brief detail to give the reader an understanding of what is being discussed.

Going back to the subject-verb pairings and their proximity, we notice that while there is a stronger amount of explanatory, supplementary writing, most sentences follow the same type of agreement as before. Most sentences exhibit strong subject-verb agreement in close proximity at the beginning of sentences, such as "High-level data are" and "this paper analyzes." The one exception to this which is not caused by a conjunction is:

> Currently, comprehensive **analysis** and **research** of quality standards and quality assessment methods for big data *are lacking.* [35]

This is a passive phrase, but unlike the passive in the first example, this one exists for a different reason. In the mathematics paper, the passive was a result of the way that mathematical terminology is defined. That is, formulas or definitions are "given" not necessarily by an actor who gives

researchers a fact, but by the established existence of such a fact which can be assumed to be true, and hence be "given" without demonstration. In this case however, the passive frontloads the two subjects of the sentence: analysis and research, particularly of quality standards and assessment methods. This is the crux of the paper and what the aim of the research focuses on. So to justify the major reason for this writing, a lack of analysis and research in this area, it is vital to ensure that the focus of the sentence is on the lacking elements: analysis and research. Quality standards and assessment can suffice in a prepositional phrase, since the reader can infer from the title of the paper that this is the essence of the research without being explicitly told in this case with a subject-burying edit such as:

> ...quality standards and quality assessment methods for big data *are lacking.* [35]

Similar to the statistical theory examples, emphasis is given to actions of the research in some subject-verb pairings. "First, this paper summarizes," "Second, this paper analyzes," "The research results enrich" are some examples. However, not only do these not make up a majority of the pairs in the abstract, but even they are far weaker in their assertiveness and rigor than in the mathematics example. Here the actions are summarizing, analyzing, and enriching. Analyzing is the strongest, most active verb here in terms of what the researchers themselves are doing. Compare this to computing or applying, which are more substantial and active than the acts of summary of existing material or enrichment which is not directly done by the researchers to the abstract theoretical scope in consideration.

15.3 Data Visualization

The results of the analysis for subject-verb proximity and word frequency are provided below. MA1 and MA2 denote the first and second mathematics papers whose abstracts were considered, in order as they were presented in the paper. DS1 and DS2 likewise follow from the first and second data science abstract presented here. Note that while AVG pertains to the averages for each table, TAVG is the average of the two averages for each domain, mathematics and data science.

MA1	
Pair#	*Distance*
1	4
2	4
3	0
4	0
5	4
6	0
7	4
8	0
9	5
10	0
11	0
AVG	1.91

MA2	
Pair#	*Distance*
1	0
2	0
3	0
4	0
5	0
6	4
AVG	0.67
TAVG	1.29

Figure 27: Frequency counts and averages for the two mathematics abstracts. AVG pertains to each individual abstract's averages. TAVG refers to the average of the two averages, so the total average.

MA1			
Subject	*Frequency*	*Verb*	*Frequency*
problem	1	is	3
theory	1	is + verb	2
assumption	1	are	1
prediction	1	are + verb	1
stretch	1	has been	1
Baxter's inequality	1	relate	1
coefficients	2	prove	1
we	1	converge	1
functions	1		
examples	1		

MA2			
Subject	*Frequency*	*Verb*	*Frequency*
paper	1	studies	1
process	1	is + verb	2
function	1	obtain	1
we	1	are	1
results	1	are + verb	1
extensions	1		

Figure 28: Subject/verb word frequency counts for the two mathematics abstracts.

DS1	
Pair#	*Distance*
1	0
2	0
3	0
4	3
5	0
6	0
7	2
8	0
9	0
10	7
AVG	1.20

DS2	
Pair#	*Distance*
1	0
2	11
3	0
4	15
5	0
6	0
7	5
8	7
AVG	4.75
TAVG	2.98

Figure 29: Frequency counts and averages for the two data science abstracts. TAVG refers to the average of the two averages, so the total average.

DS1			
Subject	*Frequency*	*Verb*	*Frequency*
we	5	illuminate	1
supply chain management	1	intersects	1
these	1	show	1
data science	1	are + verb	1
skills	1	requires	1
knowledge	1	call	1
		affect	1
		is + verb	1
		examine	1
		provide	1

DS2			
Subject	*Frequency*	*Verb*	*Frequency*
data	1	are	1
analysis	1	are + verb	1
research	1	analyzes	1
paper	3	presents	1
framework	1	formulates	1
process	1	consists	1
results	1	constructs	1
		has	1
		enrich	1
		lay	1

Figure 30: Subject/verb word frequency counts for the two data science abstracts.

The average proximity of subjects and verbs in the statistical theory mathematics papers were lower than those in the data science ones. That is, in mathematical papers, we would expect that generally there are stronger subject-verb pairings. While in data science articles, there is more of a

breakup that separates subjects from their verbs. In the examples described in this paper, it appears that a verb, rather than a subject, is more likely to get buried as a result of this. High-valued proximities tended to come from subjects which came near the beginning of a sentence or clause, but whose verbs were split up by prepositional phrases in between, or whose verbs were multiple and listed or otherwise split up.

Interestingly, subjects in the first mathematics and first data science abstracts both give a reader, without any added information, some concrete ideas about what the paper is about. In contrast, the second of each type uses more vague subjects which obscure the specific topics being discussed by the authors. There is also not as clear a distinction between the two fields in terms of how subjects vary and are used to convey meaning. Both types show instances of relying on the author or the paper as the subject, while both also sometimes rely on the subject matter speaking for itself. The above analysis shows in more detail, beyond the numbers, how the nuances of language explain some of this overlap.

To be verbs are extremely common in both works. Verbs such as "is" and variants where it is paired with another verb as a helper ("is generated" for instance) and related verbs made up a significant portion of all four abstracts. The verbs in the mathematics papers tend to be more direct, definitive, and logical ("converge," "prove," "obtain") while those in the data science articles vary between being indirect and tempered as well as direct and echoing more of the professional language of mathematics ("illuminate" and "examine" vs. "analyzes" and "constructs"). Again, the above analysis offers more insight beyond the numbers by themselves.

Note also that the subjects of data science abstracts are more likely to be concentrated on a single repeated subject, whereas there is a wider array of subjects in mathematics abstracts. The converse is true for verbs. There is more repetition of particular to be verbs in the mathematics abstracts, while there is an evenly dispersed selection of verbs in the data science ones.

Chapter 2: Data Science Design and Analysis

16 Introduction

Continuing from the previous chapter on data science as a proto-profession, this chapter concludes the Afterword with a look at the Major Qualifying Project in terms of design and a rhetorical look at analysis and the implications, biases, and expectations surrounding it. Finally, the chapter concludes on thoughts concerning the relationship epistemologically between data science and rhetoric, relative to the truth.

17 Background

We now examine the CDC Data Analytics MQP case study to highlight key elements of both data science as a rising proto-profession in an academic context, as well as to show how many of the rhetorical issues which were raised in the previous chapter manifest in an actual process of data analysis. This chapter will treat the project as an active instance of mathematicians engaging in machine learning and data analysis while confronting and having to wrestle with rhetorical junctures with regards to data pre-processing, feature selection, programming, and decision-making. Additionally, it will consider the stakeholders in question, primarily the CDC, and what their intentions and expectations are, as well as their role in framing the project as a whole. This framework includes elements such as the data made available from the CDC, the assumed expectations about their desired outcomes, and their input via our primary sponsor and other staff members throughout the project.

18 Literature Review

18.1 Ambient Research for Data Science Studies

One of the key promises big data offers researchers in all fields is the ability to make sense of immense datasets which can tell us extensive information about broad and intersecting subjects of interest. This focus on the collective and on using databases which cover a wide range of subjects and timeframes can lead to problems, however. McNely describes these kinds of "distributed databases" as having the most applicability for aggregate analysis which is "depersonalized, decontextualized, and pasted together" with other databases [36]. This is not far from the truth about how current big data analysis is conducted and what its major outcomes tend to be about.

Consider for example that in the aggregate, a company like Amazon might have a serious interest in knowing what a consumer network looks like for electronic devices, such as laptops. Simple data analysis can reveal useful but shallow statistics about purchasing from single dimensions. Looking at data from a single fiscal year of operations can reveal, for example, the average price a user paid for a laptop, which brand sold the most in that time, or what percentage of users took advantage of Amazon Prime in their decision-making. What becomes more interesting, especially for the crucial feature of product recommendations, is to ask questions about other related datasets and merge them. The necessary question to ask to develop such a network map of consumer preferences and packaged purchases is to look at other electronic devices and non-electronic items which users bundled in their purchase. Using this information, Amazon can begin to develop a probabilistic model to recommend users who, say, buy a laptop that a certain accessory item (a wireless mouse, laptop bag, etc.) is "Frequently Bought Together" with the item in question.

Even more useful for Amazon's business operations still would be to access information which is tailored to the user. Further intersecting datasets might include an individual's own electronic device purchase history or more specifically any items identified as laptops or being related to them. This lays the foundation for tailoring recommendations to specific users, who demonstrate their preferences in this way which would otherwise be impossible to deduce accurately based solely on large-scale, community-driven information. Going beyond its own services and linking in data from other sites would be even more useful. Consumers exhibit patterns and preferences off the site as much as they do on it.

McNely is interested in developing models for communication practices which address an earlier potential problem in this process, however [36]. While Amazon and other companies like it have a vested interest in getting to the heart of the "situated, local, human users" McNely describes [36], their first step in tackling this is to have a process of analysis which is calibrated for this outcome. That is, the data scientists behind the scenes have to take a different approach than simply abiding by the old adage, which advises the analyst to "let the data speak for itself." Doing so can lead the analyst into some significant pitfalls, not only from an analysis point of view, but also in terms of how those results are communicated.

Since information requires "specialized" knowledge to make sense of and become useful for analysis and communication according to Drucker, as cited in [36], the analyst's approach depends heavily on not only the knowledge one brings to the table, but also the assumptions and considerations one

makes throughout the process. Questions to ask in this regard include: are there inherent qualities of the data which need to be taken into account? What assumptions about the individuals who comprise that dataset (or produced its contents) are being made? Which results are thrown out and of those kept, are they being construed to mean something differently than what one might intuitively deduce? What is the agenda? What data is available and which data is made available by the stakeholders?

18.2 The Signal and the Noise

Nate Silver's 2012 book The Signal and the Noise offers an accessible look at contemporary topics in data science and analysis [6]. Here we look at some of the major points from the book which apply to the case study at hand and were influential to the way the project unfolded.

18.2.1 Basic Principles

Silver is quick to caution the reader that data science is not an end-all be-all solution to any significant problem, because the pitfalls of big data analysis are still real and difficult to overcome. As we grow more accustomed to trusting the data and believing that it speaks for itself, we become less in tune to the instances where it fails. And it does fail often. Silver cites examples, from instances as diverse as earthquake prediction and biomedical research results [6].

One of the biggest areas where prediction is a useful tool is politics and international events, in which these failures are no less prevalent. At the end of the 20th century, political scientist Philip Tetlock's research, a testament to this reality, concluded that most experts in relevant areas such as economics or geopolitics "had done barely any better than random chance" and on top of that, had failed to meet the standard of even the most basic statistical models for prediction in their fields [6]. It was only those who adhered to certain traits and approaches who performed well. These rules of thumb are treated as the basis for Silver's own analytical approach. That approach is, in summary:

1. *Probabilistic thinking.* Treating events in terms of likelihood in a range rather than whether or not an outcome is simply reaching a certain threshold. Silver uses an example of a plane having a 90% chance of landing versus a 99.9999% chance, to illustrate how this type of thinking is not the nave observation of a high-value prediction [6].

2. *Self-correcting, changing analysis.* A static model will fail spectacularly if it is attempting to model a volatile and changing reality. New information demands new tweaks and strategies. Silver describes this as a need to have "the best possible forecast todayregardless of what you said last week" [6].

3. *Consensus is the key.* A prediction made by a single individual is open to a host of biases and limited scope of consideration. Bringing in different perspectives and aggregating the prediction is not inherently superior, but does have benefits that an individual prediction simply cannot achieve by itself [6].

These components are necessary but not sufficient. Good data analysis requires more energy and consideration than what is offered above. But those who do not treat the outcomes of their models probabilistically, correct their models as new information becomes available, and consider the input of other people with complementary knowledge risk their models being less than sufficient for prediction.

18.2.2 Value of Qualitative Information

There are limited ways that one can deal with qualitative (categorical) data from a mathematical and computer processing standpoint. For most existing data processing algorithms, that type of data has to be transformed into numerical data for a computer to make sense of it. However, overreliance on this type of numerical data which is easy to process and predict from can lead to inherent biases and bad outcomes.

For one thing, Silver notes that the possibility of this type of transformation into quantitative data is possible and useful [6] when it comes to sports analysis in baseball as an example. Using all the information one can get is better than reducing the scope of the analysis to quantitative variables [6]. One may think that there is more of a risk for bias from a preference for this type of information which might not seem immediately useful since it cannot be manipulated in the same way as purely numerical data. However, this leads to a bias comparable to that of individuals in areas like sports analysis who might put too much stock in superficial player traits (looks, height, charisma, and the general "aesthetics" of the player as Silver describes them) rather than their measurable qualities [6]. The counteracting biases are that a statistician in a similar position might only trust quantifiable information and consider categorical data to be useless.

It is the folly that Silver describes, to assume that "if something cannot be easily quantified, it does not matter" [6].

It is important not to discount this information outright and to synthesize it with numerical data. Qualitative information can usually be transformed through an encoding scheme into a range or binary (0 or 1) set of values. Simple yes or no questions for example can easily be denoted 0 for "No" or 1 for "Yes." Furthermore, the data can be weighted in a certain way where non-numerical indicators are attributed weighting scores (plus or minus a particular value) based on their prevalence. This is a step away from simply encoding them and turning them into numerical data, but it preserves an analytic device which can then be figured into a model or equation and manipulated. Lastly, even information which cannot be quantified or transformed at all still has value and can be treated separately from the model as a way to tailor assumptions or interpret results. Aspects like ethical issues, stakeholder desires, structural discrimination, personal feelings, and the nature of experience cannot be quantified in the ways presented. However, they should not simply be discounted. Rather, data scientists should use them to think about the assumptions they are making and how their model might be influenced by these more elusive factors.

18.2.3 The Problem of Overfitting

While not all data behaves this way, from a statistical standpoint, using methods such as linear regression, one might be presented with the following example from [6]:

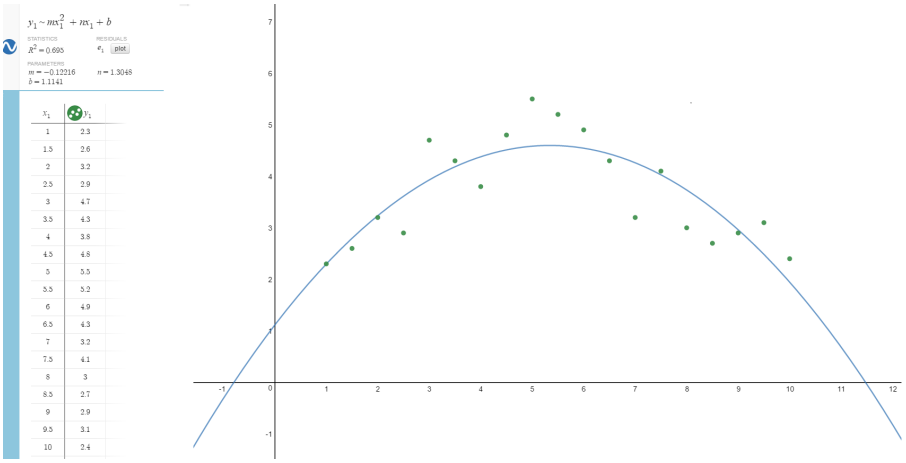

Figure 31: A non-specific example of data fitting, where the data points (green) are the raw data and the model generates an underlying pattern (the parabola) to best-fit the data points. Here quadratic regression is used to generate a parabolic model that best fits the data points, using an equation of the form $y_1 = mx_1^2 + nx_1 + b$. Adapted from [6].

Figure 31 shows a relatively straightforward approach that is commonly used in statistics. Given a set of data points, generate a curve that optimizes the distance between itself and existing points. In this way, the curve (and the function generated from it) can be used to make accurate predictions about unknown values within a certain degree of error. The concept of "overfitting" can be visualized in Figure 32 below, in which the idea of generating a curve that gets as close to as many points as possible is replaced with a goal to get as close to all points:

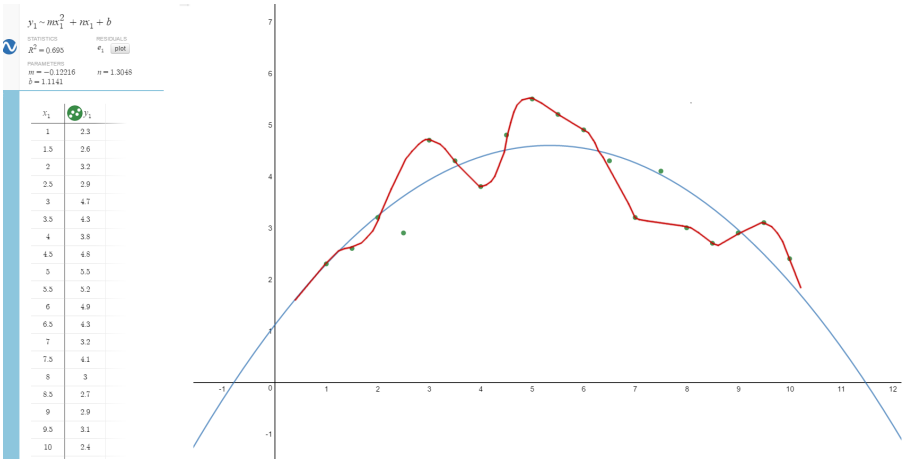

Figure 32: An overfit model using one set of data points from the above example. Note how instead of the underlying model generating a best-fit based on average proximity (blue dotted-line), the overfit model (red) bounces around wildly from point to point in an attempt to hit as many as possible. Adapted from [6].

For the purposes of modelling the existing data set, this overfit model technically performs well. It accurately captures the majority of data points with a high degree of accuracy. The problem comes when one wants to predict for a new, unknown point on the x-axis where there does not already exist a data point. Because this model is so tailored to the existing data and since there are limited points of reference to begin with, it is unlikely that this highly-specific model will have the flexibility to accurately predict a new point either in the range of values between the maximum and minimum, or especially outside that range.

19 Project Application

19.1 Data-Driven Formulations vs. Qualitative Analysis

Above all else, the question that underlies our project is the following: what can a student do during their undergraduate career at WPI to improve their career outcomes? For the purposes of our project, we consider "career outcomes" here to include graduate school (presumably, leading either to stronger credentials for private sector work or to segue into a career in academia), volunteer work, and military service. Not all of these imply a traditional job in the private sector, but all are considered to be positive outcomes from a standard four-year undergraduate experience.

What is undoubtedly a negative outcome from the Career Development Center's perspective is the approximately 9.3% of individuals surveyed who do not fall into any of the successful outcomes categories above [7]. They are unemployed, still searching for opportunities, or have unknown status. The CDC is primarily invested in helping students get employed or find financial opportunity after graduation, so outcomes including this are of great concern.

A significant potential limitation for the CDC data analytics project is that there is a lack of intersecting datasets, localized information, and qualitative analysis. The basis of our major dataset is a combination of three separate datasets provided by the CDC: 2015 self-reporting career outcomes data, CDC kiosk usage, and internships (along with co-ops) during their four yearlong undergraduate careers. Although these datasets were merged based on common user keys to link between the subjects of interest from the class of 2015 to identify their kiosk use frequency, this offers a narrow scope of potential results. In short, they do not offer a complete picture of the student and fail to cover a broad enough scope of student activity which might lead to successful career outcomes.

One critical indicator of success which is missing altogether is a detailed focus on academic success markers. Walk Ins are a highly significant predictor of Full Time job. By saying that our model predicts full time employment most accurately when using Walk In attendance as a predictor, the implication is that those who attend career fairs more often have a better chance at being employed in a full time job after graduation. There are limitations to the accuracy of this statement and just how powerful it is. One must consider that the method used here treats a subset of the data first as a training set (where the algorithm is told all the predictors and response values) and then another subset is treated as a testing set, where the predictors

are known but the algorithm must essentially guess the response values.

The reader is encouraged to review the background literature in the second Afterward chapter as well as Section 5.3.2 on the synthetic minority over-sampling algorithm (SMOTE) for more information and details on how this functions mathematically and algorithmically. But in short, this approach is, in a large sense, mimicking the existing data to produce results which imply correlation. It is different than saying that some percent of people who attend Walk In services a certain amount of times will get a full time job.

19.2 Expectations of Outcome

It was clear that we had to be conscious how we framed our results with the CDC given that innate desire to want results which are directly correlated in this way. That is, to make sure it was understood what our model was capable of doing and what it was not able to do with regards to characterizing job opportunities and postgraduate outcomes in relation to CDC usage, internships, and other basic student information such as major and graduation date.

At one point, we presented early results to our CDC sponsor and other people within the organization who had encountered our project for the first time. They had a variety of reactions to it, but one individual's was noteworthy. She stated that she pictured, as her own imagining of the project's deliverable, a pamphlet or other form of advertising the CDC could present to students with statistics about job outcomes on it. The idea was to take the results from our project and directly use them to steer positive outcomes by promoting good practices that would lead to a favourable career path, graduate school, or some other positive result.

This way of thinking is counter to the reality of what our model is able to tell the CDC about student outcomes, however. It is in line with Silver's caution to think in terms of probabilities, which behave differently than basic, intuitive statistics [6]. Suppose that we take one example from our model, which says that non-springtime graduation (fall or winter) and international status (international) were predictive of the unknown/not paid response, following the decision pathway from the internship data tree (pictured above).

But it is not accurate at all to say that this means that international students who graduate in the fall or winter are likely by design to end up without full time jobs or in graduate school. All that this means is that the model predicts this outcome given these conditions most of the time, based

on the data that it was trained on and tested with. The fact that enough students fit this description and have this outcome in a way that makes it predictable using our method should suggest something to the CDC, but it cannot be interpreted as clearly as a direct correlation with postgraduate prospects. In other words, we cannot use this model to offer a result that claims "doing x and y will make you $n\%$ more likely to get a job after college," because we are not evaluating that likelihood.

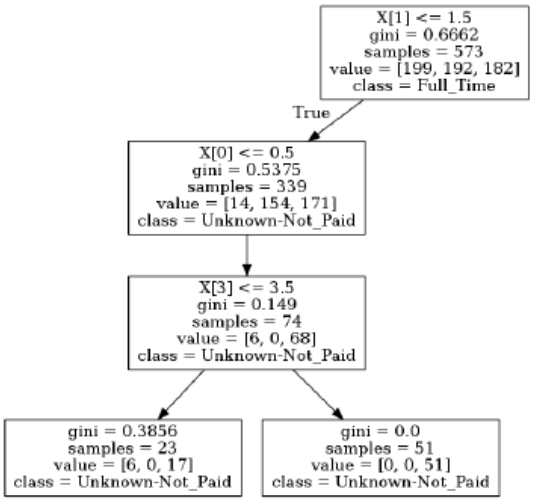

Figure 33: A decision pathway from the decision tree generated from the Internship dataset. Here X[1] refers to degree date, X[0] to international status, and X[3] to number of CDC workshops a student attended. A "True" value for X[1] <= 1.5 means that the student graduated in the fall or winter. A "False" value for X[0] <= 0.5 means that the student is an international student. And the values for X[3] are a split decision between whether a student went to 3 or less workshops or 4 or more during their academic career.

19.3 Heuristic of Invention

During the project, our team had to make several decisions about the data as it was given to us to transform it into a workable format. This included designing heuristic models based on our own knowledge and assumptions.

19.3.1 Determining Relevant Variables

One of the first discussions we had when reviewing the data was which categories to include, modify, or ignore in our analysis. Some of this was predetermined by the lack of consistency or unusable skewedness of the initial data from a processing standpoint. For instance, using city data was not feasible because there were too many individual cities and when grouped by another category (say state or country) they still remained skewed (largely in Massachusetts or the US, respectively). This raises a question of whether this information is then inherently not useful or if it is simply defined within the context of the dataset in question. It seems that it is almost always the latter case, that is, certain types of data are more or less useful based on the total consideration about the dataset in which they exist.

However, one point of consideration that put this theory into question was our consideration about the categories of gender and race. One of the main reasons we considered taking gender out of our model initially was because it was heavily skewed toward males (70.6%) versus females (29.4%). Before we implemented the over-sampling solution using SMOTE and Borderline-SMOTE, we had limited options for dealing with this imbalanced data. A simple, brute force remedy would have involved taking the imbalanced predictor out of the model all together. We discussed the relevance of gender in this case, if it would even impact outcomes.

There was some disagreement about the extent to which gender would be a useful measure of impact for the CDC, largely because it is something which the CDC could do little to influence if it were a strong predictor of outcomes. In essence, it did not appear to be what the CDC was "looking for"in the sense that it was not a modifiable variable within the CDC's scope of programs and events. It evolved from a conversation about reducing noise and making the model work into a discussion about the role of gender within this situation.

To say that Career Fairs are significant predictors and that lower use tends to predict the Not Seeking/Serving response would suggest that the CDC has a justification for making Career Fairs more accessible, more publicized, or further investigating what makes them important for job outcomes. But to say that gender is relevant and predictive of outcomes such that being male tends to predict Full Time Job more than being female would be of little value to the CDC, and might be better taken into account at a larger, administrative level by the university.

A similar discussion was had about race. Initially this was because our original dataset did not have comprehensive information about individuals'

race. But the internship and co-ops dataset we received did, and intersecting them to roll racial data into the model was feasible. Data on race was far from uniformly distributed, but was still less imbalanced than gender. Nonetheless, this was another category where we considered not including it in the model. The line of reasoning was more in parallel to the relevancy question on gender. The question we asked ourselves was whether or not racial categories would make sense as a predictor that the CDC could influence. A slightly more nuanced conversation seemed to imply that there was some uncertainty about what it would mean to include race both in terms of maintaining anonymity and suggesting race-based correlations.

19.3.2 Creating Major Predictor Classes

Because WPI offers over 50 different majors and our group wanted to capture the full spread of major disciplines in our model, we had to group them based on similarities. This includes both primary majors and secondary (double) majors. The following are the predictor categories that we created:

1. Mathematics
2. **Aerospace Engineering**
3. **Civil Engineering**
4. Science
5. **Biomedical Engineering**
6. **Chemical Engineering**
7. **Robotics Engineering**
8. Computer Science
9. Liberal Arts
10. Electrical And Computer Engineering
11. Materials
12. Business Engineering
13. **Mechanical Engineering**
14. **Interactive Media and Game Development**

15. **Fire Protection Engineering**

16. Business

Of these categories, eight only contained a single major. They are bolded in the list above. The rest included multiple majors combined under a single category identifier. The rationale and some explanation of each's limitations follow.

Mathematics included Actuarial Math, Applied Math, Applied Statistics, Financial Math, and Mathematical Sciences. This category was designed to include all undergraduate and graduate-level mathematics majors. Although there is a large distinction between applied mathematics and theoretical fields, all of the above majors are mathematically rigorous and have significant overlap. For instance, the program track for Actuarial Mathematics and Mathematical Sciences are virtually identical apart from some additional requirements specific to actuarial studies and theoretical mathematics specific to each major, respectively.

Civil Engineering contained majors within the same department of Civil & Environmental Engineering at WPI. These were Architectural Engineering, Civil Engineering, Construction Project Management, Environmental and Sustainability Studies, and Environmental Engineering. The university essentially defines these disciplines as related by being under the purview of the same department, so there was little contention grouping them together this way.

Science included the general science majors of Bioinformatics and Computational Biology, Biology and Biotechnology, Chemistry, Biochemistry, Physics, Bioscience Administration, and Engineering Physics. The rationale for this category was to consider physics, chemistry, and biology to be "general sciences,"distinct from applied engineering disciplines. One significant drawback of this is that while the three are related insofar as they overlap in practice (the foundation of chemistry is physics, and many biological processes are inherently chemical), they are distinct fields with more nuance than just being general science. In terms of content, experiment design, and applications, physics and biology are vastly different.

Computer Science grouped together Computer Science, Information Technology, and Management Information Systems. Generally, Management Information Systems (MIS) is classified within the school of Business. However, the major does include a strong focus on computer science and technical analysis.

Liberal Arts encompassed one of the largest groups of majors, including

Humanities and Arts, Professional Writing, Economic Science, Psychological Science, Society Technology and Policy, International Studies, Masters for Physics Educators, Master of Mathematics for Educators, and Learning Sciences and Technologies. This category combines major groups from the Humanities and Arts department as well as the Social Sciences department. There is reason to question at what point a category such as this becomes too weighted down with different sub-categories when one considers the presence of psychological sciences, professional writing, and educational majors in one group.

The Materials category was based on Materials Science majors, the two majors being Materials Process Engineering and Materials Science Engineering. Both of these have a strong relationship in terms of scope.

Business Engineering encompassed majors related to engineering in business, such as Industrial, Management, and Manufacturing Engineering, as well as Manufacturing Management. These majors all share strong relevance and are consistently grouped together within the school of business as WPI.

Business is distinct from the above category as it does not explicitly deal with business in an engineering setting. These disciplines included MBAs; Marketing and Technological Innovation; Management; and Operation, Design and Leadership. As above, these are well-grouped and consistent with WPI's own standards for discerning areas within its business department.

19.3.3 Re-Classifying Military Service

Toward the end of the project, we found that one reason that our Borderline-SMOTE algorithm was taking so long to process and why our results were not quite hitting a high threshold of accuracy was because of a categorical dilemma. Our original "Not Seeking/Serving" class, which covered those students who were serving in the military or neither seeking neither employment nor graduate school, was too small. The proposition was to create a category "Unknown/Not Paid" to encompass students who did not have a job, volunteer, or did not provide any additional information for primary or secondary status.

There were, however, still fourteen students serving in the military. The team concluded that because ROTC students sign a five-year contract to serve in the military after graduation, that their pathway did not make sense to include in our model. The rationale was that our model is designed to predict outcomes with a high accuracy from a selection of predictors that could be used to formulate a profile of successful and unsuccessful students, relative to their postgraduate outcomes. But in this sense, students serving

in the military in this five year period would have their outcomes predetermined by the nature of their declaration at the beginning of their academic careers. So we removed them entirely from the model. This decision resulted in our code running faster and our accuracy increasing significantly, but there remains the concern that the CDC still has an interest in working with ROTC-enrolled students, and that there might be as problematic an implication to removing them from the model, as there was for race or gender when those facets were considered as well.

20 Rhetoric and Data Science

There is a strong intersection between data science processes and rhetoric to examine. To conclude this discussion, we will consider this relationship in terms of the epistemological questions about machine learning and data science as tools for discovering and explaining truth.

20.1 Data Science, Rhetoric, and the Truth

Although a unified view of the use and application of rhetoric is difficult to pin down, scholars and authors have viewed one of its functions as being means of conveying truth. On the matter of the artful' use of rhetoric for speechmaking in his work Phaedrus, Plato speaks through an anecdotal encounter involving his mentor Socrates to outline what comprises such artful rhetoric. The first of these is to "know the truth concerning everything you are speaking or writing about" [37]. This distinction was historically as a facet of Aristotelian theories of rhetoric made in ancient times between rhetoricians and so-called sophists, whose intention for using rhetorical methods was not for persuading audiences of the truth but for other, non-truthful purposes [38]. Importantly, rhetoric does not under these definitions serve to discover truth in and of itself. Rather, it is a method of discovering ways to express truth which is derived from some other means [38].

The goal of truth underlies the general fields of mathematics and the professional aspects of data science as well. Recall that one characteristic of professional spaces are that their constituents lay claim to knowledge which is "esoteric and complex" [24]. This knowledge would be considered the truth in the context of being valuable, factual information which exists independently of the motivations or biases of its discoverers and interpreters. A mathematician in the purest sense would value the truth, in this instance, of a proof which demonstrates the existence of a robust way to simplify

solving certain equations. The entire purpose of mathematics is to come to these types of conclusions whose applicability is intimately tied to a discovery which is true.

Although the professional aspects of data science in a rhetorical sense are still not clearly defined with how new the practice is, the element of valuing truth seems to be an inseparable part of data science, especially in an academic setting. This is captured well in the common motif that the data "speaks for itself" [27]. On some level, it could be argued that through machine learning methods, a data scientist does not do much to create novel truths but rather interpret truths which are self-evident in the data that is being analyzed. The formulation and execution of algorithms based on certain assumptions and classifications of the data in question, under this interpretation, serve as means of discovering this truth and interpreting it to be conveyed to a particular audience. Here there is a strong link to the underlying goals and applications of rhetoric. Neither necessarily makes any claim to actively creating truth, but both deal directly with means of discovering, interpreting, and conveying it.

Another point of comparison between the two fields is found in their limitations regarding their pursuit and interpretations of the truth. Both are subject to biases from human actors and stakeholders with motives which might not always align with or benefit from the truth. A rhetorician, though in this sense considered a sophist, might have personal convictions or subjective experiences in conflict with a higher idea of truth. This individual might compromise a rhetorical presentation of the truth which would certainly be influenced by their worldview if he or she is not cautious about their relation to the truth being discussed. The same is true of a data scientist whose expectations about the outcomes of analysis or who is beholden to a stakeholder with a certain set of assumptions and expectations.

The data may speak for itself in theory, but in practice humans always do the talking, both at the beginning and the end of the process. Rarely is a dataset so clean, so well-classified, and so self-evident that even sophisticated methods reveal clear-cut truths without human intervention or assumptions about aspects of the data, especially when it is driven by human activity (for example, when each data point represents a student). All of these interventions put one at risk to frame analysis in such a way that the results speak to their expectations rather than an independent reality.

During our project, we were seeking what would be considered the truth about what elements of student life are most accurately predictive of post-graduate outcomes. Although we were actively designing a model which would be able to function without our direct input, say for future projects

of this type, we had to constantly make assumptions and adjustments which introduced our own outlooks and biases into the equation, as outlined in the previous section. Some of those interventions were justified in part based on the relationship of particular predictors' relevance and expected outcomes with regards to our sponsor, the CDC. There is an inescapable balance which must be struck between painting a full and truthful picture and creating a model which can be implemented to produce useful and interpretable information. Too much attention to the former can lead to an overfit and functionally useless model which too finely categorizes classes of individuals to the point where a wholly individualized model would not be able to predict generalizable results at all and be purely overfit. However, overreliance on a model which produces a narrow set of results tailored to a stakeholder's immediate and express needs can obfuscate realities that may be hard to confront but still important.

For instance, in deciding to not include racial breakdowns into our model, we justified the decision partially as an effort to preserve anonymity, but also because we did not believe it would be useful for the CDC. But there is an underlying assumption here that, while positive, might bury the lead on a major aspect of success. That is, we essentially assumed race did not matter in this context. It is not easy to grapple with the possibility that one's race might be predictive of their success or failure, but if this is true, that information could be a powerful tool for challenging and resolving racial biases in the university system. It might be beyond the CDC's direct control, but it might also be of great importance in relation to their program success if racial factors are in play.

The 'colorblind' approach might have some benefits, but it enters into a conversation about preconceived notions surrounding the complex role of social factors like race and gender in situations where the assumptions about their relevance appear to stem from an idea of fairness and a notion of equal opportunity, but may in fact be ignoring real and persisting institutional biases against such groups. Of extreme importance is the distinction between considering factors like race and gender as bindingly deterministic through a supervised learning model and further investigating the potential reasons they would influence outcomes. One must ask the question why such factors would be predictive of particular outcomes and be willing to go beyond a cause-and-effect relationship to investigate institutional hurdles uniquely faced by certain groups. Such results, while far from congenial, are of great importance toward making a fairer, better integrated, and more beneficial university experience that maximizes outcomes for students of all backgrounds.

21 Limitations and Future Work

This study was subject to a number of limitations which could have impacted results and may be adjusted in future, replicated attempts to produce outcomes with increased quality. Mainly, the time constraints, scope of the works selected, and the lack of attention to the functional activity within academic and career spaces all contributed to an incomplete picture which could be enhanced through further adjustments.

A longer term study could facilitate a wider discussion related to a wider range of papers, as well as an inquiry into how certain differences show up between professional domains like statistical theory and data science. The review was limited to four selections of journal entries from the past five years as of 2016, two from statistical and probabilistic theory fields within mathematics and two from data science articles, related to and explicitly for the field. The more of each type of article that could be sampled, the more information could be gathered and the more accurate the data could be generated. The conclusions this paper draws are contingent upon the sample size that was surveyed from the existing literature. It remains to be shown if these can be scaled with an increased selection of papers. Furthermore, it would be of great interest to investigate beyond the abstract level. Selected passages from the other components of these papers (introduction, methodology, discussion) could raise interesting questions. For example, do certain sections share more similarities across the fields? Are some much more distinct? How are they organized, is it the same or different?

The actual workings of data science as an occupation and as an academic discipline remain to be outlined in more detail. Lingering questions to consider here include the following: How is statistics taught differently than data science? What do the theoretical approaches to both look like in an academic context? How much more or less important is theory versus applicability in both cases in occupational settings? How certain are claims to knowledge, and what is done to assert them? Who dominates these spaces and how are they organized, if at all? How inclusive or exclusive are the environments and what is done to allow people in or keep people out who have a more novice knowledge base? All of these are relevant questions that would need to be answered to fully determine the professional context of data science and its related fields.

22 Conclusion

It remains undetermined exactly at what stage in professional data science writing is relative to more traditional and professionally characterized mathematical works. From the results presented above, what is clear is that data science is still emerging out of an interdisciplinary origin that continues to divide data science writing between goals-oriented supplements pre-existing industries and a self-sufficient force for knowledge that acts independently of other fields or stakeholders.

One way this conflict arises in the differences of writing are demonstrated by the ways in which confident assertions and definitive claims to knowledge are made in mathematical papers versus in data science articles, where authors, even in data science, are more likely to temper what they are saying and not use technical terminology without sufficient reason. This arises in the writing, through the ways that subjects and verbs are chosen and paired in proximity. Generally, stronger pairings in closer proximity with more active verb choices lend to stronger assertions and more rigid claims. The longer time taken to get from a subject to its verb(s) in these articles tended to be the result of an author lending more explanation and listing out concepts. In the case of data science abstracts, the larger gaps were not attributed to more technical terminology, but a need for more supplementary content for the reader.

That is not to say that there is not jargon and advanced mathematical concepts within data science findings, investigations, and research. However, even when they are relevant to the subject in question, they are seldom frontloaded in the abstract by data science authors. Mathematicians appear more confident in including them at the high level. Although this research did not dive deeply into the actual content of the papers beyond the abstract, the example from the Baxter's inequality paper shows a common theme that mathematics papers tend to have introductions which resemble the starts of proofs, more than a general overview of the topic leading into need-to-know concepts.

This is in large part because of the assumptions authors in their respective fields make about the audiences to whom they are writing. There is an assumed knowledge base for the audience in statistical theory journal articles, so information overload of mathematical lexicon is common. Even when information is being explained for the reader in an historical or mathematical sense, it still requires significant and narrow knowledge from an audience which the authors have to assume are at least interested, if not a part of, their own fields. This is why it is much more difficult to get a

general sense of what an abstract for a mathematical paper is summarizing and what its relevance is, versus the data science articles where a general audience can quite easily deduce meaning. Even if a reader is not aware of what supply chain management is, or knows the mathematical ways to measure data quality, one could likely make sense out of what the two data science articles about these topics are saying. In contrast, it takes a significant amount of work for even an educated mathematician without as much experience in the field to understand all the parts that go into Baxter's inequality or chi-square processes.

Traditional mathematicians are still comfortably making definitive claims and expanding the realm of esoteric knowledge they specialize in to a specific audience on the same plane as they are. Data scientists meanwhile are still staking out their territory and while they are differentiating from their parent fields, they are still largely acting in tandem or even in support of other industries and stakeholders who use their results for their own purposes. Even though data scientists use rigorous mathematics and computer science concepts to generate their results (statistical analysis, machine learning, significance testing, etc.) they are still in a stage where they are advancing knowledge for its own sake to improve their data science methods, while at the same time largely relying on or serving other interests to do so.

The proto-professional nature of data science does not mitigate its relationship to rhetoric's exercise in discovering and conveying persuasive information about the truth to an audience. However, it does potentially exacerbate limitations and roadblocks to successfully doing this. Such bulwarks range from the inevitable biases and complications of human involvement, to the limiting factors of necessary assumptions and classifications to avoid overfitting. By including deeply thought out and well-designed qualitative measures and confronting the implications of human-led design, the value of a model can be enhanced and humanized to speak beyond what the data is saying by itself.

References

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 6. Springer, 2013.

[2] J. Luan, "Data mining and its applications in higher education," *New directions for institutional research*, vol. 2002, no. 113, pp. 17–36, 2002.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[4] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[5] G. Lemaitre, F. Nogueira, D. Oliveira, and C. Aridas, "Smote borderline 2," 2016.

[6] N. Silver, *The signal and the noise: Why so many predictions fail-but some don't*. Penguin, 2012.

[7] S. Koppi, "Post-graduation report class of 2015," tech. rep., Worcester, MA, 2016.

[8] S. Koppi, "Post-graduation report class of 2014," tech. rep., Worcester, MA, 2015.

[9] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, pp. 498–506, 2010.

[10] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv preprint arXiv:1201.3417*, 2012.

[11] J. E. Beck and B. P. Woolf, "High-level student modeling with machine learning," in *International Conference on Intelligent Tutoring Systems*, pp. 584–593, Springer, 2000.

[12] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students'performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.

[13] S. Raschka, "Linear discriminant analysis," Aug 2014.

[14] A. Hildebrand, "Joint distributions, discrete case," 2005.

[15] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research*, vol. 2, no. Nov, pp. 45–66, 2001.

[16] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[17] Python, "Generate pseudo-random numbers," 2017.

[18] D. Hellmann, "Pseudorandom number generators," 2017.

[19] M. Matsumoto and T. Nishimura, "Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator," *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, vol. 8, no. 1, pp. 3–30, 1998.

[20] V. Paruchuri, "K nearest neighbors in python: A tutorial," Jul 2015.

[21] A. C. Muller and S. Guido, *Introduction to machine learning with Python*. O'Reilly Media, 2017.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] G. Press, "A very short history of data science," *Forbes. com*, 2013.

[24] B. Faber, "Professional identities: What is professional about professional communication?," *Journal of Business and Technical Communication*, vol. 16, no. 3, pp. 306–337, 2002.

[25] A. Beaufort and C. Bazerman, "Writing in the professions," *Handbook of research on writing: History, society, school, individual, text*, pp. 221–235, 2008.

[26] R. Agarwal and V. Dhar, "Editorialbig data, data science, and analytics: The opportunity and challenge for is research," 2014.

[27] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, communication & society*, vol. 15, no. 5, pp. 662–679, 2012.

[28] M. J. Salvo, "Visual rhetoric and big data: Design of future communication," *Communication Design Quarterly Review*, vol. 1, no. 1, pp. 37–40, 2012.

[29] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51–59, 2013.

[30] M. Meyer, T. McMurry, and D. Politis, "Baxters inequality for triangular arrays," *Mathematical Methods of Statistics*, vol. 24, no. 2, pp. 135–146, 2015.

[31] D. Pollock, "Wiener-kolmogorov filtering, frequency-selective filtering, and polynomial regression," *Econometric Theory*, pp. 71–88, 2007.

[32] P. Liu and L. Ji, "Extremes of chi-square processes with trend," *arXiv preprint arXiv:1407.6501*, 2014.

[33] P. McClean, "The chi-square test," 2000.

[34] M. A. Waller and S. E. Fawcett, "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management," *Journal of Business Logistics*, vol. 34, no. 2, pp. 77–84, 2013.

[35] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal*, vol. 14, 2015.

[36] B. McNely, "Big data, situated people: humane approaches to communication design," *Communication Design Quarterly Review*, vol. 1, no. 1, pp. 27–30, 2012.

[37] A. Nehamas, P. Woodruff, *et al.*, *Phaedrus*. Hackett Publishing, 1995.

[38] S. F. Crider, *The Office of Assertion: An Art of Rhetoric for the Academic Essay*. Open Road Media, 2014.

[39] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, *et al.*, "Xsede: accelerating scientific discovery," *Computing in Science & Engineering*, vol. 16, no. 5, pp. 62–74, 2014.

[40] C. A. Stewart, T. M. Cockerill, I. Foster, D. Hancock, N. Merchant, E. Skidmore, D. Stanzione, J. Taylor, S. Tuecke, G. Turner, *et al.*, "Jetstream: a self-provisioned, scalable science and engineering cloud environment," in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, p. 29, ACM, 2015.

23 Appendix A

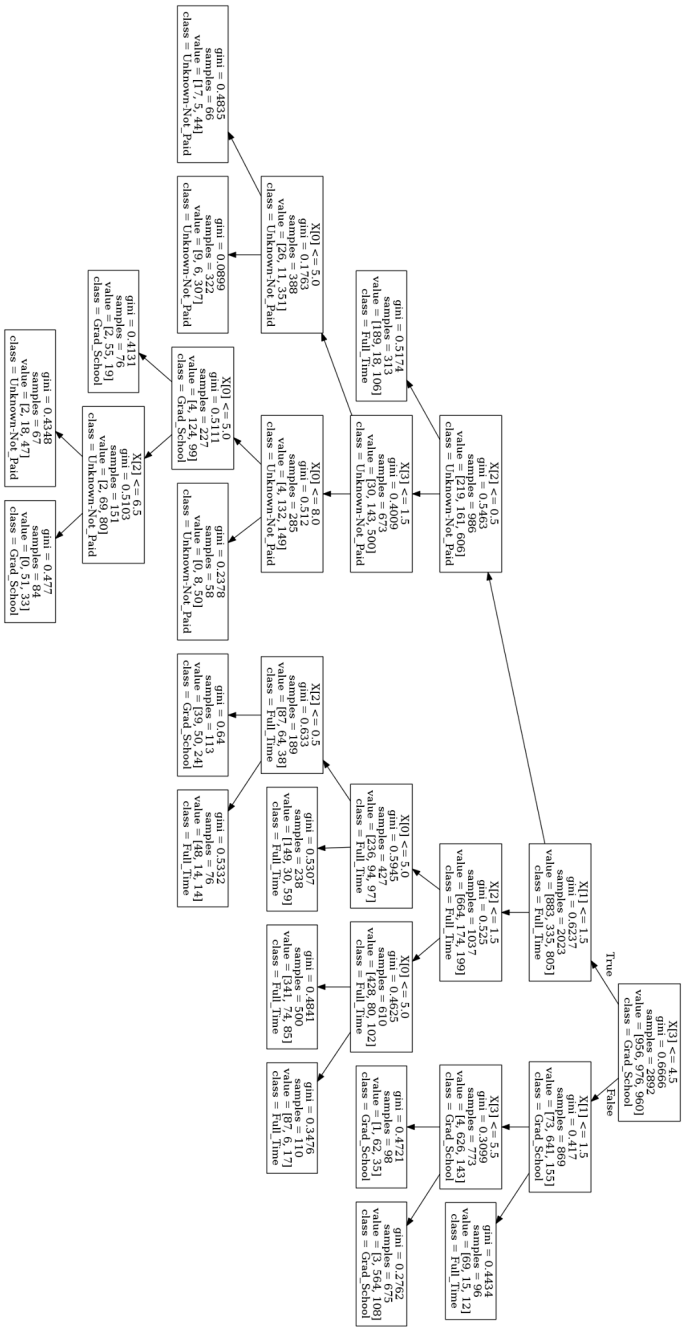

Figure 34: The entire tree made using the top predictors from Borderline-SMOTE from the Class of 2015 dataset

24 Appendix B

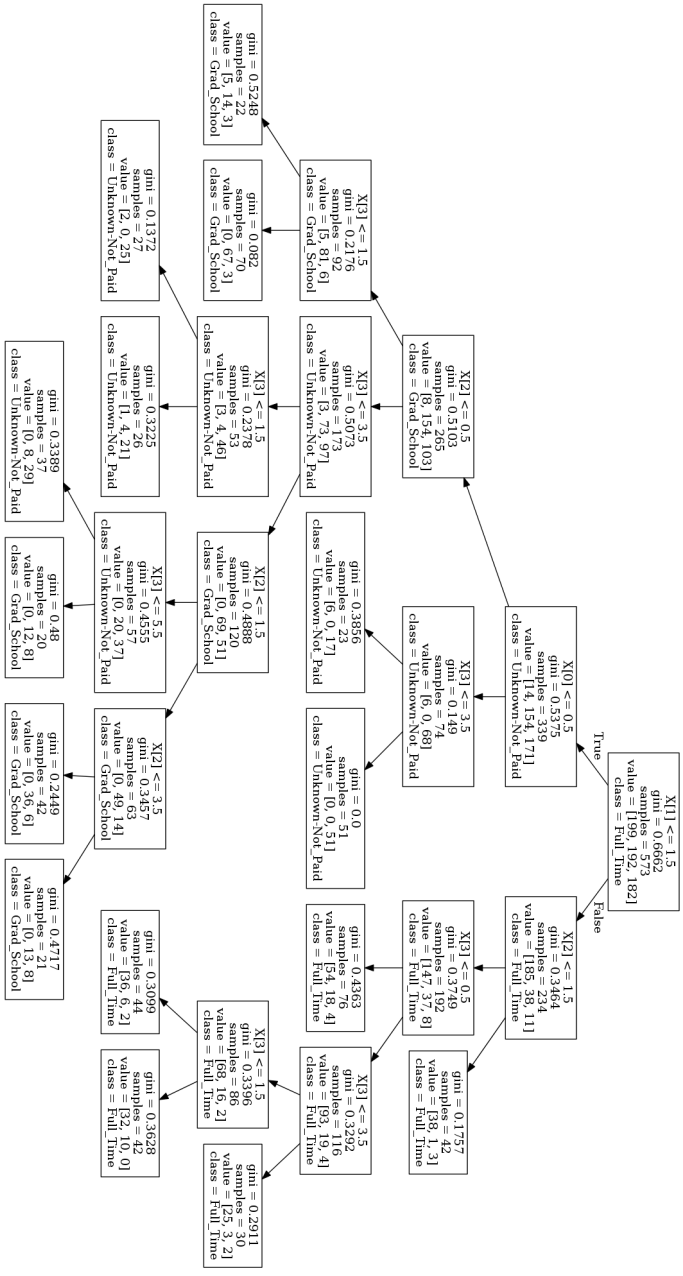

Figure 35: The entire tree made using the top predictors from Borderline-SMOTE from the Internship dataset

25 Appendix C

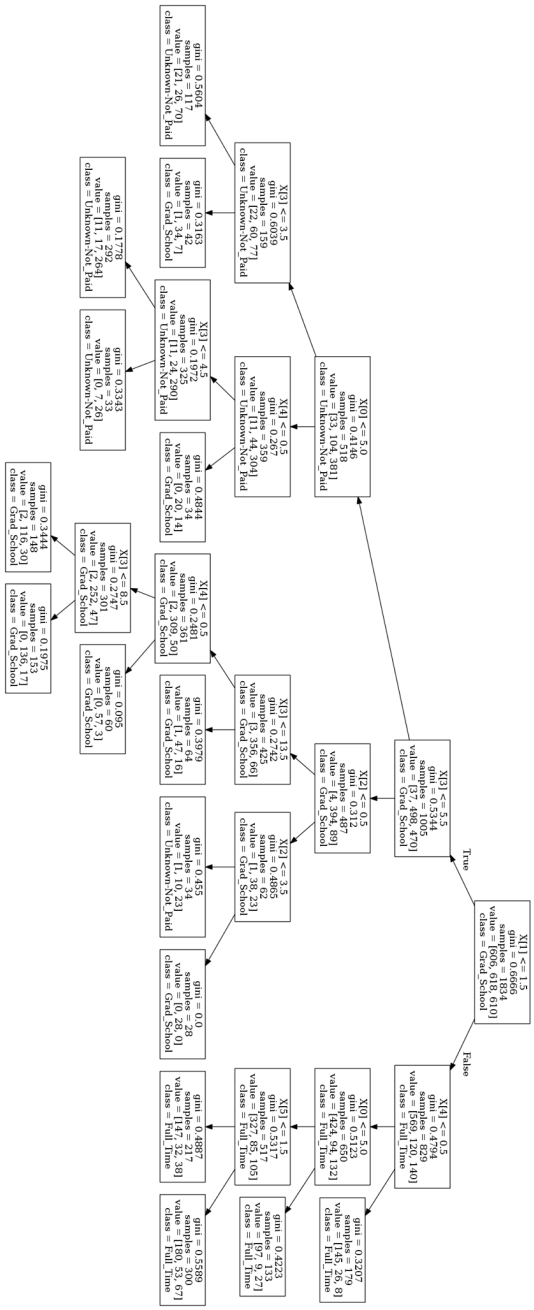

Figure 36: The entire tree made using the top predictors from Borderline-SMOTE from the CDC Usage dataset

26 Appendix D

26.1 XSEDE Supercomputer

Our project required computing large amounts of data which exceeded the processing power of any local device any member of our group had immediate access to. To resolve this, we took advantage of a cloud-based supercomputer which allowed us to greatly reduce program runtime to process our data on a more timely basis for regular analysis and to test code updates more periodically.

26.2 XSEDE and Jetstream Overview

Since our largest active dataset has dimensions 39 x 1676, it is comparatively small in the world of data processing. However, it still proved to be taxing with how many resources were required for computations. To decrease downtime, we began using a supercomputer from the Extreme Science and Engineering Discovery Environment (XSEDE) [39]. Led by the National Center for Supercomputing Applications (NCSA) at the University of Illinois, the organization bridge gaps in accessing state-of-the-art computing resources and offers researchers quality support for project collaboration and completion [39].

The particular cloud environment we used for this project was called Jetstream, which is tailored toward smaller scale, on-demand processing from smaller research teams such as our own [40] [39]. Our team was allocated computing time on the Jetstream environment located at the Texas Advanced Computing Center (TACC). Jetstream uses the specific application called Atmosphere which allows the user to launch their desired template of the virtual computer from a list. These templates are referred to as images [40] [39].

26.3 Working with Jetstream

At the beginning of this project, we began by launching the image called Ubuntu 16.04 LTS. We quickly ran into issues with this image getting stuck in the networking phase of the launch. After submitting a ticket, we received a response from XSEDE Support who explained that since the Ubuntu 16.04 LTS image was not one of Jetstream's "featured images, they could not guarantee it would properly work [39]. After careful contemplation, the XSEDE support staff, along with his colleagues, determined that this bug

was something that is inherent using Ubuntu 16.04 LTS on the Jetstream TACC cloud[39].

After this minor setback, we started running our research on the image called Ubuntu 14.04.3 Development GUI, a Linux-based environment. In order to get data from our Windows laptops into the Linux environment of Jetstream, we used the following tools:

1. **PuTTYgen** allowed us to create both public and private SSH keys in order for Jetstream to recognize our specific user login.

2. **PuTTY** allowed us to SSH into the launched virtual machine.

3. **Bash** on Ubuntu on Windows allowed Linux commands to be executed efficiently on the Windows machines.

4. **Rsync** a Linux command to copy files from the Windows machine to the launched virtual machine.

With the ability to use the supercomputer to process our data, we were fully-equipped to proceed with our analysis.