Optimizing mathematics placement: A machine learning approach comparing predictive algorithmic models

By

Rebekah Coggin

Submitted to the Department of Mathematics and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Master of Arts

Dr. Estela Gavosto, Thesis Advisor

Committee members

Dr. Weizang Huang

Dr. Rodolfo Torres

Date defended: July 14, 2017

The Thesis Committee for Rebekah Coggin certifies
that this is the approved version of the following thesis :

Optimizing mathematics placement:
A machine learning approach comparing predictive algorithmic models

Dr. Estela Gavosto, Thesis Advisor

Date approved: July 14, 2017

Abstract

Choosing the best criteria to place incoming college freshmen into the appropriate first semester courses proves to be a challenge for all subject areas, but for mathematics in particular. It is crucial that universities give students an opportunity to succeed by avoiding placing them in courses with material that is too advanced for them, but just as crucial, if not more, that universities do not place student in remedial classes when they do not need them. In this study we use data from over 21,500 algebra students at a midwestern university over eleven fall semesters to train a variety of machine learning algorithmic models to predict whether or not students will be successful in intermediate algebra and college algebra based on their high school GPA and all four individual components of the ACT. Of these five scores, we find that only GPA and Math ACT are significant predictors of success in algebra courses. We implement algorithms based in optimization, information, and metric space theories. Although they approach this problem with different perspectives, we find they all consistently give similar accuracies on the testing data and similar predictions. The main conclusion of this analysis is that a combination of GPA and Math ACT is the best predictor of success with GPA being the most important factor. We use this information to make recommendations for optimal initial mathematics courses based on an incoming student's high school GPA and Math ACT score.

Acknowledgements

To my advisor, Dr. Estela Gavosto, thank you for going so far beyond merely advising me to being a trusted mentor throughout my time at KU, even before you were my official advisor. Thank you for presenting me with this thesis idea, guiding me through the process, and equipping me with tools for my career outside of academics. Thank you for investing your time into helping me know my strengths, encouraging me to be confident in them, and finding where my strengths intersect with what I enjoy.

To Dr. Ingrid Peterson, the many conversations we have had about algebra students and how to make then successful have motivated and developed this thesis. But my gratitude to you goes beyond this research. Thank you for training me to teach math. Thank you for modeling how to be a humble leader by making every one of your employees feel valued and every one of your students feel known. Your care and compassion for students is evident throughout all you do.

To my committee, Dr. Rodolfo Torres and Dr. Weizang Huang, and professors and KU, thank you for challenging my mathematical abilities and stretching me to become the mathematician I am.

To all the algebra teaching assistants I have supervised over the past three years, all the conversations we have had about the challenges in the classroom and how to help our students succeed have motivated this paper. Thank you also for making my job as supervisor so enjoyable. Even grading algebra for hours on end was not so bad when it was with you all.

To my sweet church, Grace Evangelical Presbyterian Church in Lawrence, thank you for taking me in as family while I was away from mine.

To my dear friends, Julie and Avary, you two kept me laughing through all the challenges of grad school.

To my parents, thank you for teaching me to love learning from a young age. To my siblings, Erin and Wynn, and their spouses, Shawn and Ellie, thank you all for loving and supporting me from far away.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Math Placement

1.1 Current Issues on Math Placement

Placing incoming students into a mathematics course at the appropriate level is a challenge all higher education institutions face whether community college, public university of private school. Institutions advise students on which class is most appropriate for them by with a variety of success indicators, including high school GPA, overall standardized test (ACT or SAT) scores, ACT/SAT math portion scores, placement exams, student choice, or a combination of multiple indicators.

Clearly it is crucial to avoid over placement, which is placing student in courses for which they do not know the background material. This does not give them a chance to succeed. But it is just as critical to avoid under placement, which is placing a students in a course that is not the highest in which the student could succeed. Studies have shown that if students are under placed, they are less likely to continue in math (Bressoud & Hsu, 2015). When students do continue taking math courses, far too often placing students preparatory course does not help them succeed in subsequent mathematics courses (Bressoud & Hsu, 2015; Jaggars & Stacey, 2014). A 2015 MAA study documents several examples of low retention in math courses, even among students who were successful in the remedial courses in which they were placed. These cases include students who have intentions of pursuing a STEM major (Bressoud & Hsu, 2015; Jarrett, 2000).

A 2014 study of 132 institutions considered students who were near the cutoff criteria for

Calculus I. The study compared the Calculus I scores of the students who took Precalculus and then Calculus as opposed to going straight to Calculus. The study showed that taking the Precalculus course does not improve the outcome of Calculus I (Sonnert & Sadler, 2014). As summarized in (Bressoud & Hsu, 2015),

> *In a recent study of over 10,000 Calculus I students across the United States, Sonnert & Sadler (2014) compared grades in Calculus I of students with the same high school preparation in mathematics (courses taken, grades earned, SAT/ACT scores) who either had or had not taken a post-secondary precalculus class. They found that students below the mean composite secondary school preparation score did appear to benefit from precalculus, but by a meager and not statistically significant single point on a 100-point grading scale. They also found that for students above the mean, placement in precalculus lowered their Calculus I grade by a statistically significant average of six points. The meager gains from precalculus do not appear to offset the considerable risk that students directed to precalculus will not persist to Calculus I.*

Another study of students in Texas who were placed into college classes based on placement tests shows there is no indication that remedial courses are of benefit to students (Martorell & McFarlin, 2011). In their paper, Martorell and McFarlin state, "For a wide range of academic outcomes and across a variety of subgroups, the estimated effects of remediation are small in magnitude and statistically insignificant. If anything, we find some evidence that remediation might worsen the outcomes of some students." We can see that it is crucial to place students in the highest level course in which they might be successful.

Typically these remedial courses do not count towards a degree yet students must use their limited time and funding to complete these courses before taking the subsequent required course. If remedial courses are not increasing passing rates in subsequent courses and if schools are seeing low retention rates of students moving from remedial courses, then schools need to seriously consider whether students gain anything from placement in the remedial courses. Remedial courses are not an effective or responsible use of the school or students' resources if students are not

placed optimally. Not only could schools retain more students in mathematics if they are able to place them in the optimal course, but optimal placement would result in the allocation of financial resources to students who have higher probabilities of success.

1.2 Present study: Data from over 21,500 algebra students

In this study, we consider all algebra students at a midwestern university over a period of eleven years. The two algebra courses offered at this school are of Intermediate Algebra and College Algebra College Algebra courses. Each fall, approximately two thousand students at this school enroll in these algebra courses. Under the current system, incoming students are placed in their first mathematics course based only on their score on the math component of their ACT score or an optional placement exam. Students scoring below 18 on the math portion are placed in Intermediate Algebra, students scoring 18-22 are placed in College Algebra and any students who score above a 22 are placed into calculus classes. This placement system often places students in classes that are not at the optimal level – either a class that does not challenge them or a class for which they are not prepared.

Our goal is to find a better method of placing students into math courses. We would like to consider more information about the incoming student, including high school GPA, overall ACT score and other individual ACT component scores beyond just the math component. We experiment with machine learning algorithms to find a method that will place students with the highest chance of success. In the end we will recommend a way of weighting the most important indicators of success so that students will have optimal outcomes in their first year mathematics course.

The algebra program at this school has kept detailed records of each student enrolled in Intermediate Algebra and College Algebra. For this study we choose to use data from students in fall semesters only from Fall 2006 to Fall 2016. With approximately two thousand students per semesters, we have data on over twenty thousand students total.

During the spring semester, students enrolled in Intermediate Algebra are primarily students

retaking the course and students enrolled in College Algebra are primarily moving from Intermediate Algebra or retaking the course. Since we would like to study students enrolled in each course for the first time, we only consider fall semester students, which are primarily students taking their first college level math course. The information from the algebra student data in which we are interested is each student's high school GPA, ACT, individual ACT components and final grade in course. In Chapter 5 we analyze students with a unique record. That is, we consider only each student's first semester enrolled in algebra at this university. We drop all data on students repeating a course or students moving to College Algebra after having taken Intermediate Algebra.

1.3 Placement Methods

A variety of success indicators for mathematics courses exist because it is difficult to develop a single indicator that will serve all students well. In the following paragraphs, we discuss some advantages and disadvantages to several indicators used for placement methods.

While high school math courses and grades are not easily accessible and practical to use to place large numbers of students, high school GPA's are readily available to placement advisors. Critics of putting too much emphasis on high school GPA say that it is an inconsistent measurement because all schools assign grades and thus GPA's differently. This group claims the ACT/SAT is more reliable because it is standardized across all school. Furthermore, they fear that emphasizing GPA hurts minorities and people of low socioeconomic status. However research show that high school GPA is consistently the best predictor of college success in freshman year (Geiser & Santelices, 2007; Scott-Clayton, 2012). This is likely because it is not just a measure of a student's proficiency, but also a measure of "motivation and perseverance," (Bowen et al., 2011), a quality that is otherwise difficult to capture in a meaningful way.

Not only is GPA a better predictor than ACT/SAT, but it actually becomes an even stronger predictor of success in college grades after freshman year. Furthermore, when admission decisions are based on high school GPA it actually helps disadvantaged minorities more than when the ACT/SAT are emphasized (Geiser & Santelices, 2007). The standardized tests results reveal

that underrepresented minorities and disadvantaged students are in fact less prepared for college than people with higher socioeconomic status, and thus fewer are admitted when ACT/SAT is emphasized over high school GPA. (Geiser & Santelices, 2007; Rothstein, 2004)

In addition, placing students based on their ACT/SAT score alone does not give more predictive power than considering the students' GPA alone (Rothstein, 2004). In fact, the portion of SAT (and presumably ACT) scores which gives the more accurate predictions of a student's success in college is the portion which correlates with high school demographic characteristics. The portion orthogonal to high school demographics does not give any further predictive power. (Rothstein, 2004)

In 2011, a study actually showed a negative correlation between grades in Calculus I and ACT scores (Reddy & Harper, 2013). This could partially be because the ACT/SAT was never intended to be a placement exam, but instead a college readiness exam. The broad skill set measured by the ACT/SAT is too general to be used in placing students in specific classes (Bressoud & Hsu, 2015).

Furthermore, in a study of college drop out, George Bulman found that a student's high school GPA carries a lot of information, but claims that adding ACT/SAT scores as a predictor does not add any predictive power (Bulman, 2017). If this is the case for predicting whether a student will drop out or complete college, we should also consider high school GPA when placing students into initial math courses rather than making decisions based on ACT scores alone.

Since high school GPA and ACT/SAT scores are not often effective placement tools, many schools administer other standardized tests specifically for placement. These are online tests which adapt the questions given to a student based on which questions the student has answered incorrectly. It continues narrowing the questions until it converges on the student's mathematics level. Two such standardized placement test are the ACT Compass exam, which was phased out at the end of 2016, and the College Board Accuplacer exam.

According to a paper examining the effectiveness of these placement exams by (Scott-Clayton, 2012),

...the incremental validity of placement tests relative to high school background predictors

of success is weak, even in math. Adding test scores to a model using high school GPA/units to predict college-level grades increases the proportion of variation explained by about 6 percentage points in math...using high school GPA/units alone as a placement screen results in better outcomes than using placement test scores alone ..., and adding in placement test scores results in little additional improvement.

Furthermore, by considering multiple indicators for placement rather than only standardized tests, schools could reduce the number of students places into remedial courses by 8 to 12% while maintaining or increasing success rates in college level courses. For these reasons, ACT phased out the Compass exam at the end of 2016. As explained by ACT spokesman, Ed Colby, "A thorough analysis of customer feedback, empirical evidence and postsecondary trends led us to conclude that ACT Compass is not contributing as effectively to student placement and success as it had in the past," (Fain, 2015).

Another commonly used placement test is ALEKS, which stands for Assessment and LEarning in Knowledge Spaces. ALEKS is also an online test. It is an artificially intelligent learning system which quickly assesses exactly what the student understand and does not. This placement test differs from Compass and Accuplacer because it is free response only while the other two are multiple choice. This means that the feedback is not immediate, but often more insightful. ALEKS' learning process includes reassessment of material which the students previously answered incorrectly. ALEKS provides tutoring to help the students review material on their own and reassess. It also different in that it employs theory of Knowledge Spaces. As explain in (Bressoud & Hsu, 2015), "This theory is used to map a mathematical subject area such as the mathematics of Algebra I to a set of items, essentially problem types, and to capture the hierarchical structure of dependence among these items. It is this understanding of the dependence structure of problem types that enables it to drill down and identify the most basic level at which a student is experiencing difficulty."

The remediation component of ALEKS avoids the static cutoffs of other placement exams (Hodara et al., 2012). The responsibility of background material falls to the student with support

of review resources. Three years of data on students at the University of Illinois shows "very high correlations between mean grades over small ALEKS score ranges and range midpoints. This greatly outperforms the former placement policy. Similar correlations for ACT math scores were generally much lower, less consistent year to year, and sometimes negative." (Reddy & Harper, 2013)

Futhermore, "Data analysis indicates that ALEKS scores and some subscores correlate well with final grades and that the ALEKS-based placement program lowered failure and withdrawal rates in nearly all the placement classes in each semester." (Reddy & Harper, 2013)

Even though ALEKS is a powerful placement tool, mathematics placement exams alone are not sufficient in placing students into an appropriate mathematics course. A study of Connecticut high schoolers entering directly into college showed that other statistically significant predictors of success were number of years of mathematics, course level and gender (Moran, 2008). While placement exams are significant predictors, including high school GPA, ACT/SAT scores, and high school math records and other non-academic variables in addition to placement exams will decrease false positive rates (Drake, 2010).

In considering these various studies, we may conclude that there is no single indicator which contains all the information needed to place students. Instead, optimal placement for the highest success rates in mathematics courses requires a combination of indicators.

1.4 Innovation of present study: Training with students' background attributes

In most education studies (and all the studies referenced in this chapter), statistical analysis is done by traditional data models. However in the past several years, an explosion of research has shown that machine learning can be an extremely effective tool in building algorithmic models for predictions (Breiman, 2001). In our study we bring new insight in to the math placement problem by employing machine learning methods outlined in Chapter 2.

As Breiman states in his comparison of data models (models from traditional statistical analysis) and algorithm models (models from a machine learning approach), "An algorithmic model can produce more and more reliable information about the structure of the relationship between inputs and outputs than data models." In the end, it is the relationship between the high school GPA/ACT scores and outcome of first year college math course which we would like to know. These types of algorithmic models based on machine learning algorithms give a variety of ways to explore this relationship.

Some models like, logistic regression and decision trees, give high interpretability so that we can gain a lot of insight about what information is most important, but this usually comes at a cost of a slight loss of accuracy. Other models, like random forests and support vector machines, have almost no interpretability. We give the models the input and the algorithm gives a classification. These results do not come with probabilities and variables weights like models with higher interpretability. However we can usually increase accuracy using these models. In this study we will explore both extremes of the trade off between interpretability and accuracy in order to gain as much information as possible about predicting an incoming student's probability of success in College Algebra or Intermediate Algebra.

Variables used in training: Even though references above show that ACT score is not the optimal placement tool, because of ease of access, many schools, include the university in this study, still choose to use ACT scores alone as a placement tool. As mentioned above, the ACT was meant to be a college readiness assessment, not a placement tool, but since students must take the ACT exam to apply to colleges, universities have the scores on recored an no new tests or information is required (Bressoud & Hsu, 2015). If schools do choose to use the ACT for placement because they are not able to use ALEKS to evaluate the optimal starting math course, they can still consider how to take advantage of the all information the ACT does have to offer in regards to placement.

While it may not be intuitive initially, studies have shown the English score on standardized test are helpful indicators of success in remedial math courses. In a study of students in colleges across

Ohio, researchers Bettinger, Evans and Pope focused on correlations between college outcomes and individual components of the ACT. They found

> *...a strong correlation between higher ACT composite scores and positive college outcomes. However, this overall correlation masks an important pattern: Mathematics and English scores are much more tightly correlated with college success than are Reading and Science scores. In fact, after controlling for Mathematics and English scores, Reading and Science provide essentially no predictive power regarding college outcomes (Bettinger et al., 2013).*

More specifically, the model the researchers built "predicts that a student who gets an ACT composite score of 24 by getting a 26 each on the Reading and Science tests and a 22 each on the Mathematics and English tests is 59 percent more likely to be a first-year dropout and 43 percent more likely to drop out by the third year of college, relative to a student who gets the same ACT composite score of 24, but with a 26 each on the Mathematics and English tests and a 22 each on the Reading and Science tests." (Bettinger et al., 2013)

As discussed in Section 1.3 above, GPA is a far better indicator of success in college than ACT scores. Incidentally, ACT math and English scores are also far better predictors of high school GPA than ACT reading and science scores. Considering math and English scores and not science and reading would have a significant impact in the way school admit students. The researchers estimate that with this transition, up to 55% of students in Ohio would be attending a different level school that the one they are currently attending. Furthermore, the top schools in the state could reduce drop out rate by up to 7% by implementing this change (Bettinger et al., 2013).

A similar study was done in Texas using the Texas Academic Skills Program (TASP) scores. Researchers were interested in knowing if components of the TASP beyond math would be significant indicators of first year mathematics course outcomes. They found that reading scores were in fact significant indicators of success in elementary algebra courses but not in intermediate or college algebra. Thus for lower level math courses, reading score could help make more accurate placements. In all courses considered in the study, the combination of reading and math placed students into courses equally as well as math alone (Preast, 1998).

Based on studies like these, we use individual components of ACT scores along with high school GPA to train our placement models.

Part of this work will be further developed and published elsewhere in a joint article with Dr. Estela Gavosto.

Chapter 2

Methods of Machine Learning

Based on previous similar work which uses machine learning to answer questions about college education Marbouti et al. (2016), we choose eight algorithms to predict whether students will pass (receive a grade of A, B or C) or not pass (receive a grade of D, drop, fail or withdraw) based on their high school GPA, overall ACT exam score and scores of individual components of the ACT exam. Machine learning algorithms are developed from several different theoretical approaches. We chose models to come from a variety of these approaches. Logistic regression, linear discriminant analysis, and quadratic discriminant analysis are all developed from optimization theory. Both K nearest neighbors and support vector machines take a more geometric approach based in metric theory. Decision trees and random forests come from information theory in that their goal is to find a minimal structure which differentiates the data. In this section we give a brief overview of how each algorithm works and how we will implement it to analyze our data on algebra students.

2.1 Data Preprocessing

Throughout this section, we represent the data from one student by $\mathbf{x}$, which is a vector of features and we represent the target outcome (pass or fail) by y. We choose the class $y = 1$ to represent a passing grade and the class $y = 0$ to represent a grade of D or F, a drop, or withdraw. The feature vector $\mathbf{x} = (x_1, \ldots, x_n)$ holds numerical values of features. The data base housing data on algebra

students has information on students before beginning college, such as high school GPA and ACT scores as well as every exam score, homework scores and attendance record the student received while enrolled in an algebra course. Since we would like to optimize the placement of students, in this study, we will only use information about the students before they begin algebra courses.

In order to check the accuracy of our models, we save data for validation and testing. We reserve 20% of the data for testing data, that is, it is used for error analysis only. This data will not be seen in the training process. We also reserve 20% for the data for validation data. It not be used in the initial training, but will be used for fine tuning parameters. We train the model on the remaining 60% of the data. We will randomize the divisions in the data so that each portion has data from all eleven semesters we are analyzing.

Before training process begins, we check for abnormalities in the data. A small number of the students (less than 10 over the 11 semesters) reported GPA scores on a 100 point scale. These students were removed. A large number of students did not report ACT scores. These students were mostly international students, transfer students, or students from outside the midwest who reported SAT scores. We remove the students with no ACT score from our study. The remaining number of students in our study is 21607. Of these, 8874 were enrolled in Intermediate Algebra and 12733 were enrolled in College Algebra.

Before training each model, we normalize the input features with a mean of zero and a standard deviation of one. This is necessary because the grades and scores are on a variety of scales and we do not want to unintentionally weight them differently. Intermediate Algebra and College Algebra students are standardized separately.

In some models that we build, the outcome will be binary, that is the outcome has two possible classes: one class is earning an A, B, or C and a second class is a grade of D, drop, withdraw, or fail. However, other models actually predict the probability of being in class 1 or class 0.

2.2 Logistic regression

Logistic regression is a discriminative classifier. Similar to linear regression, it computes an optimal weight vector β and take the product with the feature vector $\beta^{\mathbf{T}}\mathbf{x}$ to make predictions. For binary classification problems, with classes 0 and 1, we would like to have a probability of the data begin in class $y = 1$. This is accomplished by scaling the predictions so that they are between 0 and 1. We can then choose a threshold between 0 and 1 (usually 0.5) such that if the probability is greater than the threshold, we assign $\mathbf{x}$ to class 1, otherwise class 0. Note that a threshold of 0.5 is equivalent to rounding to nearest whole number to determine the class. We use the sigmoid function

$$sigm(x) = \frac{1}{1+e^{-x}}$$

to accomplish this scaling to between 0 and 1. Thus the probability of being in class 1 is given by

$$h(\mathbf{x}) = sigm(\beta^{\mathbf{T}}\mathbf{x}) = \frac{1}{1+e^{-\beta^{\mathbf{T}}\mathbf{x}}}$$

and the probability of being in class 0 is $1 - h(\mathbf{x})$.

The learning problem in logistic regression is optimizing β. As shown in (Murphy, 2012), logistic regression corresponds to the model

$$p(y|\mathbf{x},\beta) = \text{Ber}(y|sigm(\beta^{\mathbf{T}}\mathbf{x}))$$

where Ber is the Bernoulli distribution. We can multiply the probabilities over all the training data to formulate the maximum likelihood estimator,

$$\prod_{i=1}^{N} sigm(\beta^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})})^{I(y^{(i)}=1)} \cdot \left(1 - sigm(\beta^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})})\right)^{I(y^{(i)}=0)}$$

where I is the indicator function. Taking the log of this expression and negating gives a negative

log likelihood of

$$NLL(\beta) = -\sum_{i=1}^{n} \left[y^{(i)} \log \left(sigm \left(\beta^{\mathbf{T}} \mathbf{x}^{(i)} \right) \right) + \left(1 - y^{(i)} \right) \log \left(1 - sigm \left(\beta^{\mathbf{T}} \mathbf{x}^{(i)} \right) \right) \right] \tag{2.1}$$

where $\mathbf{x}^{(i)}$ is the feature vector and $y^{(i)}$ is the target of the training data point i. We would like to minimize this function over β based on our training data.

Because this equations is not in closed form (Murphy, 2012), we must used unconstrained optimization methods to minimize the negative log likelihood. Common methods for this are gradient descent and Newton's method (Ng, 2016). Both require choosing a step size α. We train on the training data and use the validation data to optimize α based on the number of iterations we choose.

2.3 Linear and Quadratic Discriminant analysis (LDA and QDA)

Similar to Naive Bayes Classifier, quadratic and linear discriminant analysis are generative classifiers. Though the names are misleading, discriminant classifiers model $p(y = c|\mathbf{x})$ and simply map the data directly to the targets, while generative classifiers model $p(y = c, \mathbf{x})$, which gives information about how the data is generated. For generative classifiers, we calculate the class conditional density $p(\mathbf{x}|y = c)$ and use Bayes' Rule to estimate the posterior

$$p(y = c|\mathbf{x}) \propto p(y = c)p(\mathbf{x}|y = c).$$

Geometrically, LDA and QDA define an ellipse for each class c which includes a certain percentage of the training data that falls into that class, usually 95%. To define these ellipses, we compute the center of each class μ_c for the centroid of the ellipse. We also compute the covariance matrix for each class Σ_c. Since Σ will be a positive semi-definite matrix, we can perform singular value decomposition to find the eigenvalues and eigenvectors. These eigenvectors will give the axes of the ellipse and the eigenvalues gives how much far the ellipse is stretched along the axes.

Both LDA and QDA assume the class conditional density has a Gaussian distribution allowing us to compute the probability of the data falling into each class ellipse

$$p(\mathbf{x}|y=c) = \mathcal{N}(\mathbf{x}|\mu_c, \Sigma_c) = |2\pi\Sigma_c|^{-1/2}\exp\left(\frac{-1}{2}(\mathbf{x}-\mu_c)^T\Sigma_c^{-1}(\mathbf{x}-\mu_c)\right).$$

To classify an unseen data point $\mathbf{x}$, we use the proportion given by Bayes' Rule above. Let $p(y = c) = \pi_c$. Usually we compute $\pi_c = \frac{N_c}{N}$ where N is the total number of training examples and N_c is the number of training examples in class c. We compute $p(y = c|\mathbf{x})$ for each class c and choose the class with the highest probability. Since we are only choosing the highest, taking the log doesn't change the class we choose, but make the computation much cleaner. Then the classification problem becomes

$$\begin{aligned}\hat{y}(\mathbf{x}) &= \text{argmax}_c\left[\log(p(y=c)) + \log(p(\mathbf{x}|y=c))\right] \\ &= \text{argmax}_c\left[\pi_c - \frac{1}{2}\log(2\pi\Sigma_c) - \frac{1}{2}(\mathbf{x}-\mu_c)^T\Sigma_c^{-1}(\mathbf{x}-\mu_\mathbf{c})\right].\end{aligned}$$

Notice the middle term is constant over the new data $\mathbf{x}$, thus we can equivalently write

$$\hat{y}(\mathbf{x}) = \text{argmax}_c\left[\pi_c - \frac{1}{2}(\mathbf{x}-\mu_c)^T\Sigma_c^{-1}(\mathbf{x}-\mu_\mathbf{c})\right].$$

If the priors are constant across classes, we may drop the first term as well. Since this leaves a leading negative, we minimize over the opposite and again drop the meaningless constant of $\frac{1}{2}$

$$\hat{y}(\mathbf{x}) = \text{argmin}_c(\mathbf{x}-\mu_c)^T\Sigma_c^{-1}(\mathbf{x}-\mu_\mathbf{c}).$$

This result is the classifier for QDA. LDA makes the assumption that the covariance matrices are equivalent across classes $\Sigma_c = \Sigma$ to further simplify the computation to

$$\hat{y}(\mathbf{x}) = \text{argmin}_c(\mathbf{x}-\mu_c)^T\Sigma^{-1}(\mathbf{x}-\mu_\mathbf{c}).$$

2.4 Decision Tree

The name of the decision tree algorithm comes from the tree-like structure that arises in the way the data is split based on a series of binary questions. These splits at binary questions are called nodes. All incoming data starts at the same node, the 'root' of the tree. If the answer is yes, the data follows on path and if not, it follows the other path. Both paths lead to a second node and is split again based on the answer to this new node. Each node uses only one feature. The most important features are used early in the tree, at or near the root. The less influential features are utilized for fine tuning at the ends of the branches or often not used at all. A target is assigned at the very tip of each branch.

At the root of the tree, an algorithm uses all the data to assign a score to each feature. The feature with the highest score is used to split the root. As we progress up the tree, the algorithm to select a feature for each node is repeated, but only the subset of the data that is directed to that node is used in the algorithm.

One major difference from what we have seen in other algorithms is that the decision tree requires binary or discrete features, not continuous real valued features. Clearly the algebra student data consists of real valued features. In order to implement the decision tree, we must convert these to binary. This is not difficult by splitting each feature into multiple features. Consider the high school GPA feature for example. It holds a real value between 0 and 4. We can choose to split this real valued feature into six binary features by asking if the real value is greater than or equal to a certain value between 0 and 4. Then the first of the six new features may be "the student's high school GPA is greater than or equal to 0.5" and take the value 0 if the statement is false and 1 if the statement is true. The chart below summarized how this process works. The first column gives the original feature value and the follow six columns give the corresponds binary values the six new features take based on the original feature value.

Original feature value	≥ 0.5	≥ 1	≥ 1.5	≥ 2	≥ 2.5	≥ 3	≥ 3.5
3.78	1	1	1	1	1	1	1
2.34	1	1	1	1	0	0	0
1.75	1	1	1	0	0	0	0

2.5 Naive Bayes Classifier

The Naive Bayes approach to machine learning makes use of Bayes' Rule which gives the posterior probability as the product of the class prior probability and the likelihood.

$$p(y = c|\mathbf{x}) = \frac{p(\mathbf{x}|y = c) \cdot p(y = c)}{p(\mathbf{x})}$$

where y is the outcome and $c \in C$ is one of the possible classes C the outcome can fall into.

Navie Bayes further makes the 'naive' assumption that all features are conditionally independent from one another, that is

$$p(x_j, x_k|y = c) = p(x_j|y = c)p(x_k|y = c)$$

for all $j,k \in \{1,\ldots,n\}$. This impacts the likelihood and the equation above becomes

$$\begin{aligned} p(y = c|x_1,\ldots,x_n) &= \frac{p(x_1,\ldots,x_n|y = c) \cdot p(y = c)}{p(\mathbf{x})} \\ &= \frac{p(x_1|y = c) \cdot (x_2|y = c) \cdot \ldots \cdot p(x_n|y = c)p(y = c)}{p(\mathbf{x})} \\ &= \frac{\left(\prod_{i=1}^{n} p(x_i|y = c)\right) p(y = c)}{p(\mathbf{x})} \end{aligned}$$

We now have a way of calculating the posterior probability $p(y = c|\mathbf{x})$ given the priors $p(y = c)$ and $p(\mathbf{x})$ and the likelihoods $p(x_i|y = c)$. To classify a new data point $\mathbf{x}$, we simply calculate $p(y = c|\mathbf{x})$ for each $c \in C$ and choose the highest resulting probability.

It is common practice (Murphy, 2012) to use the Gaussian distribution to estimate $p(x_i|y = c)$.

Given the training data, we select only data points which result in $y = c$. From these points, we calculate a mean μ_{ic} and standard deviation σ_{ic} for each feature x_i for $i \in \{1, \ldots, n\}$. The we can compute

$$p(x_i|y = c) = f(x_i)$$

where $f(x)$ is given by the probability density function of the normal distribution $\mathcal{N}(\mu_{ic}, \sigma_{ic})$.

To calculate $p(y = c)$ from the training data, we simply divide $p(y = c) = \frac{N_c}{N}$ where N is the total number of data points in the training data and N_c is the number if data points in the training data which result in $y = c$. To further simplify the computation, since $p(\mathbf{x})$ will be a constant scalar for each new data point $\mathbf{x}$ which we would like to classify, it will not influence our choice of class. That is, since

$$\begin{aligned} &\frac{p(y = c_j)\prod_{i=1}^{n} p(x_i|y = c_j)}{p(\mathbf{x})} < \frac{p(y = c_k)\prod_{i=1}^{n} p(x_i|y = c_k)}{p(\mathbf{x})} \\ &\implies p(y = c_j)\prod_{i=1}^{n} p(x_i|y = c_j) < p(y = c_k)\prod_{i=1}^{n} p(x_i|y = c_k) \end{aligned}$$

we have no need to compute the predictor prior probability in the denominator. We simply compute the numerator and assign the new data $\mathbf{x}$ to the class resulting in the highest probability.

2.6 K Nearest Neighbor

The K nearest neighbor classifier (KNN) compares the features of new data with the features of all the training data. It chooses the K most similar training data points N_K and looks at the classes these K training data points fall into. The algorithm classifies the new data by simply choosing the class that appears most frequently in the N_K training data points. Let Y_K be the set of classes that the N_K data takes. The likelihood can be written in terms of the Kronecker delta functions as

$$p(y = k|x, D, K) = \frac{1}{K}\sum_{y \in Y_K} \delta_{yk}.$$

The algorithm selects N_K by treating the data features as vectors and computing the euclidean distance between the new data point $\mathbf{x_{new}}$ and each training data point $\mathbf{x_i}$. We keep the K training data points with the smallest distances

$$dist(\mathbf{x_{new}}, \mathbf{x_i}) = \left(\sum_j (x_{new,j} - x_{i,j})^2 \right)^{\frac{1}{2}}$$

where i denotes the training data point and j denotes the feature.

The algorithm may be straight forward, but choosing a good value for K is not. If we choose a value too small, we overfit the training data. If we choose a value too large, we over-simplify the training problem and assign all new data to the most popular class in the training data.

KNN makes the assumption that all nearby data have the same target, which may not necessarily be the case, but we hope it is true or learning may not be possible at all. In contrast to decision tree, KNN weights all features evenly. This is usually not helpful. In many situations, some features are much more important in determining the target than others. KNN does not have the power to make use of this fact. However since we have only three features in the algebra student data, this should pose a problem for our research.

2.7 Support Vector Machine

There are many techniques to find an optimal hyperplane which separates completely separable data. The optimal separating hyperplane is defined to be the hyperplane which separates the data and maximized the distance between the plane and the closest data point from each class. In the two-class case, let the classes be represented by -1 and 1. Let β be the vector of coefficients defining the optimal hyperplane, with β_0 being the intercept. Assuming $(\mathbf{x}_i, y_i)$ be the i^{th} training point where $\mathbf{x}_i$ is a vector of features and y_i is the class to which the i^{th} training example belongs. We can then turn the optimal hyperplane question to an optimization problem as follows

$$\max_{\beta, \beta_0, ||\beta||=1} M$$

$$y_i(x_i^T\beta + \beta_0) \geq M,\ i = 1,\ldots,N.$$

This is a convex optimization problem, so typically it is solved with Lagrangian multipliers. Details can be found in Section 4.5.2 of reference Hastie et al. (2009).

In the optimization problem above, M is called the margin and is the distance between the hyperplane and the closest data points. Thus $2M$ is the width of path between the data. The data points which lie exactly on the margin are called support vectors.

The support vector machine (SVM) provides a way to extend this separating hyperplane idea to data that is not separable. It defines a was of measuring the overlap and minimizing it. We define a new slack variables $\xi_i \geq 0$ for each training data point. This variable gives the distance a point must move to be on the correct side of the margin and thus is zero if the point is on the correct side of the margin. The SVM then adds a new constraint on the sum $\sum_{i=1}^{N} \xi_i < C$ for some constant C. The optimization problem above can be rewritten to eliminate the margin variable M as

$$\min_{\beta,\beta_0} ||\beta||$$

$$y_i(x_i^T\beta + \beta_0) \geq 1,\ i = 1,\ldots,N.$$

The SVM then allows for some points to lie on the wrong side of by modifying the second constraint to

$$y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i,\ i = 1,\ldots,N.$$

The algorithm assigns a class to new data by simply considering on which side of the hyperplane the data lies.

2.8 Random Forest

As the name 'forest' implies, a random forest is a collection of decision trees. The 'random' comes in the way in which the trees are grown. First, the data used to train each tree is randomized. If there are n training examples, n are randomly selected with replacement. Once the training data

for a particular tree is selected, at each node, a randomized subset of the features used in training is selected to make the decision for that node. The number of features selected is typically the floor of $\sqrt{p}$, where p is the number of features in the entire data set, although as few as one feature can possible be selected. Once a number is selected, it is held fixed through the entire training process. No pruning is done on the trees Breiman (1999). The number of trees grown for the forest is typically chosen using the validation data to test how many trees are necessary before the accuracy levels off.

To classify new data, each tree in the forest predicts a class and casts a vote for this class. The class with the most votes is the class to which the random forest assigns the new data. This algorithm almost always produces higher accuracy that a single decision tree. Each individual tree is nearly unbiased, but trees are known to be very noisy. Giving each tree one vote towards the final classification averages out the noise. The advantage in random forests come in that the individual trees are all independent and identically distributed. When i.i.d. random variables are averaged, the independence means that the variance is reduced from the variance of a single random variable σ^2 to $\frac{1}{N}\sigma^2$ where N is the number of random variables Hastie et al. (2009). The final variance for the random tree model takes the positive pairwise correlation of the features into account, but is reduced by the same order of $\frac{1}{N}$.

When analyzing the error of random forests, we find a trade-off between correlation between trees and strength of individual trees. Accuracy of the model improves with decreased correlation between trees and increased strength of individual trees, however correlation between trees and strength of individual trees vary together with number of features considered at each node. Thus when one increases so does the other. The model optimizes the number of features used for each node based on the training examples that were *not* selected for a particular tree in the randomized selection process.

Because random forests typically consist of hundreds, sometimes thousands, of trees, the data is untrackable through the classification process. However, computation can be made to quantify the importance of each variable in the data set. This can be done by computing what is called the

Gini impurity index at each node. The impurity index is highest when the data at the node is split evenly between classes and lowest when the data is all of a single class. In other words, a node is more "pure" when previous splits have isolated data from a single class as much as possible.

To compute the importance of a variable within the data set, we consider all the nodes that utilize that particular variable to split the data. The importance of that variable at each node can be calculated by subtracting the Gini impurity index of the two subsequent nodes from the Gini impurity index of the parent node. The importance per node is averaged over the entire forest and results in the overall importance of the variable. We compare these average Gini decreases of all variables. This can be used for dimensionality reduction by training a random forest model with say M variables and selecting $m << M$ of the most important variables according to the average Gini decrease results. We then train a new random forest with only these m variables.

2.9 Feature Engineering and Dimensionality Reduction: Principal Component Analysis

Often manipulating the data before inputting it into a machine learning algorithm can raise accuracy of a model. This process is called feature engineering and is most often approached by constructing new features from the original features. These new features can produce more separability in the data so that patterns can be more easily recognized by algorithms. In this study we consider a method of feature engineering called principal component analysis (PCA).

PCA is an unsupervised method, meaning it is not necessary to know to which class a data point is assigned to implement the method (unlike all previous algorithms, which require the class labels to train models). PCA is derived from singular value decomposition (SVD) Hastie et al. (2009). Let $\mathbf{X}$ be an $n \times p$ matrix with all input feature information but not class labels where n is the number of data points we have and p is the number of features. Assume the features have been normalized and thus are centered so that means are zero. Then each row of $\mathbf{X}$ contains all p feature values for a particular training example.

We can use SVD to write

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$$

where $\mathbf{U}$ is $n \times p$, Σ is $p \times p$, and $\mathbf{V}$ is $p \times n$. The columns of $\mathbf{U}$ and $\mathbf{V}$ form orthonormal bases The column spaces of $\mathbf{U}$ and of $\mathbf{V}$ span the column and row space of $\mathbf{X}$ and row space of $\mathbf{X}$ respectively. The matrix Σ is a diagonal matrix.

Consider the covariance matrix

$$\mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{n-1}.$$

Substituting the SVD we can obtain the eigen decomposition of $\mathbf{X}^T\mathbf{X}$

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= \left(\mathbf{U}\Sigma\mathbf{V}^T\right)^T \mathbf{U}\Sigma\mathbf{V}^T \\ &= \mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T \\ &= \mathbf{V}\Sigma^2\mathbf{V}^T.\end{aligned}$$

Thus we can see that the columns of $\mathbf{V}$ are the eigenvectors, called principal components, of the covariance matrix. The eigenvalues λ_j can be written in terms of the singular values s_j of Σ by $\lambda_j = \frac{s_j^2}{n-1}$.

The first principal component gives the direction which captures the maximum variance in the data. That is, if the data were projected onto this vector, it would have the maximum separability. The principal components are listed in descending order of variance explained. We can plot the principal components along the x axis and fraction of variance explain on the y axis in what is called a scree plot. This visual is used to determine how many of the principal components are need to explain the variance in the data.

The data in $\mathbf{X}$ is mapped to the orthogonal space defined by the principal components. The values of each principal component become new feature values for each data point. Thus we have engineered new features which are linear combinations of the original features. At this point the number of principal components is p, but because they are now listed in order of variance

explained, we may drop features, beginning with the last one, without losing valuable information. In fact, we often eliminate meaningless variance when we drop portion of these new features so that only the most informative features remain.

Chapter 3

Machine Learning applied to data on college students

Although traditional data models still make up the vast majority of predictive models in the field of education, many have had success with algorithmic models as well. One study (Marbouti et al., 2016) similar to our study was conducted in 2016. Researchers analyzed machine learning algorithmic models which predict at-risk college students in standards based grading courses. They trained binary classification models to predict whether a student will pass or fail a class.

Similar to our study, the researchers experimented with logistic regression, support vector machines, decision trees, Näive Bayes classifiers, and K nearest neighbor (KNN) clustering. They considered multi-layer perceptron (neural networks) in addition to the models we use in our study. They used logistic regression (which can also be considered as a data model) as their baseline model with which to compare all other models. Logistic regression alone predicted 92.6% of students correctly. The researchers were able to improve on this with several models, the best of which was KNN which predicted 94.9% of students correctly.

When we consider accuracy of pass and fail predictions separately, KNN performs with 99.7% accuracy on students who pass the class. The 0.3% of students who were predicted to fail when they actually passed all passed with only a C. However with only 34.5% accuracy in students

predicted to fail, KNN is the worst model for predicting students who will not pass. While helpful in identifying students who will pass the course, it is not effecting in predicting at-risk students. Table 3.1 shows a similar summary for other models.

Model	Overall Accuracy	Pass Accuracy	Fail Accuracy
Logistic Regression	92.6%	95.3%	58.6%
KNN	94.9%	99.7%	34.5%
Multi-layer Perceptron	93.1%	96.7%	48.3%
Decision Tree	92.3%	96.1%	44.8%
Näive Bayes'	86.9%	87.0%	86.2%
Support Vector Machine	87.2%	88.4%	72.4%
Ensamble	92.1%	93.9%	69%

Table 3.1: Model accuracy comparison (Marbouti et al., 2016)

Another such study looked at engineering students in a dynamics course (Huang & Fang, 2013). The goal was to predict the binary outcome (pass or fail) of the course mid-semester. The training data consisted of test grades for the dynamics course, GPA, final grades in four pre-requisite courses and of course the pass/fail result of the dynamics course for 323 students over four semesters. The researchers considered four machine learning models: multiple linear regression, multi-layer perceptron (neural network), radial basis function neural network, and support vector machine. The study shows that if the goal is to predict the success of the class overall, the simplest model, multiple linear regression, with only GPA as a predictor will do just as well as other models. However if the goal is to predict whether an individual student will be successful in the course, the SVM model with all predictors available before the midterm exam has the highest overall accuracy at 64%.

The study trained model for multiple points throughout the semester. Prior to the start of the semester, a model with students' GPA and final grades of four prerequisite courses does not predict significantly better that GPA alone. Both predict near 50% accuracy. After the first of three midterm exams, the SVM model predicts with 59.1% accuracy. After the second midterm, the

accuracy increases to 61.3% and after the third midterm the accuracy jumps to 64.0%.

Another recent study which employed machine learning algorithms explored how different students respond to collaborative learning–the "flipped classroom" style of teaching (Cen et al., 2016). "We explore the predictability of academic performance based on the mechanics of interactions during live collaborative learning. The aim is to predict how well the group is likely to perform given all available individual and group historical evidence as well as live interaction patterns." The study divides students into classes based on how the learn from this style and make predictions on how new students will learn in this environment based on other characteristics of the students.

A similar study predicted how student would learn from education games (Barata et al., 2016). The researcher's used algorithmic models and feature engineering to answer the following questions:

1. Is there a subset of relevant features that can be used to predict the student type in the data sample?

2. Can the relevant feature set be used to predict the students? class in another instance of the course?

3. Can student types be predicted by midterm?

Chapter 4

Results and Method Comparison

4.1 Descriptive Statistics

Before normalizing the data and examining the variety of predictive models discussed in Chapter 2, we would like to gain insight into the data by comparing GPA scores and ACT scores of student groups. We separate the students based on whether or not they passed a course and compute the average, median and standard deviations of GPA and Math ACT scores. These are showing in Table 4.1. The differences in the ACT averages and percent changes are shown in Tables 4.2 and 4.3. In contrast, we see a higher separability in the students who passed and failed when we consider GPA. This is shown in the Table 4.4.

Score and student group	Average	Median	Standard deviation
GPA of all Int. Algebra students	3.03	3.08	0.56
GPA of Int. Algebra failing students	2.86	2.9	0.56
GPA of Int. Algebra passing students	3.18	3.22	0.50
GPA of all College Algebra students	3.22	3.3	0.57
GPA of College Algebra failing students	3.01	3.05	0.59
GPA of College Algebra passing students	3.39	3.46	0.49
Math ACT of all Int. Algebra students	18.6	19	2.1
Math ACT of Int. Algebra failing students	18.2	18	2.2
Math ACT of Int. Algebra passing students	18.9	19	1.9
Math ACT of all College Algebra students	22.4	23	2.7
Math ACT of College Algebra failing students	21.7	22	3.0
Math ACT of College Algebra passing students	22.9	23	2.3

Table 4.1: Descriptive statistics for student groups

ACT Component:	Math	English	Science	Reading
Difference in average score (out of 36):	0.65	0.75	0.57	0.16
Percent difference:	1.8%	2.1%	1.6%	0.5%

Table 4.2: Differences in average ACT scores of Intermediate Algebra students who passed vs. failed

ACT Component:	Math	English	Science	Reading
Difference in average score (out of 36):	1.25	0.91	0.72	0.38
Percent difference:	3.6%	2.5%	2.0%	1.1%

Table 4.3: Differences in average ACT scores of College Algebra students who passed vs. failed

	Average GPA (pass)	Average GPA (fail)	Difference	Percent difference
Intermediate Algebra:	3.18	2.86	0.32	8.0%
College Algebra:	3.39	3.01	0.39	9.6%

Table 4.4: Differences in average GPA scores students who passed vs. failed

To help visualize the data, we plot the GPA against the ACT scores of individual students in Figures 4.1 and 4.2. Teal points correspond to students who passed the course and coral points correspond to students who did not. The large dots represent the average of all the corresponding students. The ellipses designate the area in which each type of student lies with 95% confidence. The plots are done with normalized data so that GPA and ACT scores are on the same scale. This way we can more easily compare the separability due to GPA and due to ACT scores.

Intermediate Algebra 95% confidence ellipses

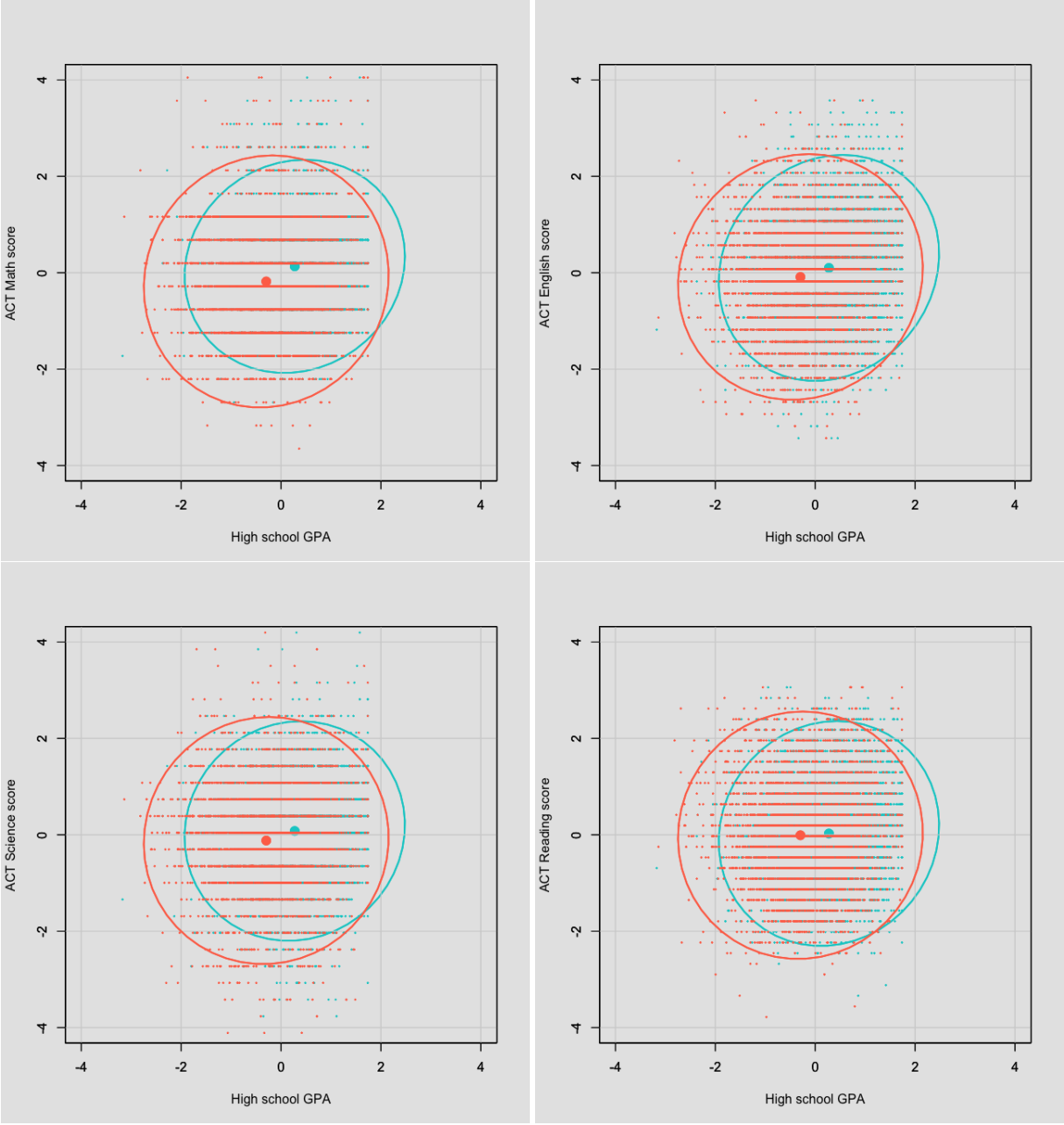

Figure 4.1: Ellipses show separation between students who passed and students who failed Intermediate Algebra. The x axis shows normalized high school GPA and the y axis shows normalized ACT scores. The ellipse represents 95% confidence.

College Algebra 95% confidence ellipses

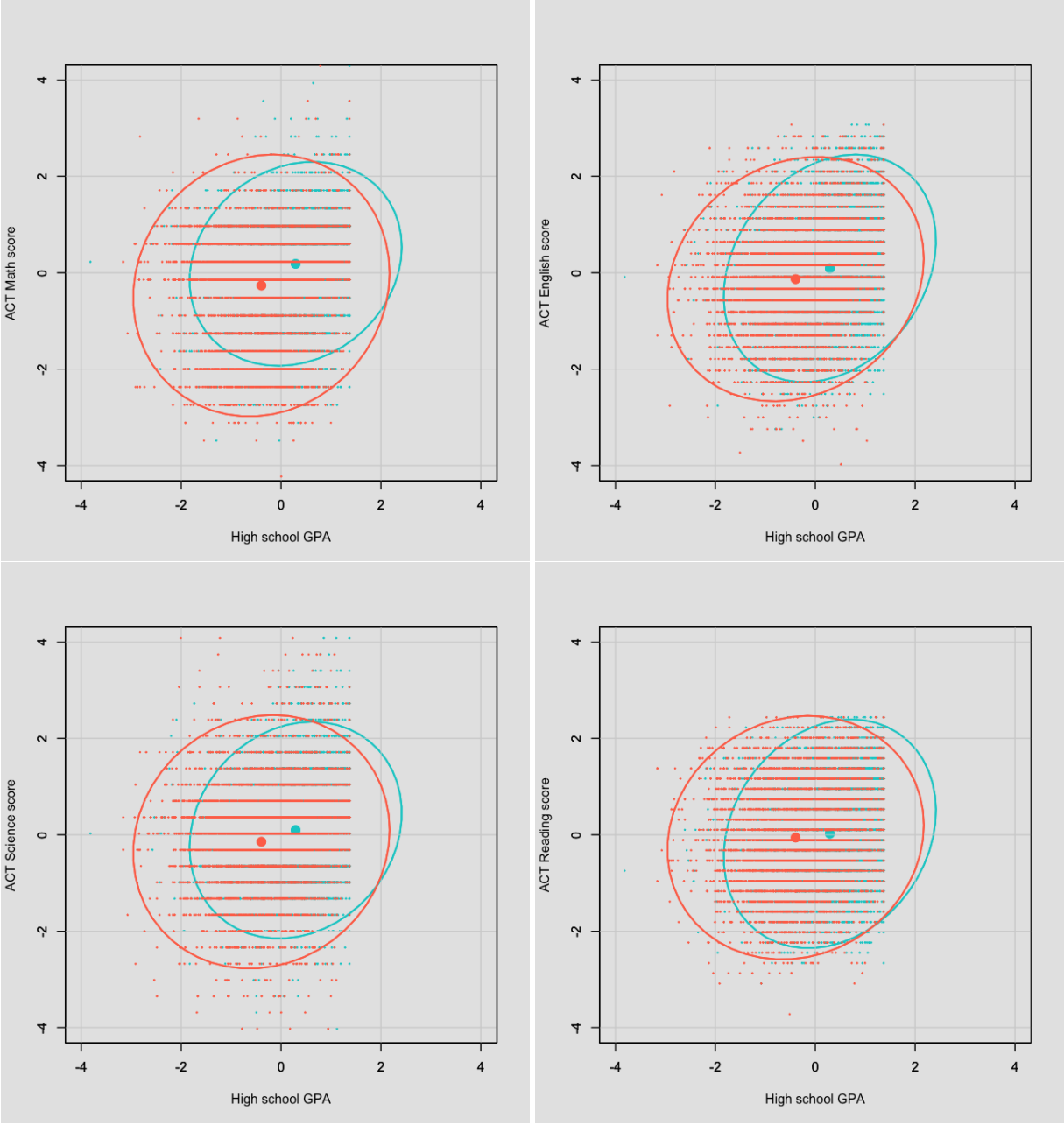

Figure 4.2: Ellipses show separation between students who passed and students who failed College Algebra. The x axis shows normalized high school GPA and the y axis shows normalized ACT scores. The ellipse represents 95% confidence.

Student data convergence to average: To more easily view all the data points together, we separate the students based on whether or not they passed the course, and take averages of small counts of students. We then plot a single data point for each of these averaged groups. This results in data that is much more separable, shown in Figures 4.3 - 4.8.

Averages of 5 students

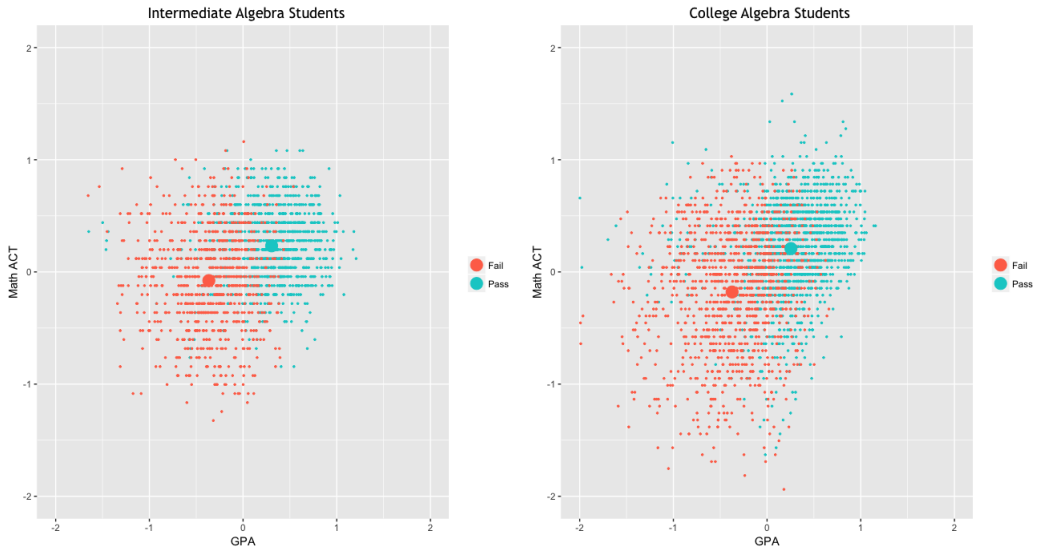

Figure 4.3: Each data point represents the average of 5 students, all of which passed or all of which failed.

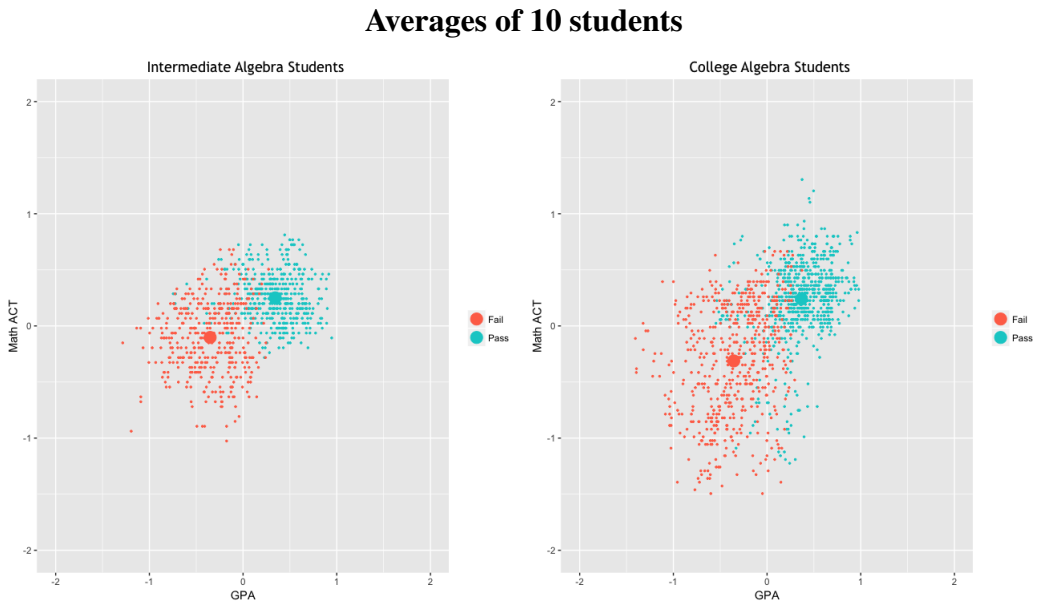

Figure 4.4: Each data point represents the average of 10 students, all of which passed or all of which failed.

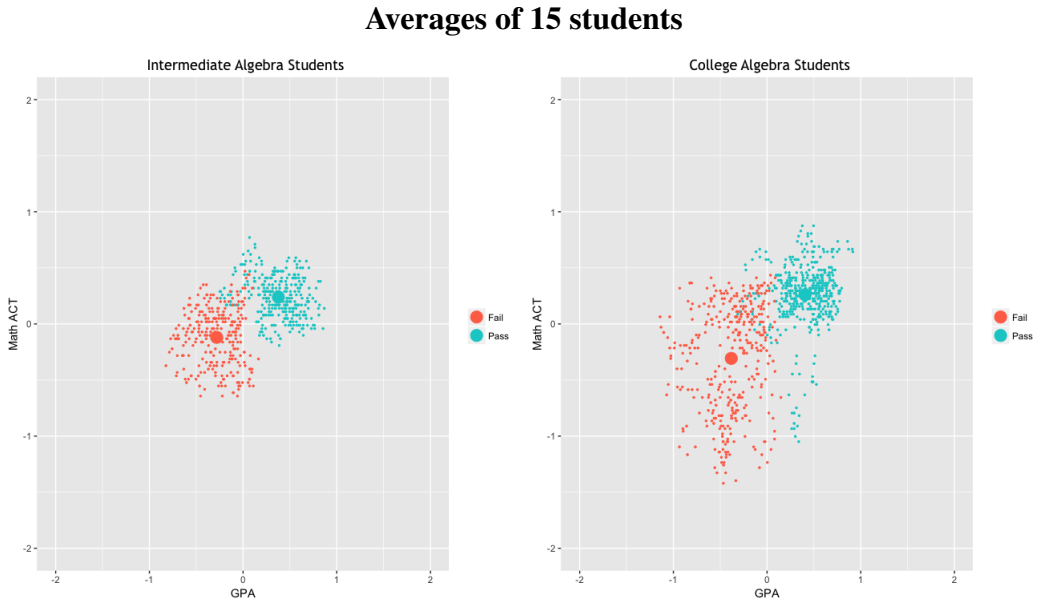

Figure 4.5: Each data point represents the average of 15 students, all of which passed or all of which failed.

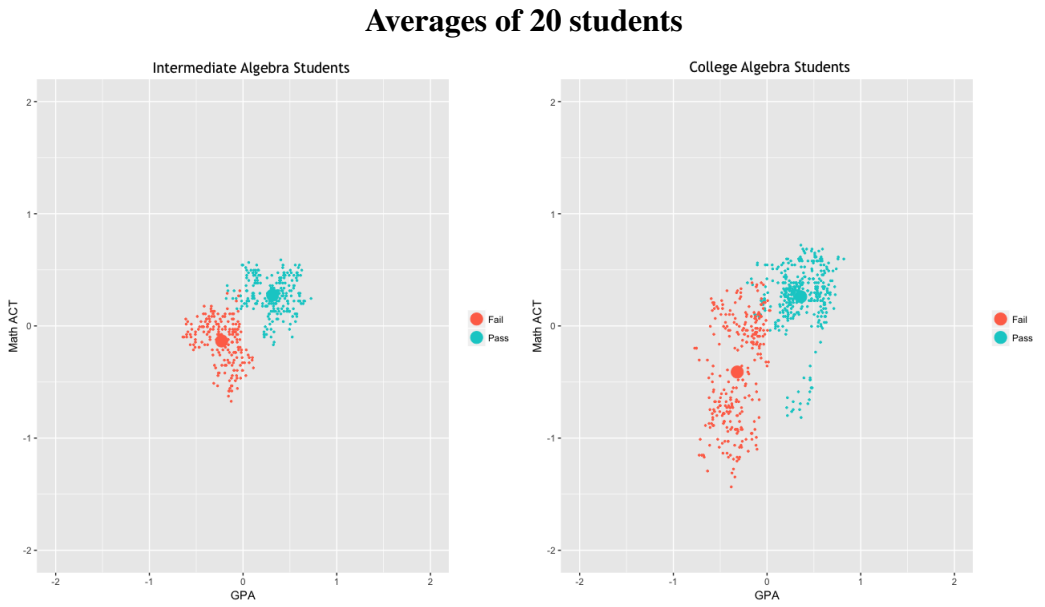

Figure 4.6: Each data point represents the average of 20 students, all of which passed or all of which failed.

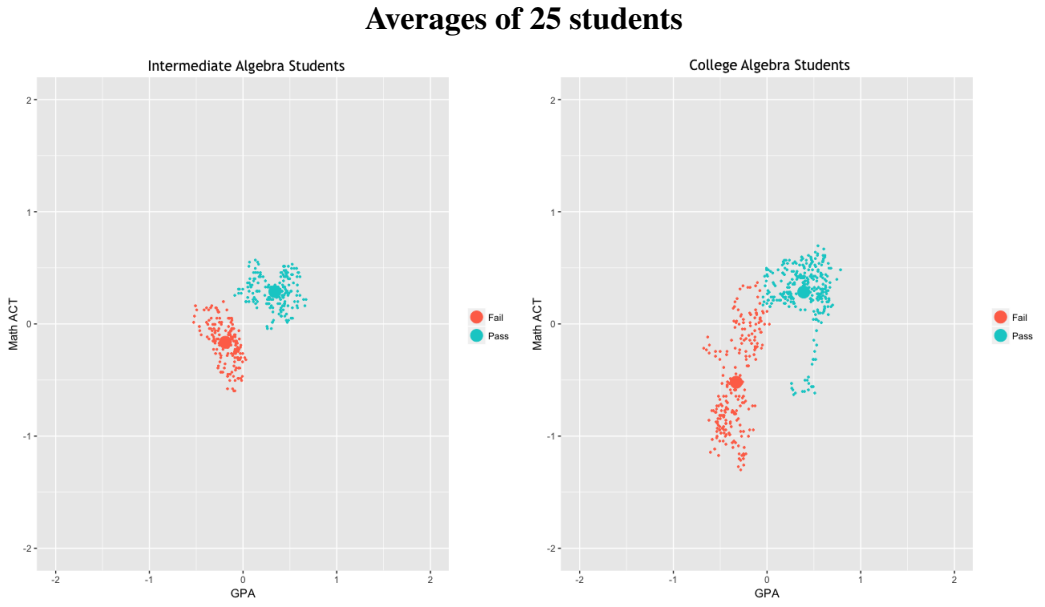

Figure 4.7: Each data point represents the average of 25 students, all of which passed or all of which failed.

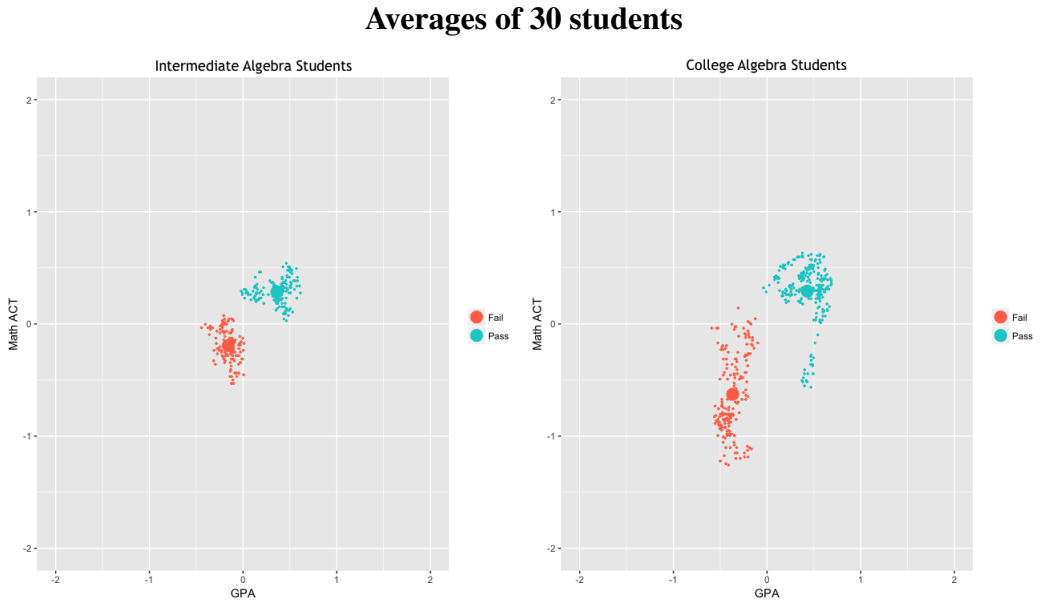

Figure 4.8: Each data point represents the average of 30 students, all of which passed or all of which failed.

It is also helpful to divide students based on their Math ACT and GPA scores and consider what percent of each group passed and failed. These numbers are given in bar graphs in Figures 4.9 - 4.12. In these graphs we can see that the percent of students passing increases slightly with increasing ACT but increases dramatically with increasing GPA. The total number of students are given in the gray graphs. Because most of the students in College Algebra scoring 21 of less are students who previously took Intermediate Algebra, we do not consider them in the increase/decreasing bar graphs.

Intermediate Algebra students divided by Math ACT score

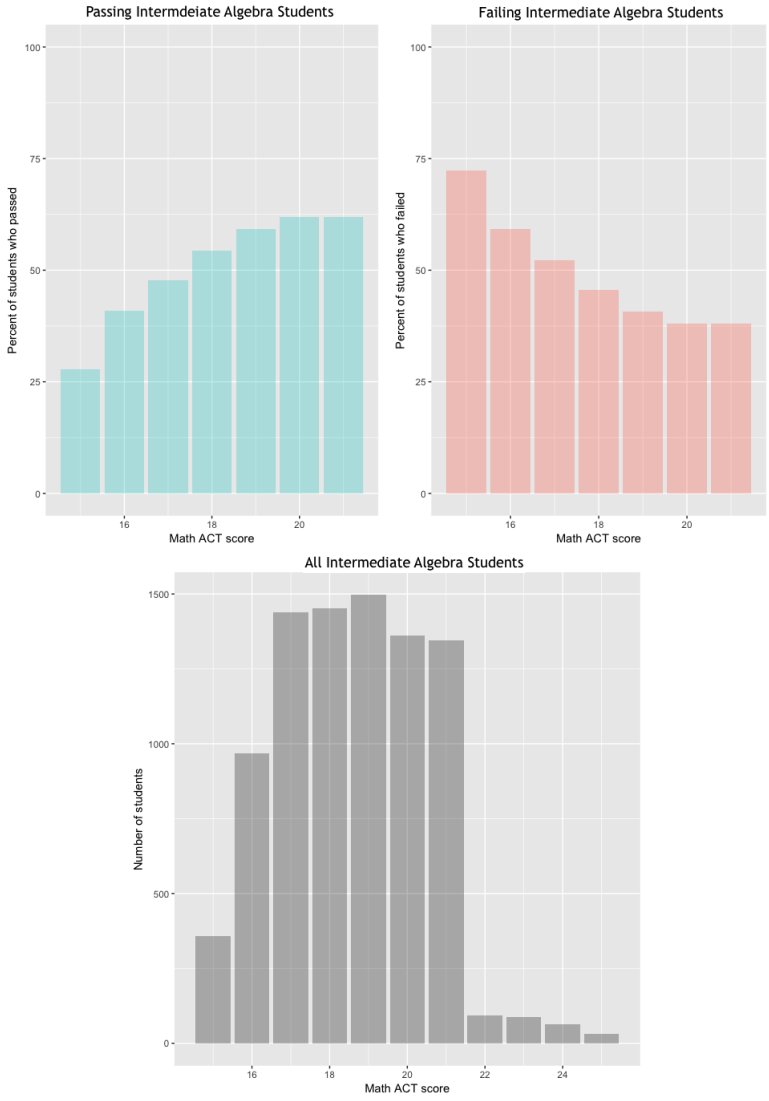

Figure 4.9: Percent of Intermediate Algebra students who pass and fail divided by Math ACT scores. The percent of students passing increases from 27.7% of students with a score of 15 to 62.0% of students with a score of 25.

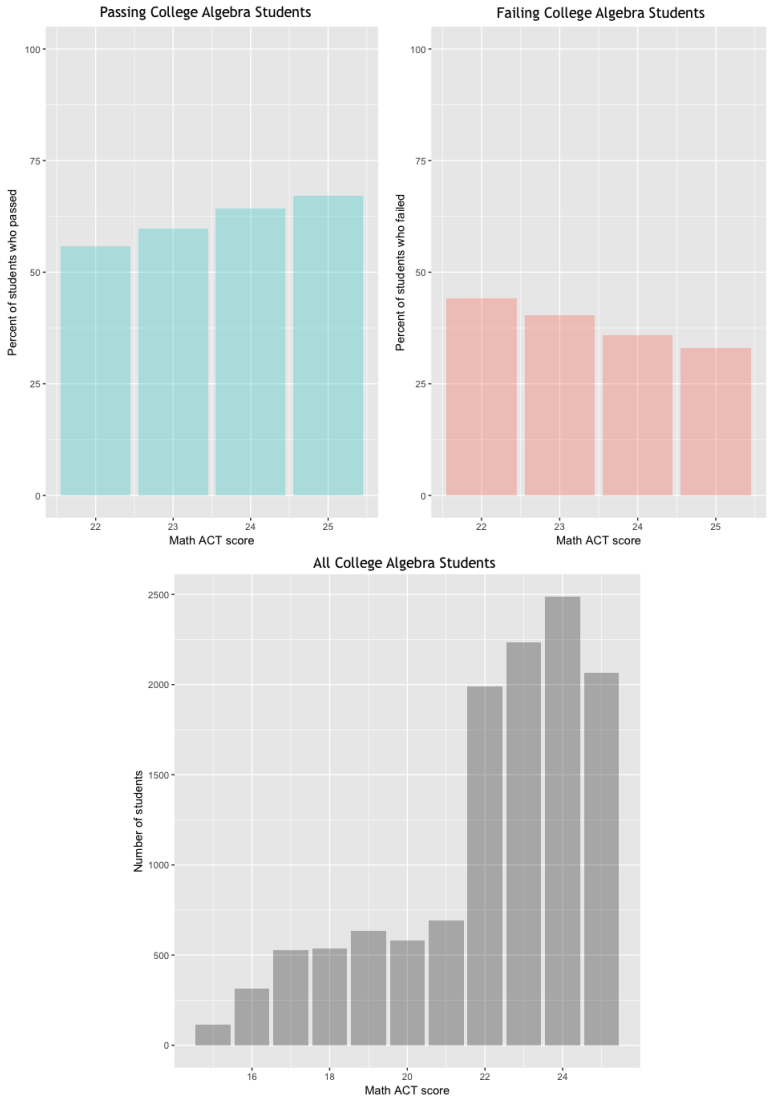

Figure 4.10: Percent of College Algebra students who pass and fail divided by Math ACT scores. The percent of students passing increases from 55.8% of students with a score of 22 to 67.1% of students with a score of 25.

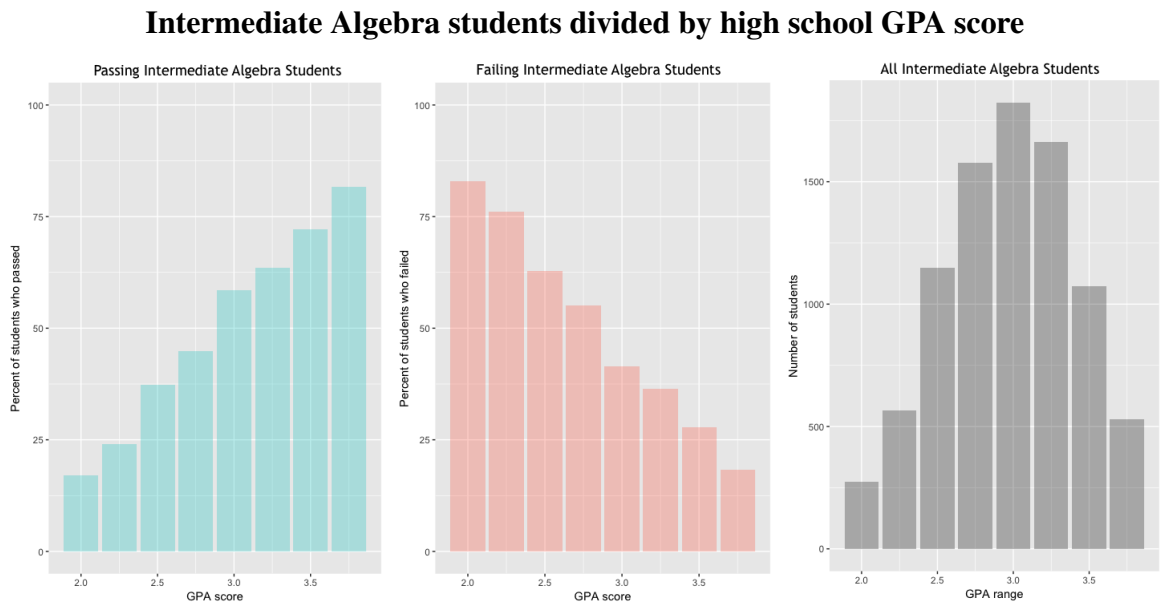

Figure 4.11: Percent of Intermediate Algebra students who pass and fail divided by GPA score ranges. The percent of students passing increases from 17.1% of students with a GPA of 2.0 to 2.25 to 81.7% of students with a score of 3.75 to 4.0.

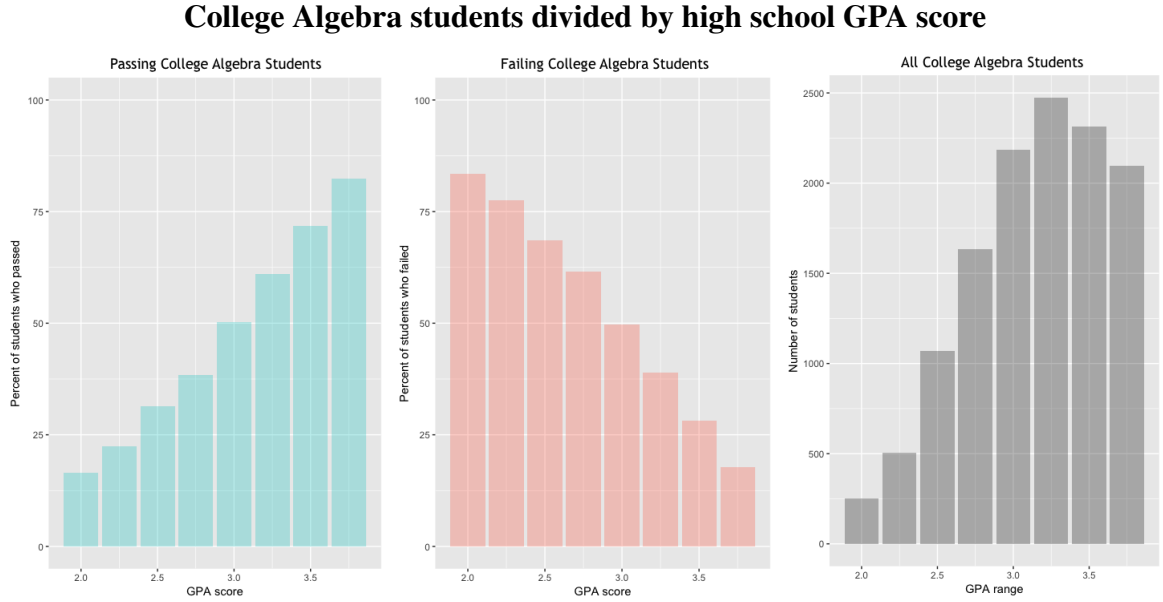

Figure 4.12: Percent of College Algebra students who pass and fail divided by GPA score ranges. The percent of students passing increases from 16.5% of students with a GPA of 2.0 to 2.25 to 82.3% of students with a score of 3.75 to 4.0.

	15	16	17	18	19	20	21
3.75 - 4.0	0.1%	0.3%	0.6%	0.9%	1.1%	1.1%	1.4%
3.5 - 3.74	0.3%	1.0%	1.7%	2.1%	2.1%	2.1%	2.3%
3.25 - 3.49	0.7%	1.8%	2.9%	3.4%	3.3%	2.9%	2.8%
3.0 - 3.24	0.7%	2.4%	3.5%	3.3%	3.8%	3.3%	2.9%
2.75 - 2.99	0.9%	2.2%	3.3%	3.0%	2.8%	2.4%	2.3%
2.5- 2.74	0.7%	1.7%	2.3%	2.2%	1.9%	1.9%	1.8%
2.25 - 2.49	0.3%	0.8%	1.2%	1.0%	0.9%	0.9%	0.8%
<2.25	0.3%	0.6%	0.8%	0.6%	1.0%	0.9%	0.8%

Table 4.5: Percent of Intermediate Algebra students with Math ACT/GPA combinations. This table represents 95% of the Intermediate Algebra data. The remainder either had a Math ACT score greater than 21 but opted to take Intermediate Algebra anyway or had a Math ACT lower than 15, but this was a negligible number of students.

Putting the information from the bar graph together, we create a table with Math ACT scores along the x axis and GPA ranges along the y axis. At each cell in the grid, we give the percentage of the total students that fall into that cell. This is show in Tables 4.5 and 4.6. Then we compute the probability that the students in each particular cell will pass the course. This is shown in Tables 4.8 and 4.9.

The analysis in this section all indicates that GPA separates passing students from failing students more definitely than Math ACT score. We see this first from comparing the percent change in scores in averages among passing and failing students. While average Math ACT score changes by only 1.8% for Intermediate Algebra and 3.6% for College Algebra, average GPA changes by 8.0% for Intermediate Algebra and 9.6% for College Algebra. The higher separability that GPA gives compared to any other ACT component can be visualized in Figures 4.1 and 4.2. Notice that there is more separation horizontally than vertically, especially in ACT English and ACT Reading but also in ACT Math and ACT Science.

In the averaged student data convergence, we notice that GPA converges much faster than Math ACT. Especially in Figure 4.8 we see in College Algebra the passing and failing students each fall

	22	23	24	25
3.75 - 4.0	2.1%	3.0%	3.8%	4.3%
3.5 -	2.8%	3.0%	4.1%	3.6%
3.25	3.2%	3.6%	3.9%	2.8%
3.0	3.0%	3.0%	3.1%	2.1%
2.75	2.2%	2.2%	2.1%	1.5%
2.5	1.3%	1.4%	1.3%	1.0%
2.25	0.6%	0.9%	0.7%	0.4%
<2.25	0.6%	0.6%	0.7%	0.4%

Table 4.6: Percent of College Algebra students with Math ACT/GPA combinations. Just under 70% of all students in College Algebra fall into the Math ACT/GPA grid shown here. Just over 5% have Math ACT scores greater than 25 and the remaining students have Math ACT scores below 22, meaning they likely took Intermediate Algebra post high school at the university of our present study or elsewhere

	$x \geq 75\%$
	$0.75 > x \geq 50\%$
	$50\% > x \geq 25\%$
	$25\% > x$

Table 4.7: Legend for color scheme where x is the percent of students in each cell who pass the course.

	15	16	17	18	19	20	21
3.75	42.86%	76.00%	82.14%	87.34 %	84.85%	89.69%	84.68%
3.5	45.16%	64.13%	73.47%	73.77%	68.48%	82.98%	76.35%
3.25	33.90%	55.83%	53.08%	61.07%	71.53%	69.29%	74.70%
3.0	31.75%	43.87%	56.49%	55.02%	67.07%	69.90%	64.45%
2.75	31.58%	31.82%	40.55%	45.32%	52.02%	54.76%	57.21%
2.5	13.33%	31.76%	32.00%	42.35%	37.95%	43.45%	43.95%
2.25	12.90%	13.33%	24.27%	40.77%	25.00%	29.87%	28.38%
<2.25	20.00%	24.07%	18.67%	26.00%	32.61%	12.99%	18.67%

Table 4.8: Percent of Intermediate Algebra students who pass Intermediate Algebra in each of the Math ACT and GPA combinations.

	22	23	24	25
3.75	83.97%	83.81%	83.20%	84.27%
3.5	72.85%	73.11%	74.03%	75.60%
3.25	61.08%	65.21%	68.48%	66.20%
3.0	51.73%	57.52%	57.47%	57.99%
2.75	38.55%	42.18%	51.52%	51.87%
2.5	28.83%	34.86%	43.37%	42.86%
2.25	22.22%	21.30%	25.58%	28.30%
<2.25	21.33%	21.92%	20.00%	21.43%

Table 4.9: Percent of College Algebra students who pass College Algebra in each of the Math ACT and GPA combinations.

into a narrow GPA column but there is still large variation in Math ACT with a large overlap. This shows that GPA more precisely identifies passing and failing students. The bar graphs with percent of students passing and failing with each ACT score/GPA range further confirm that GPA is a more important predictor of success. Notice that students have a much more dramatic increase in passing rates with increase GPA than with increasing Math ACT. Finally, when we put this information together in Tables 4.8 and 4.9 we see that while higher Math ACT scores does positively impact percent of students passing the course, the general cutoffs are horizontal, meaning that GPA is a heavier weight in passing rates than Math ACT.

4.2 Principal Component Analysis

Using principal component analysis (PCA), we can map the original variables to new variables which are orthogonal. This maximizes the separability in the data. The new variables, which are features called components, are linear combinations of the original variables. They are listed in terms of how much of the variance in the data each one explains. The direction of the first component is chosen to maximize the explained variance. PCA results from the Intermediate Algebra and College Algebra data are shown in Tables 4.10 and 4.11 respectively and scree plots for the new features are shown in Figure 4.13.

Intermediate Algebra	PC1	PC2	PC3	PC4	PC5
Proportion of Variance	0.49	0.20	0.14	0.10	0.07
Cumulative Proportion	0.49	0.69	0.83	0.93	1.00000

Table 4.10: Intermediate Algebra variance explained by PCA features

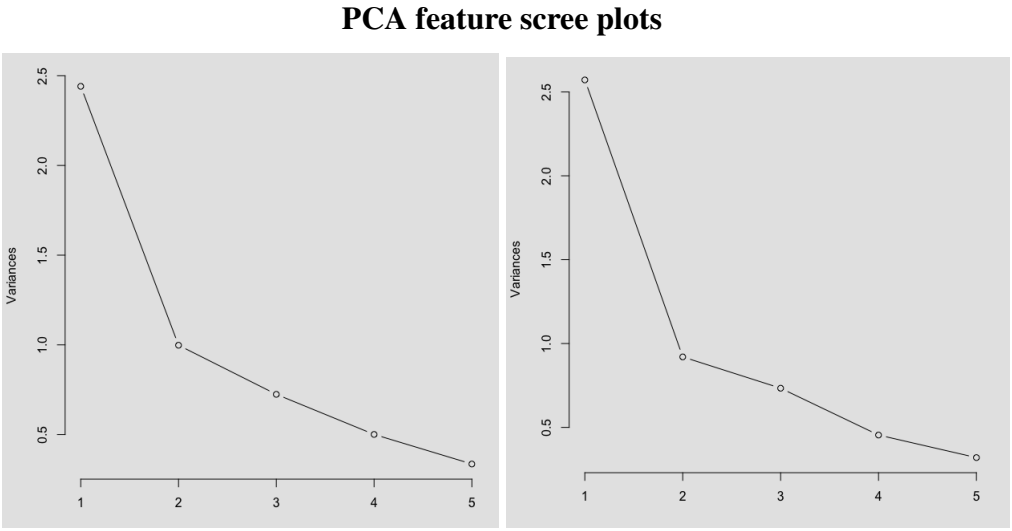

Figure 4.13: Intermediate Algebra scree plot is on the left and College Algebra scree plot is on the right. Scree plots show the fraction of total variance explained by each principal component.

College Algebra	PC1	PC2	PC3	PC4	PC5
Proportion of Variance	0.51	0.18	0.15	0.09	0.06
Cumulative Proportion	0.51	0.70	0.85	0.94	1.00

Table 4.11: College Algebra variance explained by PCA features

Using the new PCA variables, we again would like to visualize the data to see if more separation between students who pass and students who fail appears. The maximum amount of separation we will see in two dimensions occurs in plotting principal component (PC) 1 and PC 2. For separability comparison, we also plot PC 3 against PC 4 in Figure 4.14. In Section 4.4 we further analyze models resulting from training with PCA components. We will compare these models to the models trained on normalized features.

PCA plots showing separability of data

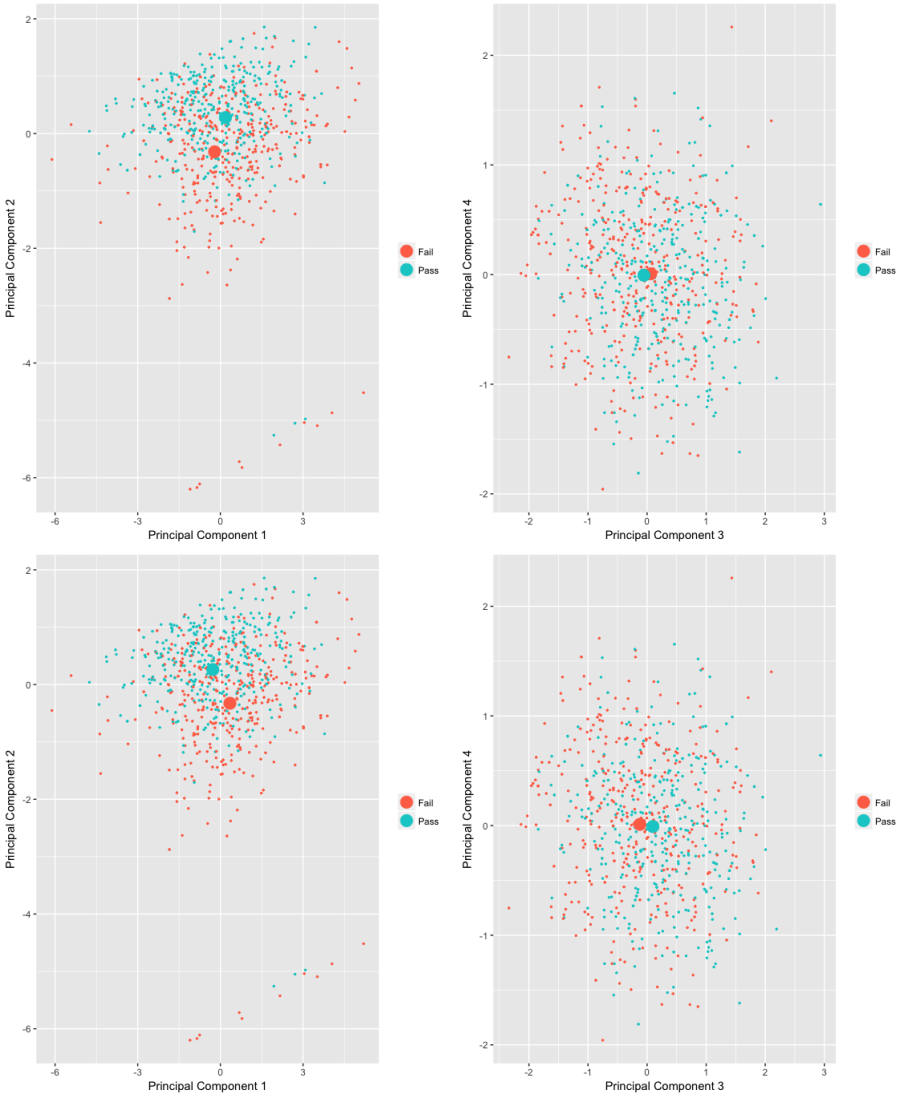

Figure 4.14: PCA features for Intermediate Algebra data are shown on the top row and for College Algebra data are shown on the bottom row. Since PCA features are ordered in terms of amount of variance explained, the most separability in two dimensions comes in plotting the first two features and almost no separability comes in plotting the next two features.

4.3 Model Comparison

We train each of the models discussed in Chapter 3 on 60% of the data from 2006 to 2016. Some models require additional parameters and these are selected based on the validation data. We then make predictions on the testing data with each model and check these predictions against the actual outcomes. We compute accuracy by dividing the number of testing data points that the model predicted correctly by the total number of testing data points. Table 4.12 shows the resulting accuracies.

Model:	Int. Algebra accuracy	College Algebra accuracy
Logistic Regression	65.5%	68.5%
Linear Discriminant Analysis	65.3%	68.4%
Quadratic Discriminant Analysis	64.7%	67.3%
Decision Tree	65.2%	66.9%
Naive Bayes	64.8%	57.5%
K Nearest Neighbors	65.1%	68.5%
Support Vector Machine	65.3%	68.4%
Random Forest	64.9%	68.4%

Table 4.12: Model accuracies with normalized data

For Intermediate Algebra, 65.5% is the maximum accuracy we reach in the testing data. This was achieved by logistic regression. However the lowest accuracy was 64.7%, a difference of only 0.8% in accuracy, so all Intermediate Algebra models perform approximately equivalently.

For College Algebra, 68.5% accuracy is the maximum accuracy, achieved by both logistic regression and KNN. The LDA, SVM, and random forest models are all within 0.4% of the maximum accuracy. The only model that performs significantly worse than the other models is Naive Bayes. This is likely because the main assumption necessary for Naive Bayes is that all features are independent. Clearly in our case, GPA is correlated with ACT scores and individual ACT scores are correlated with each other. Thus the independence assumption is false and we expect Naive

Bayes to perform poorly.

In other studies referenced in Chapter 3, there was more variance in the accuracies of difference models. With the exception of Naive Bayes in College Algebra, all models for both Intermediate Algebra and College Algebra are very consistent in accuracies. This is likely because of the massive data set we use in training.

We would like to consider not only overall accuracies of models, but accuracy in predicting passing students and accuracy in predicting failing students individually. As shown in a study referenced in Chapter 3 by Marbouti et al. (2016), even though Naive Bayes performs the worst overall, it is by far the best model in predicting failing students. We show results of passing and failing accuracies from our study in Tables 4.13 and 4.14.

Model	Overall Accuracy	Pass Accuracy	Fail Accuracy
Logistic Regression	65.5%	73.1%	57.3%
Linear Discriminant Analysis	65.3%	75.2%	54.5%
Quadratic Discriminant Analysis	64.7%	79.0%	48.0%
Decision Tree	65.2%	70.5%	59.5%
Näive Bayes'	64.8%	79.3%	48.9%
K Nearest Neighbors	65.1%	72.3%	57.3%
Support Vector Machine	65.3%	72.9%	56.8%
Random Forest	64.9%	73.4%	56.8%

Table 4.13: Intermediate Algebra model accuracies of passing students and of failing students compared to overall accuracy.

For Intermediate Algebra, all models give passing accuracies between 70% and 80% with the maximum accuracies achieved by QDA and Naive Bayes. However these two give the lowest failing accuacies. All failing accuracies are between 48% and60% with the highest achieved by the decision tree.

For College Algebra, all models give passing accuracies between 75% and 82% except the decision tree, which has accuracy of 67.5%. However the decision tree gives the highest failing

Model	Overall Accuracy	Pass Accuracy	Fail Accuracy
Logistic Regression	68.5%	77.3%	56.6%
Linear Discriminant Analysis	68.4%	79.1%	53.9%
Quadratic Discriminant Analysis	67.3%	81.2%	49.4%
Decision Tree	66.9%	67.5%	66.0%
Näive Bayes'	57.5%	81.0%	49.0%
K Nearest Neighbors	68.5%	76.6%	57.7%
Support Vector Machine	68.4%	78.0%	55.6%
Random Forest	68.4%	75.0%	58.5%

Table 4.14: College Algebra model accuracies of passing students and of failing students compared to overall accuracy.

accuracy at 66.0%. The closest to the decision tree is the random forest with 58.5% accuracy. Again, the models with the highest passing accuracies are also the models with the lowest failing accuracies.

Model	Precision	Recall	F1 Score
Logistic Regression	0.65	0.77	0.70
Linear Discriminant Analysis	0.64	0.75	0.69
Quadratic Discriminant Analysis	0.63	0.79	0.70
Decision Tree	0.65	0.70	0.68
Näive Bayes'	0.63	0.79	0.70
K Nearest Neighbors	0.65	0.72	0.68
Support Vector Machine	0.65	0.73	0.69
Random Forest	0.65	0.74	0.69

Table 4.15: Comparing Intermediate Algebra models with F1 scores

To further compare the performance of our models, we compute the precision, recall and F1 score defined by

$$\text{precision} = \frac{\text{true positives}}{\text{true positives + false positives}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives + false negatives}}$$

$$\text{F1 score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision + recall}}.$$

Precision gives proportion of students predicted to pass who actually did pass. Recall gives the proportion of students who passed the class that the model predicted correctly. The F1 scores gives the harmonic mean of the two proportions. We would like to find the model that maximizes both precision and recall, and thus we would like to maximize the F1 score. Tables 4.15 and 4.16 gives these results.

Like we saw in comparing accuracies in models, the F1 scores are incredibly consistent between models. This is probably a result of the large data set.

Model	Precision	Recall	F1 Score
Logistic Regression	0.71	0.77	0.74
Linear Discriminant Analysis	0.70	0.79	0.74
Quadratic Discriminant Analysis	0.68	0.81	0.74
Decision Tree	0.73	0.68	0.70
Näive Bayes'	0.68	0.81	0.74
K Nearest Neighbors	0.71	0.77	0.74
Support Vector Machine	0.70	0.78	0.74
Random Forest	0.71	0.77	0.74

Table 4.16: Comparing College Algebra models with F1 scores

4.4 Principal Component Analysis Results

We train each of the models discussed in Chapter 2 again with the PCA features. Results are shown in Table 4.17. Most models show a slight improvement with PCA features. The random forest model gives the best result for both Intermediate Algebra and College Algebra with 67.1% and 70.0% respectively. However the slight improvements in accuracy are not enough to make up for the loss in interpretability. Since each PC is a linear combination of the original variables, we do not gain information about algebra students by seeing how models weight PC's or by the importance models give to PC's. By keeping original variables it is clear which scores are most important in placing students based on results of models like logistic regression and LDA. For this reason we continue analysis with original variables and not PCA features.

In the remaining sections of this chapter, we look more closely at most of the models. We choose to do no further analysis on QDA because any insight this model give would be similar to LDA, but LDA has slightly higher accuracy. We also choose to do no further analysis on Naive Bayes because the model does not provide more interesting insight.

Model:	Int. Algebra accuracy	College Algebra accuracy
Logistic Regression	65.2%	69.4%
Linear Discriminant Analysis	65.0%	69.2%
Quadratic Discriminant Analysis	64.1%	67.7%
Decision Tree	64.1%	67.2%
Naive Bayes	53.8%	68.1%
K Nearest Neighbors	64.1%	68.2%
Support Vector Machine	65.8%	69.9%
Random Forest	67.1%	70.0%

Table 4.17: Comparison of model accuracies with PCA features

4.5 Interpreting Logistic Regression

Intermediate Algebra: We consider several combinations of available features for training a logistic regression model and making predictions on the testing data. Table 4.18 shows the resulting coefficients and accuracies of each model. After the second model with GPA and all four ACT components, variable are removed in order of lowest absolute z-score. Thus Science ACT is removed first because it is the least significant. The asterisk denotes variables that are statistically significant (absolute value of z-score > 2).

Note that because the data was normalized before training, the negative coefficients assigned to the ACT reading score do not signify a negative correlation, but instead a smaller increase in odds. In the second model with GPA and all four individual ACT components, all variables except the Science ACT are statistically significant, however the accuracy is maximized at 65.5% with the model using GPA and Math ACT only. For simplicity of model and since we do not lose any accuracy, we will do further analysis with the GPA and Math ACT only model. Notice that when we train a model using GPA only, we lose only 1.5% accuracy over the GPA and Math ACT model, but when we train a model using Math ACT only, we lose 8.2% accuracy compared to the GPA and Math ACT model.

GPA	ACT	Math	English	Reading	Science	Accuracy
0.75*	0.13*	-	-	-	-	64.3%
0.73*	-	0.26*	0.17*	-0.29*	0.11	64.8%
0.73*	-	0.29*	0.19*	-0.25*	-	64.3%
0.74*	-	0.33*	-	-0.14*	-	65.0%
0.74*	-	0.29*	-	-	-	65.5%
0.76*	-	-	-	-	-	64.0%
-	-	0.33*	-	-	-	57.3%

Table 4.18: Each row in the table represents a logistic regression model trained on Intermediate Algebra data. We train the models with different combinations of variables and compare accuracies. The numbers are the resulting coefficients for variables.

The full resulting logistic regression model for Intermediate Algebra with only GPA and Math ACT is

$$\hat{y} = 0.13 + 0.74 \cdot \text{GPA} \ + 0.29 \cdot \text{Math ACT}.$$

Since optimizing in the logistic regression algorithm requires taking the log of the maximum likelihood estimate (MLE), we must consider the coefficients of the final model as exponents of e in order to pull meaning from them. This computation give in increase in odds for a since unit increase in the corresponding variable. However the variables were normalized, so a since unit increase in the trained model corresponds to the standard deviation of that variable (that is an increase of one unit unstandardized). Then we can find the increase in odds which results in an increase of one unit of the original variable by dividing the increase in odds by the unstandardized increase of one unit. Thus when a student's GPA increases by one points, say from a 2.25 to 3.25, the student's odds in passing Intermediate Algebra increase 3.75 times. When a student's Math ACT score increases by one point, say from a 19 to a 20, the student's odds in passing Intermediate Algebra increase 0.64 times. This is summarized in the Table 4.19.

College Algebra: As with the Intermediate Algebra models, we consider several combinations of available features for training a logistic regression model and making predictions on the testing

	Coefficient	Increase in odds	Un-standardized increase of one unit	Slope
GPA	0.74	$e^{0.74} = 2.10$	0.56	3.75
Math ACT	0.29	$e^{0.29} = 1.34$	2.08	0.64

Table 4.19: Interpreting coefficients in Intermediate Algebra logistic regression model

GPA	ACT	Math	English	Reading	Science	Accuracy
0.88*	0.11*	-	-	-	-	65.9%
0.88*	-	0.37*	0.02	-0.22*	0.06	69.4%
0.88*	-	0.37*	-	-0.21*	0.06	69.5%
0.88*	-	0.40*	-	-0.18*	-	68.9%
0.85*	-	0.34*	-	-	-	68.5%
0.92*	-	-	-	-	-	66.5%
-	-	0.45*	-	-	-	61.9%

Table 4.20: Each row in the table represents a logistic regression model trained on College Algebra data. We train the models with different combinations of variables and compare accuracies. The numbers are the resulting coefficients for variables.

data. Table 4.20 below shows the resulting coefficients and accuracies of each model. The asterisk denotes variables that are statistically significant (absolute value of z-score > 2).

In the second model with GPA and all four individual ACT components both the English ACT and Science ACT are not significant contributors to the model. Again, we remove these components one at a time in order of least significant (lowest absolute z-score). The model which maximized the accuracy at 69.5% with GPA and three ACT components, however this is only 1% higher than the GPA and Math ACT only model, so for simplicity, we do further analysis with the GPA and Math ACT only model. With GPA alone we lose 2% accuracy and with ACT math only we lose 6.6% accuracy.

The resulting logistic regression model for College Algebra with only GPA and Math ACT is

$$\hat{y} = 0.22 + 0.85 \cdot \text{GPA} \ + 0.34 \cdot \text{Math ACT}.$$

	Coefficient	Increase in odds	Un-standardized increase of one unit	Slope
GPA	0.85	$e^{0.85} = 2.34$	0.56	4.18
Math	0.34	$e^{0.34} = 1.40$	2.70	0.52

Table 4.21: Interpreting coefficients in College Algebra logistic regression model

Weights of coefficients in logistic regression models

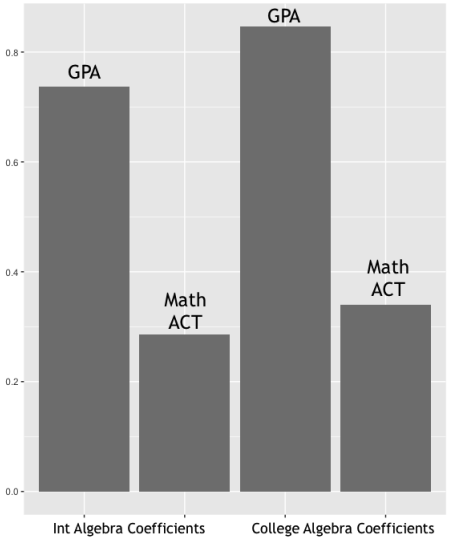

Figure 4.15: GPA is weighted 2.6 times more than Math ACT in the Intermediate Algebra model and 2.5 times more than Math ACT in College Algebra.

We interpret the coefficients here in the same way as the Intermediate Algebra coefficients–we compute the increase in odds of passing per increase of a single unit of each variable. Results are shown in Table 4.21. Figure 4.15 shows the weights of each coefficient in the logistic regression models.

Based on the models with only GPA and ACT Math, we compute the probabilities of passing Intermediate Algebra and 101 for a variety of GPA/ACT Math combinations. Results are shown in Table 4.23 and Table 4.24. A legend for the color scheme in these two tables and the remainder of the paper is given in Table 4.22.

The Intermediate Algebra model is valid for students with a Math ACT < 22 because it has little or no training data for students with a Math ACT $\geq$ 22. Since students in algebra courses at the university of our present study have always been placed based on Math ACT score and 22 was

	$x \geq 0.75$
	$0.75 > x \geq 0.50$
	$0.50 > x \geq 0.25$
	$0.25 > x$

Table 4.22: Legend for color scheme where x is the probability of passing a course.

	15	16	17	18	19	20	21
3.75	0.64	0.67	0.70	0.73	0.76	0.78	0.80
3.5	0.56	0.60	0.63	0.66	0.69	0.72	0.75
3.25	0.48	0.52	0.55	0.58	0.62	0.65	0.68
3.0	0.40	0.43	0.47	0.50	0.54	0.57	0.60
2.75	0.33	0.36	0.39	0.42	0.46	0.49	0.52
2.5	0.26	0.28	0.31	0.34	0.38	0.41	0.44
2.25	0.20	0.22	0.25	0.27	0.30	0.33	0.36
2.0	0.15	0.17	0.19	0.21	0.24	0.26	0.29

Table 4.23: Probability of passing Intermediate Algebra based on logistic regression model. Math ACT is given along the x axis and GPA score along the y axis.

	22	23	24	25
3.75	0.72	0.75	0.77	0.79
3.5	0.64	0.67	0.70	0.72
3.25	0.55	0.58	0.61	0.64
3.0	0.46	0.49	0.52	0.55
2.75	0.37	0.40	0.43	0.46
2.5	0.28	0.31	0.34	0.37
2.25	0.21	0.24	0.26	0.29
2.0	0.16	0.18	0.19	0.22

Table 4.24: Probability of passing College Algebra based on logistic regression model. Math ACT is given along the x axis and GPA score along the y axis.

the cutoff for College Algebra, only students with ≤ 22 were placed into Intermediate Algebra. Thus models have very little or no training data for students with Math ACT > 22.

We see a similar issue in the College Algebra model. Students with a Math ACT score ≥ 26 are able to enroll in calculus. Thus we do not have enough training data to make accurate predictions about students scoring above a 25. Students in College Algebra who have a Math ACT score below 22 must have taken Intermediate Algebra or Intermediate Algebra at another institution, so our training data does not have examples of students who are taking a college math course for the first time. Thus we show only the probabilities of passing College Algebra for students with Math ACT scores greater than 21.

PCA features	Int. Algebra accuracy	Col. Algebra accuracy
1	56.9%	60.6%
2	64.6%	67.4%
3	65.2%	69.4%
4	65.1%	69.3%
5	64.8%	69.4%
6	64.8%	69.5%

Table 4.25: Optimizing number of PCA features in logistic regression model

In hopes of minimizing some of the variance in the data, we train another logistic regression model for each course using the data from the average of 10 students shown in Figure 4.4. This data produced the models

$$\hat{y} = -0.90 + 5.40 \cdot \text{GPA} \ + 7.06 \cdot \text{Math ACT}$$

and

$$\hat{y} = -0.22 + 7.90 \cdot \text{GPA} \ + 2.22 \cdot \text{Math ACT}$$

for Intermediate Algebra and College Algebra respectively. Both models are on a much larger scale than the original models. The ratio of the College Algebra coefficients is 3.56 which is larger than the ratio the ratio of the original coefficient. The Intermediate Algebra model actually gives more weight to Math ACT than GPA. This is the only time we see this in all our analysis. The weight given to Math ACT is 1.3 times the weight given to GPA. The accuracies of these models on the test data are 64.4% and 67.9% respectively. Both are accuracies are approximately 1% lower than the accuracies produced by the original logistic regression models.

We also consider logistic regression models with PCA features. We use validation data to optimize the number of PCA features to use in training the model. Table 4.25 shows accuracies each number of features produces. Features are removed in reverse order so that features explaining the least variance are removed first. We find that using the first three of the six principal components

is optimal. We do not gain any more accuracy by adding principal components 4, 5, and 6.

As we expected from Section 4.1 analysis, logistic regression models weight GPA heavier that Math ACT. We see the importance of GPA in multiple way in logistic regression, most convincingly from the facts that the GPA only model is more accurate than the Math ACT only model and the coefficients given to GPA are multiple times larger than the coefficients given to Math ACT.

4.6 Interpreting Linear Discriminant Analysis

In Table 4.26, we can see that a model trained with GPA and all four ACT components does not do significantly better than a model trained with only GPA and Math ACT. We choose to do all analysis in this section with the model trained with only GPA and Math ACT. The simplicity of this model with only two variables allows us to continue the GPA/Math ACT prediction grids in two dimensions.

	GPA and Math ACT only	**GPA and all four ACT components**
Intermediate Algebra	65.3%	64.0%
College Algebra	68.4%	69.0%

Table 4.26: Compare accuracies of LDA models trained with only GPA and Math ACT vs. GPA and all four ACT components

The resulting model trained with normalized Intermediate Algebra data is

$$\hat{y} = 0.92 \cdot \text{GPA} + 0.42 \cdot \text{Math ACT}.$$

The model trained with normalized College Algebra data is

$$\hat{y} = 0.90 \cdot \text{GPA} + 0.45 \cdot \text{Math ACT}.$$

The GPA coefficient is 2.2 times the Math ACT coefficient in the Intermediate Algebra model and 2.0 times the Math ACT coefficient in the College Algebra model. Thus the Math ACT is

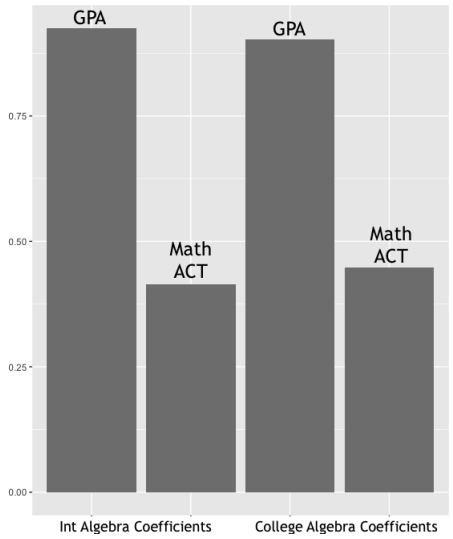

Figure 4.16: GPA is weighted 2.2 times more than Math ACT in the Intermediate Algebra model and 2.0 times more than Math ACT in College Algebra.

less important when deciding if a student will pass Intermediate Algebra than College Algebra. Figure 4.16 shows the weights of each coefficient.

Just as we did for the logistic regression models, we compute the probabilities of passing Intermediate Algebra and College Algebra for a variety of GPA/ACT Math combinations. Results are shown in Table 4.27 and Table 4.28.

As in logistic regression, LDA assigns coefficients to GPA which are more than double the coefficients assigned to Math ACT, confirming that GPA is the more important predictor.

Table 4.27: Probability of passing Intermediate Algebra based on LDA model. Math ACT is given along the x axis and GPA score along the y axis.

	15	16	17	18	19	20	21
3.75	0.62	0.65	0.68	0.71	0.74	0.76	0.79
3.5	0.55	0.58	0.62	0.65	0.68	0.71	0.73
3.25	0.48	0.51	0.55	0.58	0.61	0.64	0.68
3.0	0.41	0.44	0.48	0.51	0.54	0.58	0.61
2.75	0.34	0.37	0.41	0.44	0.47	0.51	0.54
2.5	0.28	0.31	0.34	0.37	0.40	0.44	0.47
2.25	0.23	0.25	0.28	0.31	0.34	0.37	0.40
2.0	0.18	0.20	0.23	0.25	0.28	0.31	0.34

Table 4.28: Probability of passing College Algebra based on LDA model. Math ACT is given along the x axis and GPA score along the y axis.

	22	23	24	25
3.75	0.71	0.74	0.76	0.79
3.5	0.54	0.57	0.61	0.64
3.25	0.52	0.56	0.59	0.62
3.0	0.44	0.48	0.51	0.54
2.75	0.40	0.43	0.46	0.50
2.5	0.29	0.32	0.35	0.38
2.25	0.23	0.25	0.28	0.31
2.0	0.20	0.22	0.24	0.27

4.7 Interpreting Decision Trees and Random Forests

No matter how many of the ACT features are given to the model, the same decision tree grows using only GPA and Math ACT to make decisions at nodes. The shape of the trees for Intermediate Algebra and College Algebra are the same. They both divide based on GPA at the root and then again by GPA on the branch which students with a higher GPA follow. The final splits is based on Math ACT.

For Intermediate Algebra, the first node splits the students at a normalized GPA of -0.2, below the normalized mean of zero. This translates to a GPA of 2.91 on the 4.0 scale. The students who have a GPA < 2.91 are all predicted to fail Intermediate Algebra. This student group is 37% of Intermediate Algebra students. Of the students that have a GPA > 2.91, they are splits again based on a normalized GPA of 0.64, which translates to 3.39 on a 4.0 scale. The students with a GPA > 3.39 are all predicted to pass Intermediate Algebra. This is 25% of all Intermediate Algebra students. The remaining students with 2.91 < GPA < 3.39 are split once more based on a normalized Math ACT score of -1, which corresponds to a score of 17 on the 36 point scale. Students with a Math ACT $\geq$ 17 are predicted to pass Intermediate Algebra. 32% of the Intermediate Algebra students follow this branch. Students with a Math ACT < 17 are predicted to fail. Only 6% of Intermediate Algebra students follow this branch. This tree has an accuracy of 63.2% on the testing data.

For College Algebra, the first node splits the students at a normalized GPA of -0.078, which translates to a GPA of 3.27 on the 4.0 scale. The students who have a GPA < 3.27 are all predicted to fail College Algebra. This student group is 42% of College Algebra students. Of the students that have a GPA > 3.27, they are splits again based on a normalized GPA of 0.6, which translates to 3.57 on a 4.0 scale. The students with a GPA > 3.57 are all predicted to pass College Algebra. This is 29% of all College Algebra students. The remaining students with 2.27 < GPA < 3.57 are split once more based on a normalized Math ACT score of -1.1, which corresponds to a score of 19 on the 36 point scale. Students with a Math ACT $\geq$ 19 are predicted to pass College Algebra. 25% of the College Algebra students follow this branch. Students with a Math ACT < 19 are predicted to fail. Only 5% of College Algebra students follow this branch. This tree has an accuracy of 67.9%

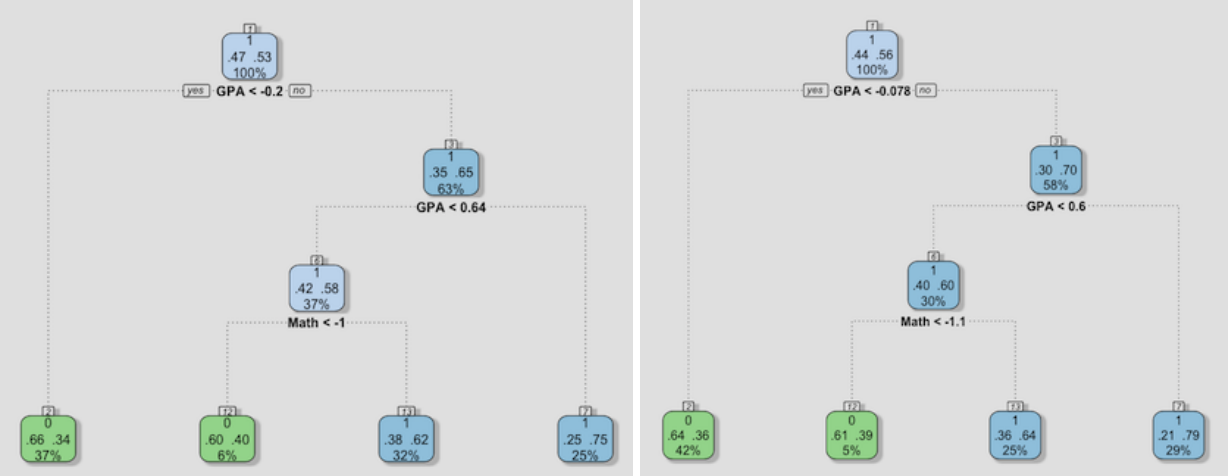

Figure 4.17: The tree for Intermediate Algebra is on the left and the tree for College Algebra is on the right.

on the testing data.

	GPA and Math ACT only	GPA and all four ACT components
Intermediate Algebra	64.9%	64.6%
College Algebra	68.4%	70.0%

Table 4.29: Compare accuracies of random forest models trained with only GPA and Math ACT vs. GPA and all four ACT components

Even though the decision tree is exactly the same whether we use only GPA and Math ACT or if we use GPA and all four components of the ACT, the random forest will not be exactly the same because only a randomized subset of the variables are used to train each node. The differences are shown in Table 4.29. Since using all four components does not improve Intermediate Algebra accuracies at all and only a 1.6% difference in College Algebra accuracies, we do predictions using the model trained on only GPA and Math ACT. However we do analyze importance of features with GPA and all four ACT components at the end of this section.

Unlike the models we have interpreted so far, the random forest models give only the final prediction, not a probability that the prediction is correct. In Table 4.30 and 4.31 we show the predictions for the same GPA and Math ACT combinations we have considered in other models. A result of 1 means that a student with this combination of GPA and Math ACT is predicted to pass and a result of 0 means they are not.

	15	16	17	18	19	20	21
3.75	0	1	1	1	1	1	1
3.5	0	1	0	1	1	1	1
3.25	0	0	1	1	1	1	1
3.0	0	0	1	1	1	1	1
2.75	0	0	0	0	0	0	1
2.5	0	0	0	0	0	0	0
2.25	0	0	0	0	0	0	0
2.0	0	0	0	0	0	0	0

Table 4.30: Intermediate Algebra outcome predictions based on random forest model. Math ACT is given along the x axis and GPA score along the y axis.

Both models point toward GPA being more significant than Math ACT. The decision tree uses GPA for the first two splits and Math ACT only for fine tuning at the tip of one branch. Classifi-

	22	23	24	25
3.75	1	1	1	1
3.5	1	1	1	1
3.25	1	1	1	1
3.0	0	1	0	0
2.75	0	0	0	0
2.5	0	0	0	0
2.25	0	0	0	0
2.0	0	0	0	0

Table 4.31: College Algebra outcome predictions based on random forest model. Math ACT is given along the x axis and GPA score along the y axis.

cations in Table 4.30 and especially in Table 4.30 indicate that splits between passing and failing students are almost completely horizontal, meaning almost completely dependent on GPA.

4.8 K Nearest Neighbors Results

As with other models, we begin by comparing the model accuracies of a model trained with only GPA and Math ACT vs. a model trained with GPA and all four components of the ACT. We find again, as shown in Table 4.32, that adding the extra components does not significantly impact the accuracy of the model so we do further analysis with the GPA and Math ACT only model.

	GPA and Math ACT only	**GPA and all four ACT components**
Intermediate Algebra	65.1%	64.0%
College Algebra	68.5%	68.2%

Table 4.32: Compare accuracies of KNN models trained with only GPA and Math ACT vs. GPA and all four ACT components

Figures 4.33 and 4.34 show predictions KNN gives for the GPA and Math ACT grids.

	15	16	17	18	19	20	21
3.75	1	1	1	1	1	1	1
3.5	1	1	1	1	1	1	1
3.25	0	1	1	1	1	1	1
3.0	0	0	1	0	1	1	1
2.75	0	0	0	0	0	1	0
2.5	0	0	0	0	0	0	0
2.25	0	0	0	0	0	0	0
2.0	0	0	0	0	0	0	0

Table 4.33: Intermediate Algebra outcome predictions based on KNN model. Math ACT is given along the x axis and GPA score along the y axis.

The separation in both the Intermediate Algebra and College Algebra tables are almost completely horizontal, which indicates that the outcomes depend much more heavily on GPA and on Math ACT.

	22	23	24	25
3.75	1	1	1	1
3.5	1	1	1	1
3.25	1	1	1	1
3.0	0	1	0	1
2.75	0	0	0	0
2.5	0	0	0	0
2.25	0	0	0	0
2.0	0	0	0	0

Table 4.34: College Algebra outcome predictions based on KNN model. Math ACT is given along the x axis and GPA score along the y axis.

4.9 Support Vector Machine Results

As with other models, we begin by comparing the model accuracies of a model trained with only GPA and Math ACT vs. a model trained with GPA and all four components of the ACT. We find again, as shown in Table 4.35, that adding the extra components does not significantly impact the accuracy of the model so we do further analysis with the GPA and Math ACT only model.

	GPA and Math ACT only	GPA and all four ACT components
Intermediate Algebra	65.3%	65.1%
College Algebra	68.4%	68.7%

Table 4.35: Compare accuracies of SVM models trained with only GPA and Math ACT vs. GPA and all four ACT components

Even though the SVM is a model with low interpretability, we will include Tables 4.36 and 4.37 with predictions of passing or failing given particular combinations of GPA and Math ACT because SVM it produces the best accuracy in the testing data. The SVM does not return a probability of passing like logistic regression and LDA do. It returns only the final prediction of pass or fail.

A result of 1 means that a student with this combination of GPA and Math ACT is predicted to pass and a result of 0 means they are not. The 0's in the lower right corner of the Intermediate Algebra predictions are most likely a result of having very little training data in this region. Since students with Math ACT > 22 are placed into College Algebra, the model does not have many

(maybe no) examples in training.

	15	16	17	18	19	20	21
3.75	1	1	1	1	1	1	1
3.5	0	1	1	1	1	1	1
3.25	0	0	1	1	1	1	1
3.0	0	0	0	1	1	1	1
2.75	0	0	0	0	0	1	1
2.5	0	0	0	0	0	0	0
2.25	0	0	0	0	0	0	0
2.0	0	0	0	0	0	0	0

Table 4.36: Intermediate Algebra outcome predictions based on SVM model. Math ACT is given along the x axis and GPA score along the y axis.

	22	23	24	25
3.75	1	1	1	1
3.5	1	1	1	1
3.25	1	1	1	1
3.0	0	1	1	1
2.75	0	0	0	0
2.5	0	0	0	0
2.25	0	0	0	0
2.0	0	0	0	0

Table 4.37: College Algebra outcome predictions based on SVM model. Math ACT is given along the x axis and GPA score along the y axis.

SVM produces the same pattern we have seen with other models. Intermediate Algebra data separation between passing and failing students depends heavily on GPA but is also impacted by Math ACT score. College Algebra data separation between passing and failing students depends almost exclusively on GPA.

Chapter 5

Overall Conclusions and Further Analysis

5.1 Remove repeating students from data set

A number of students retake algebra courses multiple times. For our purpose of placing new, incoming students, we could be more accurate if we consider a data set with only first-time college algebra students at the university. Based on the 11 years of fall semester data we use in this study, 9.3% of Intermediate Algebra students are repeaters and 11.9% of College Algebra students are repeaters. More significant than students repeating the same course are students moving from Intermediate Algebra to College Algebra. If we would only like to consider first-time algebra students, we must remove repeaters and all the students from the College Algebra data set that have taken Intermediate Algebra. This subset consists of 22.6% of all the College Algebra students considered in this study. This changes the make-up of the College Algebra students by mostly removing the students who have a Math ACT scores < 22 because these are the students originally placed in Intermediate Algebra. Figure 5.1 shows the new counts of College Algebra students by Math ACT scores. The counts of students with Math ACT scores from 22 to 25 decreased slightly, but the counts of students with Math ACT scores less than 22 each decreased by about half.

Table 5.1 shows descriptive statistics for Intermediate Algebra students without repeaters and College Algebra students without repeaters or students previously enrolled in Intermediate Algebra. Tables 5.2 and 5.3 show the differences between average scores of passing students and failing

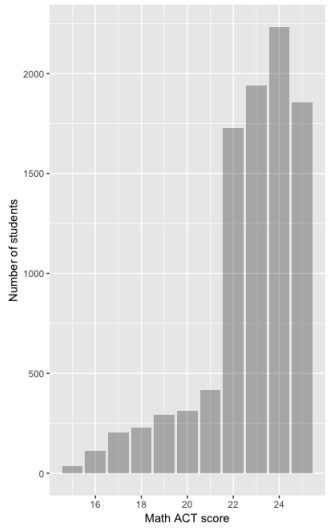

Figure 5.1: The number of students with Math ACT scores less than 22 decreases when we remove students who have taken Intermediate Algebra.

students. They give the percents change on the 4.0 scale for GPA and 36 point scale for Math ACT

When we train a logistic regression model on the remaining 90.7% of the Intermediate Algebra data we find the following equation

$$\hat{y} = 0.18 + 0.68 \cdot \text{GPA} \ + 0.27 \cdot \text{Math ACT}.$$

The GPA coefficient is now 2.5 times the Math ACT coefficient, compared the the original model in Chapter 4 which weighted GPA 2.6 times more than Math ACT. Thus removing the students who are retaking Intermediate Algebra does not change the model significantly. This model produces 67.0% accuracy on the testing data. This is an increase from the 65.5% accuracy the original model in Chapter 4 achieved. We then make predictions on students with GPA/Math ACT combinations shown in Figure 5.4. Notice that this table is computed differently from the tables in Chapter 4 because GPA predictions are shown in a range. This was achieved by taking predictions at 6 points within each range and averaging the probabilities. The GPA intervals between points were 0.05 and both endpoints are included in each averaged prediction.

When we remove all the repeating students and the students who have taken Intermediate

Score and student group	Average	Median	Standard deviation
GPA of all Int. Algebra students	3.06	3.09	1.2
GPA of Int. Algebra failing students	2.91	2.91	1.5
GPA of Int. Algebra passing students	3.20	3.24	0.8
GPA of all College Algebra students	3.30	3.35	1.2
GPA of College Algebra failing students	3.05	3.11	0.6
GPA of College Algebra passing students	3.46	3.50	2.0
Math ACT of all Int. Algebra students	18.6	19	2.1
Math ACT of Int. Algebra failing students	18.3	18	2.2
Math ACT of Int. Algebra passing students	18.8	19	1.9
Math ACT of all College Algebra students	23.0	23	2.3
Math ACT of College Algebra failing students	22.6	23	2.5
Math ACT of College Algebra passing students	23.2	24	2.0

Table 5.1: Descriptive statistics for student groups without repeat students

	Average GPA (pass)	Average GPA (fail)	Diff.	Percent of 4.0
Int. Algebra:	3.20	2.91	0.29	7.3%
Col. Algebra:	3.46	3.05	0.41	10.3%

Table 5.2: Differences in average GPA scores without repeat students

	Average Math ACT (pass)	Average Math ACT (fail)	Difference	Percent change
Int. Algebra:	18.8	18.3	0.5	1.4%
Col. Algebra:	23.2	22.6	0.6	1.7%

Table 5.3: Differences in average GPA scores students without repeat students

	15	16	17	18	19	20	21
3.76 - 4.0	0.68	0.70	0.73	0.76	0.78	0.80	0.82
3.51 - 3.75	0.61	0.64	0.67	0.70	0.72	0.75	0.77
3.26 - 3.5	0.53	0.57	0.60	0.63	0.66	0.69	0.71
3.01 - 3.25	0.46	0.49	0.52	0.55	0.59	0.62	0.65
2.76 - 3.0	0.38	0.41	0.45	0.48	0.51	0.54	0.58
2.51 - 2.75	0.31	0.34	0.37	0.40	0.44	0.47	0.50
2.26 - 2.5	0.25	0.28	0.30	0.33	0.36	0.39	0.42
2.01 - 2.25	0.20	0.22	0.24	0.27	0.30	0.32	0.35

Table 5.4: Predictions on first time Intermediate Algebra students. Every cell is the average of six probabilities computed based on six points (including end points) in the corresponding GPA range.

Algebra from the College Algebra data set and train a logistic regression model on the remaining 77.3% of the data we find the following equation

$$\hat{y} = 0.52 + 0.76 \cdot \text{GPA} + 0.20 \cdot \text{Math ACT}.$$

The GPA coefficient is now 3.8 times the Intermediate Algebra coefficient. Since the original model in Chapter 4 weighted the GPA coefficient 2.5 times more than the Math ACT coefficient, this is a significant change. This model achieves 69.1% accuracy compared to the 68.5% accuracy achieved by the original model. Table 5.5 gives probabilities the of passing College Algebra given by this model. Probabilities here were averaged in exactly the same way as Table 5.4.

	22	23	24	25
3.76 - 4.0	0.80	0.81	0.82	0.83
3.51 - 3.75	0.74	0.75	0.76	0.78
3.26 - 3.5	0.66	0.68	0.70	0.71
3.01 - 3.25	0.59	0.60	0.62	0.64
2.76 - 3.0	0.50	0.52	0.54	0.56
2.51 - 2.75	0.42	0.44	0.46	0.47
2.26 - 2.5	0.34	0.36	0.37	0.39
2.01 - 2.25	0.27	0.28	0.30	0.31

Table 5.5: Predictions on first time Intermediate Algebra students. Every cell is the average of six probabilities computed based on six points (including end points) in the corresponding GPA range.

5.2 Recommendations for algebra course placement

We would like to make recommendations to improve the success rate of students in these courses. All models show that high school GPA is more important than Math ACT scores when predicting whether or not a student will pass Intermediate Algebra or College Algebra. Since the university is currently places students based only on Math ACT scores, we recommend that they take high school GPA into account as well when making placement decisions. This could be done by choosing a College Algebra cutoff chance of passing and using the percent chances of passing based on logistic regression model in Figure 4.24 or LDA model in Figure 4.28. For example, if the GPA and Math ACT combination falls into a cell with a chance of passing that is higher than 50%, then the student will be placed in College Algebra and otherwise in Intermediate Algebra.

In Table 5.7 we make recommendations based on the results of the logistic regression models in Section 5.1. We choose this model because we would like to assume that we are placing students taking algebra for the first time at the university. A color legend with course we recommend the student takes along with criteria we base this choice on is given in Table 5.6.

For comparison we also include tables showing how students were placed historically. Table 5.8 shows how students were placed before Fall 2016 and Table 5.9 shows the recent change in

Color	Recommendation	Criteria
	Calculus Enhanced	≥ 0.75 probability of passing Col. Algebra
	Col. Algebra	≥ 0.50 probability of passing Col. Algebra or ≥ 0.75 probability of passing Int. Algebra
	Col. Algebra Enhanced	≥ 0.32 probability of passing Col. Algebra or ≥ 0.50 probability of passing Int. Algebra
	Int. Algebra	< 0.32 probability of passing Col. Algebra or ≥ 0.32 probability of passing Int. Algebra
	Int. Algebra with supplemental work	< 0.32 probability of passing Int. Algebra

Table 5.6: Legend for color scheme for student placement and criteria for placement

	≤ 15	16	17	18	19	20	21	22	23	24	25
3.76 - 4.0	0.68	0.70	0.73	0.76	0.78	0.80	0.82	0.80	0.81	0.82	0.83
3.51 - 3.75	0.61	0.64	0.67	0.70	0.72	0.75	0.77	0.74	0.75	0.76	0.78
3.26 - 3.50	0.53	0.57	0.60	0.63	0.66	0.69	0.71	0.66	0.68	0.70	0.71
3.01 - 3.25	0.46	0.49	0.52	0.55	0.59	0.62	0.65	0.59	0.60	0.62	0.64
2.76 - 3.0	0.38	0.41	0.45	0.48	0.51	0.54	0.58	0.50	0.52	0.54	0.56
2.51 - 2.75	0.31	0.34	0.37	0.40	0.44	0.47	0.50	0.42	0.44	0.46	0.47
2.26 - 2.50	0.25	0.28	0.30	0.33	0.36	0.39	0.42	0.34	0.36	0.37	0.39
≤ 2.0	0.20	0.22	0.24	0.27	0.30	0.32	0.35	0.27	0.28	0.30	0.31

Table 5.7: Recommendation for student placement based on logistic regression models with only students taking algebra for the first time at the university. Each cell with an ACT Math score from 15 to 21 shows the probability of passing Intermediate Algebra and each cell with an ACT Math score from 22 to 25 shows the probability of passing College Algebra.

	≤ 15	16	17	18	19	20	21	22	23	24	25
3.76 - 4.0											
3.51 - 3.75											
3.26 - 3.50											
3.01 - 3.25											
2.76 - 3.0											
2.51 - 2.75											
2.26 - 2.50											
≤ 2.0											

Table 5.8: Placement method used before Fall 2016

placement that moves in the direction of our recommendations.

	≤ 15	16	17	18	19	20	21	22	23	24	25
3.76 - 4.0											
3.51 - 3.75											
3.26 - 3.50											
3.01 - 3.25											
2.76 - 3.0											
2.51 - 2.75											
2.26 - 2.50											
≤ 2.0											

Table 5.9: Placement method used Fall 2016

5.3 Future Work

In the same way that we considered individual components of the ACT in this project, it may be helpful to consider students' GPA from each year of high school individually. Studies have shown that in general, the later years of high school have far more predictive power than earlier years

(Bulman, 2017).

In Bulman's study, he found that, "one additional GPA point in 11th grade is associated with an increase of 16 percentage points in the probability of graduating from a state university, compared to an increase of 5 percentage points for a GPA point in 9th grade. Similarly, the later GPA point is five times more predictive of dropping out of college within two years. Giving greater weight to more recent performance is likely to improve the expected outcomes of selected students...Giving equal weight to each grade level necessarily discards a large amount of information stemming from substantial variation in student performance over the four years of high school. (Bulman, 2017)"

The university in our study may find that this applies to mathematics courses too and incorporating only junior and senior year high school GPA's may give even better predictions that a single GPA over all four years of high school.

We may be able to further improve the models given in this paper by eliminating a subset of the training data. In remedial courses in institutions across the country, student participation is an issue. The data base of algebra students for our present study contains attendance records for all students. It could be useful to calculate the percent of student who never attend class in the first month of the semester. This is likely to be a significant percentage. We would like to remove this subset of students and retrain all models given in this study. This may reduce some of the variation and help make more accurate predictions for the student who do choose to attend class. Since we cannot predict which students will be the ones who enroll but do not attend class, we simply train the model on the students who do participate in the course and make predictions about incoming students assuming they attend class.

We would also be interested in making predictions mid-semester about students' success in courses. The algebra student database contains data on attendance, homework and exams for every student. The algebra courses contain five units throughout the semester. We would be interested to know whether we can predict if a student will pass the course with information from only the first unit. We would also be interested to know if there is a way to predicted whether an Intermediate Algebra student might pass College Algebra.

References

Barata, G., Gama, S., Jorge, J., , & Gonçalves, D. (2016). Early prediction of student profiles based on performance and gaming preferences. *IEEE Transactions on Learning Technologies*, (3), 272 – 284.

Bettinger, E. P., Evans, B. J., & Pope, D. G. (2013). Improving college performance and retention the easy way: Unpacking the act exam. *American Economic Journal: Economic Policy*, (2), 26–52.

Bowen, W. G., Chingos, M. M., & McPherson, M. S. (2011). *Crossing the Finish Line: Completing College at America's Public Universities*. Princeton University Press.

Breiman, L. (1999). Random Forests–Random Features, Technical Report 567. *Statistics Department, University of California Berekley*.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, (3), 199–231.

Bressoud, D. & Hsu, E. (2015). Maa national studies of college calculus, chapter 5 placement and student performance in calculus i. *Mathematical Association of America*.

Bulman, G. (2017). Weighting recent performance to improve college and labormarket outcomes. *Journal of Public Economics*.

Cen, L., Ruta, D., Powell, L., Hirsch, B., & Ng, J. (2016). Quantitative approach to collaborative learning: performance prediction, individual assessment, and group composition. *International Journal of Computer-Supported Collaborative Learning*, 11(2), 187 – 225. An optional note.

Drake, S. (2010). Placement into first college mathematics course: A comparison of the results of the michigan state university proctored mathematics placement examination and the unproctored mathematics placement examination. *Michigan State University, ProQuest Dissertations Publishing*.

Fain, P. (2015). Finding a new compass. *Inside Higher Ed*.

Geiser, S. & Santelices, M. V. (2007). Validity of high-school grades in predicting student success beyond the freshman year:high-school record vs. standardized tests as indicators of four-year college outcomes. *Center for Studies in Higher Education, University of California, Berkeley*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning DataMining, Inference, and Prediction. *Springer Series in Statistics*.

Hodara, M., Jaggars, S. S., & Karp, M. M. (2012). Improving developmental education assessment and placement: Lessons from community colleges across the country. *Community College Research Center, Teachers College, Columbia University*.

Huang, S. & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*.

Jaggars, S. S. & Stacey, G. W. (2014). What we know about developmental education outcomes. *Community College Research Center, Teachers College, Columbia University*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics.

Jarrett, E. (2000). Evaluating the persistence and performance of "successful' precalculus students in subsequent mathematics courses. *Thesis in Mathematics, Degree of Master of Science*.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers and Education*.

Martorell, P. & McFarlin, I. (2011). Help or hindrance? the effects of college remediation on academic and labor market outcomes. *Review of Economics and Statistics*, (2), 436–454.

Marwick, J. D. (2002). Charting a path to success: How alternative methods of mathematics placement impact the academic success of community college students. *Graduate Dissertations and Theses at the University of Illinois*.

Miller, S. R. (1998). Shortcut: High school grades as a signal of human capital. *Education Evaluation and Policy Analysis*, (4).

Moran, K. (2008). Examining the mathematical preparedness of first year college students entering college directly from traditional high schools in connecticut. *Research and Teaching in Developmental Education*, (1), 30–44.

Murphy, K. P. (2012). *Machine Learning A Probabilistic Perspective*. The MIT Press.

Ng, A. (2016). Computer science 229 machine learning.

Pistilli, M. D. & Arnold, K. E. (2010). In practice: Purdue signals: Mining real-time academic data to enhance student success. *About Campus: Enriching that Student Learning Experience*, (3), 22–24.

Preast, K. L. (1998). Placement of students in mathematics courses according to tasp test scores and course reading levels (Texas academic skills program). *Ed.D. dissertation, Texas A & M University–Commerce*.

Reddy, A. A. & Harper, M. (2013). Mathematics placement at the University of Illinois. *PRIMUS*.

Rothstein, J. M. (2004). College performance predictions and the sat. *Journal of Econometrics*, (1).

Scott-Clayton, J. (2012). Do high-stakes placement exams predict college success? *Community College Research Center, Teachers College, Columbia University*.

Sonnert, G. & Sadler, P. M. (2014). The impact of taking a college pre-calculus course on students' college calculus performance. *International Journal of Mathematical Education in Science and Technology*, (8), 1188–1207.

Appendix A

Appendix

All programming for this thesis was done in R via RStudio. The packages used to clean, process, visualize, and model data are the following:

- tidyr
- xtable
- car
- caret
- stats
- MASS
- rpart
- rpart.plot
- rattle
- e1071
- plyr
- class
- randomForest
- ggplot2
- psych
- reshape2