Developing a Machine Learning Framework for Student Persistence Prediction

by

Alexis Wade

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved April 2021 by the
Graduate Supervisory Committee:

Esma Gel, Co-Chair
Hao Yan, Co-Chair
Theodore Pavlic

ARIZONA STATE UNIVERSITY

May 2021

ABSTRACT

Student retention is a critical metric for many universities whose intention is to support student success. The goal of this thesis is to create retention models utilizing machine learning (ML) techniques. The factors explored in this research include only those known during the admissions process. These models have two goals: first, to correctly predict as many non–returning students as possible, while minimizing the number of students who are falsely predicted as non–returning. Next, to identify important features in student retention and provide a practical explanation for a student's decision to no longer persist. The models are then used to provide outreach to students that need more support. The findings of this research indicate that the current top performing model is Adaboost which is able to successfully predict non–returning students with an accuracy of 54 percent.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

At most universities around the world, including Arizona State University (ASU), student retention is a critical metric. Positive retention rates signal that students are successfully completing the programs they entered and further suggests that the university is providing the resources that students need to thrive. High retention is attractive to degree–seeking students who are in the process of deciding between universities as selecting a university with low retention may result in failure to complete the program. In Fulton Schools of Engineering (FSE), retention is of even greater concern. The infamously difficult course loads of Science, Technology, Engineering, and Math (STEM) majors can threaten a student's choice to remain both in engineering and at ASU.

In FSE, the office of Academic and Student Affairs (ASA) is interested in this problem as a means of ensuring students' academic and social success at ASU. Prior to COVID–19, student retention in FSE was stagnant. In general, around 88 percent of first–time freshman students returned for their second semester at ASU. While these metrics are already high, the ASA team wanted to better understand which students were leaving in order to make data–driven decisions to improve students' overall experience. ASA has spent the last several years working with student research teams at ASU to better understand retention. Unfortunately, due to the quantity and complexity of the data, not much progress has been made outside of data preparation. This research branches in a new direction and hopes to employ machine learning techniques to advance ASA's retention modeling progress. This research aligns with those goals established by ASA. The first goal is create statistical models to create an

early identification tool for students who are at risk of not returning. The second goal is to better understand these students: what exactly is driving these students to leave? In theory, the same models that flag students are able to identify features relevant to a student's choice to stay or leave. Thus, these models are applied, evaluated, and compared to gather information on which features are top performers.

The models discussed in this research utilize machine learning methods to better handle the complexity of the data and the problem itself. For simplification purposes, only domestic, first–time students are considered. In addition, the factors explored in each of the models consider only those characteristics known about students prior to the beginning of their first fall semester.

The primary purpose of this research is to establish a methodology for successfully predicting students who are not likely to return with an accuracy of 50 percent or greater. Though machine learning techniques are utilized frequently in this field, researchers typically select only a few methods to analyze. This research steps away from that mold by recognizing that all machine learning techniques have strong potential for prediction. Moreover, there is no "best" method. Thus, in this research, many different methods are tested and evaluated. In addition, some techniques that have not seen much use in this field are employed, including Adaboost and Gradient Boosting.

Beyond the technical scope, several interesting variables are included in this analysis. For example, ASU uses a Calculated Index (CI) Score which provides a measure of students' high school academic performance. By including this feature, this research validates its continued use by the university. Moreover, as this research discusses STEM students, a variable that tracks students' enrollment in math classes at ASU is included. While it seems logical that strong academic students are more likely to succeed, it is important to verify whether this is actually the case in FSE.

Overall, this research hopes to provide analysis on many potential machine learning techniques in this field, including several which are underutilized, expand the classical set of features considered by retention modelers, and provide an additional application for early-identification retention modeling in a STEM setting. To accomplish this, this research provides a history of retention modeling, a discussion of which features have been explored in the past and are still of interest today, and how machine learning techniques have been applied in previous applications. Following this, the full data set is discussed, including the methodology for processing the data and which features, as compared to those discussed in the literature review, were ultimately selected. Then, the machine learning models utilized in this application are presented, alongside several modeling strategies to improve the prediction quality. Finally, the results of this research are shared, the best performing model is selected, and the features selected by each of the models are discussed in further detail in an attempt to understand the results practically.

Chapter 2

LITERATURE REVIEW

Student retention is a widely studied field with research dating back to the early 20th century. However, universities did not consider retention as an impactful metric until the 1970s when college enrollment began to decline (Seidman, 2005). By this time, universities gradually began to perceive retention as a strategy and wanted to understand which students were leaving and why.

The first model of student retention came from Spady (1967), who argued that social factors were the most relevant to a student's choice to persist at university. One of the most cited retention researchers of this period is Vincent Tinto. His work, in agreement with Spady, discusses the impact of students' social characteristics on attrition. Tinto and Spady argued that, above all other factors, a student's successful integration into college life and their intent to continue are the strongest predictors for their persistence in a program (Tinto, 1975, 2017). Moreover, freshmen students are at the highest risk of leaving due to their lack of progress in their respective field (Tinto, 2013). Students who have already completed several core requirements are more likely to stay due to an internal cost–benefit analysis; if the student leaves late in their academic career, it would be a waste of the extensive resources they have already devoted their success at university.

Though several researchers have proposed adjustments to Tinto's notions, there is a dearth of evidence that refutes his conclusions. A model proposed by Bean (1980) suggests that there is little theoretical justification for Tinto and Spady's work. However, the alternative models proposed by Bean fail to account for a large portion of the variation in persistence (Seidman, 2005). More recently, several extensions

of Tinto's work have shown promising results. These applications have revealed that student satisfaction, confidence, student–university fit, and other social characteristics are all pertinent to student success (Schreiner and Nelson, 2013; Wright *et al.*, 2013; Bowman and Denson, 2014; Bowman *et al.*, 2019; Berger and Braxton, 1998). The amount of research that supports Tinto's claims suggests that, in retention modeling, factors should attempt to capture behavior both within and outside of the classroom setting.

While it is clear that social characteristics have an impact on student retention, the question then becomes: how can these characteristics be quantified? This research assumes that academics play an important role in STEM student integration. As such, high school academic performance and a student's entering math class are used as potential predictors of successful integration. Here, the assumption is made that students who did well in high school and have placed into math classes that meet FSE's requirements are likely to continue their education because these students fit into the mold that is expected of most STEM students. Alternatively, students who do not meet these expectations may feel separated from their peers. In addition to these academic factors, two more social factors are utilized to assess student integration. First, whether a student attended an engineering camp the summer prior to enrollment, which provides students with early opportunities to meet peers and professors, and second, if the student resided in on–campus housing. A first generation flag is also considered, as students who are the first in their family to attend university may have a harder time integrating for a number of reasons.

With improvements in technology, recent models in the field of retention include far more than merely social characteristics. One of the most extensive areas of research regards static characteristics, such as race and sex. In terms of race, several studies agree that underrepresented minority (URM) students are more likely to succeed

when strong support networks are available to them (Baker and Robnett, 2012; Chang *et al.*, 2014). Moreover, diversity seems to play a role in the success of the student body (Chang *et al.*, 2004). As with race, successful integration positively affects the retention of the sexes (Ayers, 2017). In general, research on static characteristics shows that better support networks for higher risk students increases the probability of attrition. This research includes URM status and gender as identified by the student.

Another factor discussed frequently in retention analysis is the economic status of students. According to one study, students who transition immediately from high school to college are more likely to graduate, assuming that their financial needs are met (Cabrera, 2003). At a glance, this finding appears to suggest that age upon college entrance and transfer status may affect retention. However, according to the same study, it is high financial need that serves as the cause both for transferring and late arrivals to an institution. Though this finding is not unique, research suggests that financial aid can reduce the risk of non–persistence. For example, any financial offering to a student increases their chance of attending college by nine percent. Beyond that, Pell Grants, loans, and work–study all have a positive impact on enrollment (Chen and DesJardins, 2008; McKinney *et al.*, 2015). Thus, financial need status and financial aid offerings, including Pell Grants, are also considered as features in these retention models.

Many characteristics, even beyond those discussed, have been identified as important to student persistence. As such, a wide range of factors are included in the modeling performed in this research. In summary, this includes static characteristics, like race and sex, financial characteristics, like scholarship and financial need, and several social and academic characteristics, which help define student integration. The one limitation to the factors included is that no factors beyond those known before

the start of a student's first fall term as a freshman college student are analyzed.

Of all the research performed on student retention, there is no clear methodology for developing retention models. Earlier research, such as that of Tinto, Spady, and Bean, relied on theoretical models and simple regression techniques. With the development of better user interfaces and algorithms, more complex statistical and machine learning (ML) methods have been employed in recent research. Research on student retention is heavily saturated with the application of machine learning models in predicting student persistence (Chen *et al.*, 2018; He *et al.*, 2018; Muncie, 2020). These models aim to effectively identify students who are at risk of not returning and develop frameworks that highlight key features relevant to a student's decision to stay or leave. It is not uncommon for researchers to present a single machine learning method that produces a prediction of at–risk students. The general procedure follows that researchers utilize one or several machine learning techniques, apply these to the freshman student population, and supply a model for early identification of at–risk students. Many ML methods have been explored, with the most common being neural networks, decision trees, logistic regression, and support vector machine (Sepulveda, 2020). A clear gap exists in the application of Adaptive and Gradient Boosting methods. This may be due to the fact that methods such as logistic regression provide results that offer straightforward interpretations on the qualitative impacts of certain features on retention, though this research will investigate this further.

This research, similar to these previous efforts, has the primary intention of establishing models for the early identification of at–risk students. A distinguishing factor for this research is that only factors known prior to the induction of students will be included in analysis, providing that a set of at–risk students can be determined before the semester begins. This research also explores the impact of Adaptive and Gradient Boosting methods, which are not commonly used. Moreover, this research investi-

gates the similarities and differences between multiple models to identify and analyze important features of student retention at ASU. The intention of this research is not to argue in favor of one modeling method over another. Rather, many techniques are used to create many different models. The best performing predictive model is then selected using Area Under the Curve (AUC) and F1 score, acknowledging that any change in the sample set, features, or other inputs might alter which modeling method performs best. Moreover, the remaining models are evaluated for interpretability and utilized to explain any findings further.

Chapter 3

DATA

3.1 Dataset Description

The data set consists of 9,483 domestic, first–time freshman FSE students from 2015 to 2018. Each year represents a new cohort of students. This data set is a supervised learning set in which the label is whether or not the student returned for their first spring semester. Across this four–year range, only 463 students did not persist for their first spring. This indicates that the data set is heavily unbalanced which requires detailed analysis that is described later on. Though there are plenty of features available, the data itself is relatively simple. There are primarily binary features in the set with very few continuous features. To aid in visualization of the data set, data for a dummy student is provided in Figure 3.1, using fabricated data.

The data utilized in this project was retrieved directly from ASU'S Management Information Analysts. The 2015 to 2017 data is the product of several student teams' capstone and Fulton Undergraduate Research Initiative (FURI) projects.

3.2 Preprocessing

Much of the data utilized in this project was collected and processed by student teams who had previously worked on this project. The features chosen for analysis were based on those that they had already collected. Regardless, some feature engineering was required to continue work. Many of the features available for analysis were binary, meaning that they had only two values. To prepare this data, the features were transformed from their original text characteristics to values of zero and

Feature	Dummy Student Data
URM?	0
Gender?	1
First Gen?	0
ASU Scholarships?	1
CI Score	138
No CI?	0
Attended E2?	1
Meets MAT Requirements?	1
Resident?	1
Ever in Housing?	1
Enrolled in ASU 101?	1
Honors?	0
Pell Eligible?	0
Very High	0
High	0
Moderate	0
Low	1
Very Low	0

Figure 3.1: An Example Observation of the Set, Using Fabricated Data

one. Zero implies that the student does not meet the characteristic described by that feature, whereas one implies that the student does meet the characteristic described by that feature. For example, when considering a student's URM status, zero was assigned to students who did not meet one of the URM categories, and one was assigned to the students who met any of the URM categories. In addition, a scaling factor was applied to the data set to prevent variables such as `CI Score`, which is on a much larger scale than the binary data, from being weighted unnecessarily high by the models.

This set also contains the label, which is whether or not the student enrolled in spring. Since it is more important to correctly predict students who choose not to re–enroll than students who do, one was assigned to students who do not persist, and

zero to students who do persist. This is opposite to the structuring for the features in this set.

3.3 Missing Data

Because ASU has teams that continuously track student data, there are not many issues with missing data. Typically, if data is not present, it indicates that the student does not match that criteria. For example, when a student has no scholarships listed, it is because they simply do not have scholarships, not that there is missing scholarship information. Nonetheless, visualizations were performed to identify if the proportion of blanks in the set appeared greater than what might be expected. None of the available features appeared to have an issue, save for Calculated Index (CI), where missing data is expected. Missing `CI Score`s are relatively common; since `CI Score` is calculated from standardized testing scores and high school performance, any student with a different background is at risk of not having a `CI Score`. For this reason, only domestic students were considered in this analysis – international students often do not have the same standardized testing/high school performance data available to calculate a `CI Score`. However, `CI Score` is also difficult to track for students who have non–traditional high school backgrounds, such as those who were home–schooled. As a result, approximately three percent of `CI Score`s were missing for the overall set. To combat this, the missing values were replaced with zeroes and a new variable, `No CI?`, was created to identify those students who were missing `CI Score`s. This feature was created to determine if there was anything unique about students without `CI Score`s that might affect their enrollment. A similar issue appears again when considering financial data. Students are not required to apply for aid, but to do so, they must submit their Free Application for Federal Student Aid (FAFSA). If a student chooses not to submit the FAFSA, they cannot be included in any of

the financial need brackets. Thus, the financial need criteria were broken into several columns – a student who does not meet any of these criteria is identified by a zero in all of the financial need features.

3.4 Correlation

Since several features were available for analysis, many of which have logical connections to one another, it was important to determine any potential correlations between features prior to analysis. As such, a correlation matrix was created, as seen in Appendix A. A meaningful correlation that appeared was a strong correlation between `CI Score` and `No CI?`. After consideration, both were selected to remain in the analysis. This decision was made after testing models first with `No CI?`, next with `CI Score`, and finally with both variables included. The results showed that both were often necessary in the final models. Another correlation that appeared in the models was a moderate relationship between `Pell Eligible?` and `Very High/High` financial need. Since the correlation was not as extreme as the previous example, the variables were kept and monitored throughout the modeling process.

3.5 Features

There are approximately 108 features available for processing in this set. Since the scope of this research only covers factors that are known prior to a student's enrollment, this set diminishes to 18 features of interest. A majority of these features are binary and have been coded with values of zero and one, where zero indicates that a student does not have the feature in question and one indicates that a student does have that feature. The full set of features used in this analysis is provided in Appendix B.

3.5.1 Label

The class of interest in this research is `Enrolled Spring?`, which is a binary variable that returns zero if a student enrolls in classes in the spring following their first fall, and one if they do not re–enroll. This data set is unbalanced as only 463 out of 9,483 students do not persist to the first spring.

3.5.2 ASU Scholarships?

ASU Scholarships? is a binary variable with a value of one if the student has any scholarships and zero if otherwise.

3.5.3 CI Score

`CI Score` is the only continuous variable included in this set. A `CI Score` are automatically calculated by the university and are calculated as the intersection between Arizona Board of Regents (ABOR) Grade Point Average (GPA)/high school rank and Scholastic Assessment Test (SAT)/American College Testing (ACT) scores. If any of these variables are missing, a `CI Score` is not calculated. As such, many international or home–schooled students do not have a `CI Score`. Thus, international students were removed from the data set. Only about three percent of students in this set do not have a `CI Score`. Students without a `CI Score` were assigned a `CI Score` of zero.

3.5.4 No CI?

Instead of removing domestic students without a `CI Score`, a binary variable `No CI?` was created with a value of one if the student has no `CI Score` and zero if otherwise. This variable was included to identify if not having a `CI Score` is an

important feature in re–enrollment.

3.5.5 Meets MAT Requirements?

`Meets MAT Requirements?` is a binary variable that checks if the student is enrolled in the required (or higher) math class for their first semester. If the student meets or exceeds this requirement, they receive a value of one and zero if otherwise. Difference in requirements across both years and majors were considered when calculating this variable. However, this variable is dynamic – a student might choose to enroll late or drop a class early in the semester, meaning that this variable must be tracked over time. This data set considers the math class the student was enrolled in at the start of the first fall semester.

3.5.6 Financial Need

Financial need is not a single variable in the set, but represented by five individual variables: `Very High`, `High`, `Moderate`, `Low`, and `Very Low` – each indicating a student's level of financial need. These variables are all binary. A student only falls into one of these brackets, and as such, will have a one in the level of financial need they qualify for, and a zero in the remaining features. Since financial need is determined after a student submits their FAFSA, there is some missing data based on students who chose not to submit. This was identifiable by all features equaling zero.

Chapter 4

MODELING

4.1 Intent

The goal of this research is two–pronged: first, to provide a model that predicts at least 50 percent of those students who will not return successfully. Second, to understand what features play a role in a student's decision to stay or leave. The final model needs to maintain a good balance of both interpretability and predictive power. However, models that do not perform well predictively are still useful for their ability to explain relevant features to a wider audience.

4.2 Initial Steps

As stated previously, several groups worked on this project prior to the start of this research. These groups primarily collected and processed the data, but also attempted early analysis using basic regression techniques. Due to the binary outcome and the complexity of the data, these teams were able to achieve only approximately 10 percent prediction accuracy. This project began with several attempts to create Logistic Regression models in the statistical software, JMP. Logistic Regression is a popular modeling technique that allows its users to find the probability that a data point belongs to a certain class using a structure similar to linear regression (James *et al.*, 2013). It is often employed due to its ability to model binary outcomes. Unfortunately, due to the complexity of the software, prediction accuracy only improved to approximately 30 percent. Several other softwares were explored, such as Matlab and RapidMiner, but the need for direct control over model parameters greatly out-

weighed the convenient graphical user interfaces. Thus, the transition to Machine Learning in Python was made.

4.3 Feature Selection

Feature selection is a unique challenge, particularly when modeling a classification problem. Most ensemble methods, such as Gradient Boosting and Adaboost, do not require feature selection due to their ability to select random subsets of features on each iteration. Regardless, Logistic Regression with an L1 penalty was applied to identify if any features needed to be removed prior to analysis. This is one of the only models that is able to perform feature selection (Ng, 2004). The results, which are discussed in detail later, showed that no features needed to be removed from the set.

4.4 Tuning Parameter Selection

K–fold cross validation is a technique that splits the training data into subsets, K – one of which are used for training a model, and one of which is used as a validation set (de Rooij and Weeda, 2020). Each subset is tested as a validation set and the results identify the best tuning parameter. First, the original data set was split into a 75 – 25 training and testing split. All modeling methods applied ten–fold cross validation to determine their best model tuning parameters for the final iteration of each model. Ten folds were chosen in order to minimize model variance. Only Gradient Boosting employed five–fold cross validation, as ten–fold required too much time to compute. This method also provided a way to calculate feature importance which was utilized when no coefficients were available for a model.

4.5 Threshold Selection

Since each model outputs a prediction probability for the observations in the test set, a threshold is utilized to choose how the model classifies each observation. The standard threshold is .50, which indicates that an observation with a prediction probability greater than .50 is classified as a one, or a non–returning student, and an observation with a prediction probability less than .50 is classified as a zero, or a returning student. Changing the value of the threshold is very important and often improves the performance of the model. The threshold for all models was adjusted both above and below .50 to find the best value of the threshold. The best threshold was determined by the value of the F1 score for that model, which estimates the model's accuracy.

4.6 Model Evaluation

As mentioned previously, the unbalanced nature of the data set poses a problem for the model output. To combat this, evaluation techniques that consider unbalanced techniques were applied. The primary evaluation technique was F1 score, which considers both precision and recall when measuring modeling accuracy (Tharwat, 2018). Precision evaluates the model directly and identifies how accurate the results are for both values of the label. Recall indicates how well the model correctly predicts the class of interest, or in this case, correctly predicts non–returning students. The higher the F1 score, the better the model performance. The second technique for evaluation was AUC, which aided in the identification of the best possible threshold. For each model, the AUC and Receiver Operating Characteristic (ROC) curve were employed primarily to determine if the model was performing better than a coin toss. Similar to F1 score, the higher the AUC, the better the performance.

4.7 Modeling Techniques

Many techniques were applied in this research, both with low interpretability and high interpretability. There are different advantages to both low interpretability and high interpretability methods. High interpretability methods are often very easy to understand - in the case of student retention, they not only produce predictive models, but information about how features impact retention. For example, a high interpretability model not only provides predictive information, but how much and in what direction a certain feature impacts retention outcomes. Alternatively, it is difficult to achieve this same information with low interpretability models, as there often is not a way to easily understand how the features impact the model. However, these types of models often perform much better predictively. As such, low interpretability models are typically suitable for providing early identification of at-risk students, while high interpretability models are better for making qualitative assumptions about how the features ultimately impact students.

Of the high interpretability methods, Naive Bayes, Logistic Regression, and Decision Trees were analyzed. The Naive Bayes algorithm is a very simple model, which makes the broad assumption that the features are independent. Following this assumption, the probability that a given data point is of a particular class is calculated according to Bayes Theorem (Domingos and Pazzani, 1997). Logistic Regression is another popular technique, as mentioned previously. For this analysis, both an L1, which involves feature selection, and L2 penalty were used. Finally, Decision Trees start with the entire data set at the "root" of the tree and create splits from this root, called nodes, with the goal of producing homogeneous sets in the final nodes, called leaves (James *et al.*, 2013).

Of the low interpretability methods are Random Forest, Gradient Boosting, and

Adaboost. Bagging, also called bootstrap aggregating, is an ensemble method that applies decision trees to randomly sub–sampled sets of the data and and outputs the average results of those trees (Breiman, 1996). Random Forest is an enhanced version of Bagging with a similar procedure, the primary difference being that it randomly sub–samples both data and features for each of its decision trees (Tin Kam Ho, 1995). Gradient Boosting and Adaboost, or Adaptive Boosting, are the last two ensemble models utilized in this research. They are very similar techniques – both rely on weak learners, such as decision trees, to 'boost' performance by creating new weak learners to accommodate the gaps in earlier learners. The primary difference between the two methods is that Adaboost adjusts the performance of its learners by weighting previously misclassified observations more heavily; Gradient Boosting adds learners to the model that minimize the overall loss (Freund *et al.*, 1999; Friedman, 2001).

Finally, the question remains as to how each of these methods handle unbalanced data. Logistic Regression, Decision Trees, and Random Forest all struggle to handle data with unbalanced classes. To counteract these issues, a weight was applied to allow each class even consideration by each of the models. Naive Bayes, Gradient Boosting, and Adaboost do not have these functionalities. However, the nature of their techniques ensures that they are much better at handling the class difference. For example, Naive Bayes outputs a likelihood for both classes individually, meaning each class is considered fully. In addition, Gradient Boosting and Adaboost both apply equally heavy weights to any observation that is misclassified, regardless of which class it originally came from. This means that there is no need to transform the data beforehand, as each of these methods handles the imbalance internally.

4.8 SHAP Values

As stated earlier, the goal of this research is two-pronged: first, provide a method for early identification of at-risk students that has at least 50 percent accuracy, and second, provide qualitative analysis of how the features impact the data set. In the past, high interpretability methods have been utilized to accomplish these goals, as they are capable of providing both predictions and information about features, while low interpretability methods typically lack the latter capability. However, the increase in model performance seen in low interpretability methods is also valuable. As such, Shapley Additive Expanations, or SHAP values, are utilized on the low interpretability techniques in an attempt to gain similar qualitative information about the features, as seen in high interpretability techniques. SHAP values provide an estimation of how much each feature contributes to low interpretability models by calculating the impact of features on each prediction, which in turn explains the impact of each feature in terms of magnitude and direction (Lundberg and Lee, 2017). Thus, SHAP values are utilized in this research to better explain low interpretability methods and bridge the gap in qualitative information learned between low and high interpretability models.

Chapter 5

RESULTS

Model performance is summarized in Table 5.1 with the top features for each model provided in Appendix C. The accuracy, which describes how well the models meet the overall goal, is described in Table 5.2. The Type I and Type II errors for each model are summarized in Table 5.3. As expected, the high interpretability methods tend to perform worse predictively while the low interpretability methods perform better, based on F1 score. In order of performance, from worst to best, are: Decision Trees, Naive Bayes, Logistic Regression with L1 and L2 penalty (tie), Random Forest, Adaboost, and Gradient Boosting. The results of this research indicate that ensemble methods, including Random Forest, Adaboot, and Gradient Boosting are promising for modeling student retention in this application.

Though accuracy is not a holistic method for evaluating the models, it provides information on how close the models are at achieving the original goal of successfully predicting more than 50 percent of students who are not likely to return. At the optimal F1 score, which helps minimize Type I and Type II error, the accuracy of all the models explored in this analysis is greater than the goal of 50 percent (5.2). This means that any of the machine learning models has the capability to meet the original requirements of this research. However, when further analyzing Type I and Type II errors, it is clear that some models perform better than others. In this example, Type I error refers to the percentage of returning students who are predicted as non–returning, and Type II error refers to the percentage of non–returning students who are predicted by the model as returning. Thus, it is ideal to have very low Type I and Type II errors. The models that perform the best here are Gradient Boosting

Table 5.1: Model Performance and Top Features

Technique	AUC	F1 Score	Optimal Threshold
Random Forest	0.725	0.171	0.54
Gradient Boosting	0.733	0.209	0.07
Adaboost	0.723	0.196	0.45
Naive Bayes	0.68	0.155	0.53
Logistic Regression – L1	0.719	0.158	0.5
Logistic Regression – L2	0.719	0.158	0.5
Decision Trees	0.711	0.15	0.53

Table 5.2: Accuracy

Technique	Accuracy
Random Forest	0.60
Gradient Boosting	0.52
Adaboost	0.54
Naive Bayes	0.58
Logistic Regression – L1	0.65
Logistic Regression – L2	0.65
Decision Trees	0.61

and Adaboost. Recall that the original data set is very unbalanced, indicating that there are more returning than non–returning students. This means that, although lower Type II errors are more relevant to the overall goal, a higher Type I error means a significantly larger number of students are over–predicted as non–returning. Since the number of non–returning students is so small, having a Type I error of 30 percent

Table 5.3: Type I and Type II Error

Technique	Type I Error	Type II Error
Random Forest	.28	.40
Gradient Boosting	.18	.48
Adaboost	.22	.46
Naive Bayes	.30	.42
Logistic Regression – L1	.34	.35
Logistic Regression – L2	.34	.35
Decision Trees	.32	.39

roughly translates to 30 percent of the students in the entire data set being predicted as non–returning, when they do actually return. This makes it much more difficult to target non–returning students that need help and clarifies why Gradient Boosting and Adaboost are better predictively, despite their lower accuracies.

Referring back to feature selection, which was performed through the Logistic Regression with the L1 penalty, the top model chose a very high value of C. High values of C indicate lower penalties on the features. Thus, no features were removed from the model. This finding validates the assumption to include all features for the remaining models, as the best performing model occurs with a lower L1 penalty. Interestingly, there was virtually no difference between the results of either Logistic Regression models. Both had the same F1 score and chose the same top features based on coefficient values.

The Decision Tree was by far the worst method applied, despite its potential uses in this application, with an F1 score of .15. Aside from its low F1 score, its results were variable – each time the Decision Tree was run, a different set of parameters

were output by the cross validation. The result was completely different trees upon each run. It is very likely that this model is overfitting, and for the time being, is not considered a good model for this application.

The Random Forest method was one of the first ML techniques applied in this research, but performs the worst of the low interpretability methods. As mentioned previously, both Gradient Boosting and Adaboost performed the best, with F1 scores of .209 and .196 respectively. Both Adaboost and Gradient Boosting performed feature selection, where Adaboost removed six features and Gradient Boosting only removed `Pell Eligible?`. This result is of interest as Adaboost chose to remove `Pell Eligible?`, `High`, and `No CI?`. It appears that the Adaboost chose to remove all strong correlations from the model, both between `CI Score` and `No CI?` and `Pell Eligible?` and `Very High/High` financial need. Gradient Boosting, however, only removed the correlation from the latter. This appears to indicate that removing these features has a positive impact on performance, despite other models ignoring these relationships. As mentioned earlier, both Adaboost and Gradient Boosting achieve over 50 percent accuracy while minimizing the Type I error. None of the other methods were able to achieve the same level of accuracy while minimizing the quantity of students falsely predicted as non–returning.

For all models, none of the F1 scores are too far apart, indicating that all of the techniques have some potential. In addition, for the best three models, `CI Score`, `ASU Scholarships?`, and `Meets MAT Requirements?` were always included in the top five features when ranked by feature importance. Moreover, `CI Score` was always in the top two features when ranked by feature importance for the ensemble methods, including Decision Trees, Random Forest, Adaboost, and Gradient Boosting. It was also included as the top predictor in both Logistic Regression models, followed by CI. Even for Naive Bayes, which did not include `CI Score` in the top five features, `No`

`CI?` was chosen as the top feature.

The only static characteristic in the top five features is `Sex?`, which appeared as the fourth most important feature in Adaboost. Both the Logistic Regression and Naive Bayes models found several financial need variables in their top five, though the other models did not find the same importance. Still, though all models did not acknowledge financial need level as a factor, at least one financial variable was present in the top five features for all of the models.

For the high interpretability models, the coefficient size and their odds ratios provides information on how the features impact student retention. For example, it is clear that, according to the Logistic Regression results, `CI Score` strongly impacts retention. To obtain the same information for the low interpretability methods, the SHAP values were calculated and visualized. Unfortunately, compatability with Adaboost is not yet available, so SHAP values were not calculated for this method. The SHAP visualization for Random Forest is located in Appendix D and the visualization for Gradient Boosting is located in Appendix E. For Gradient Boosting, it appears that students with low values of `ASU Scholarships?` are much more likely to not return, with different groups of student experiencing this more severely than other groups. The same can be said for `CI Score`. Similar is seen with the SHAP values for Random Forest, which show low values of `ASU Scholarships?` having a greater impact on non–returning likelihood. However, honors students and those with higher values of CI appear to have a greater likelihood of being retained by the model, which is not as evident in the Gradient Boosting model. Both models show a similar result for Meets MAT Requirements?, where a student who does not meet their major's requirements is more likely to not return, according to the model.

Chapter 6

CONCLUSION

6.1 Overall Summary

The results of this research provide meaningful information in regards to student retention. The top three models – Random Forest, Gradient Boosting, and Adaboost – all included `CI Score`, `ASU Scholarships?`, and `Meets MAT Requirements` in the top five features when ranked by feature importance. Moreover, CI appeared in the models in some way, either as `CI Score` or `No CI?`. This suggests that CI in particular may be an indicator of student success. According to the SHAP values for Gradient Boosting, which was the best performing model overall, lower values of `CI Score` and ASU Scholarships? have a stronger impact on a student's likelihood of not returning. Considering that `CI Score` is determined by SAT/ACT scores and high school GPA/class rank, this warrants a deeper analysis of how these features affect retention. Given that financial aid appears in all models, it also has some impact – though in what way, whether it be as scholarships or a student's financial need, is not clarified by the SHAP values. Since the top models include the same three features, it seems logical to conclude that a student's `CI Score`, their possession of scholarships, and their math placement might have some impact on their likelihood to continue for their first spring.

Unexpectedly, static characteristics do not appear as top features in any of the models except Adaboost. This may be due to their relatively small proportions – for example, the quantity of URM and female students is small compared to those who do not meet this criteria, so it is possible that the models struggle to include these

features in the final results.

For choosing a model to utilize for early identification of at–risk student, this research indicates that Gradient Boosting in combination with SHAP values is a strong contender. It had the highest F1 score, exceeded the original goal of correctly predicted more than 50 percent of non–returning students, and severely limited the Type I error, which is much better than previous analysis in JMP or any of the other methods. In addition, the inclusion of SHAP values in this analysis allow greater interpretation of the features, similar to a high interpretability method, like Logistic Regression. Other methods had higher accuracy at the expense of a higher Type I error, and though Adaboost performed similarly, there was no way to evaluate its results intuitively, as it does not yet have compatibility with SHAP values. It is important to note that this does not mean the results of the other methods are not useful – if interest in the future shifts more from identification to feature interpretation, Logistic Regression in particular is a well-suited tool.

Another important consideration is the actionability of the features. Knowing that `CI Score` impacts retention outcomes, what can be done to help students with low or no CI? The answer to this question is unclear, as more than advisor outreach may be required to lead to positive enrollment outcomes. This calls into question the decision to include these variables, though it is still useful to understand the impact of these features on a student's decision to persist. Features that the university can act more directly upon may be more prudent to future analysis.

These models are now available to provide early identification of students at–risk of not returning. The models, specifically Gradient Boosting, are run by the student retention team on incoming student data. The models output a list of students who are flagged as "at risk" by the university. This list of students is shared with the university who perform outreach to the flagged students and provide support

resources, in hopes of improving their likelihood of returning.

6.2 Future Work

Now that the framework is complete, there are many directions that this project can take moving forward. First, as with any system, there exists the need to analyze any gaps between this framework and other, similar implementations. Since the use of machine learning in student retention analysis is growing, lessons from other researchers can improve the implementation developed in this research.

To improve the analysis performed above, more relevant data is needed. This means adding data from the 2019 and 2020 cohorts and removing older cohorts, such as the 2015 and 2016 sets, from the analysis. In the past, five cohorts were chosen for analysis, which provided a sample size of approximately 10,000 students. This selection was made to maximize the amount of data available for examination, but no minimum or maximum number of cohorts is necessary so long as the sample size is appropriate. A potential concern when adding this new data is if there exists some type of significant difference between cohorts. Adding year as a feature or even performing clustering analysis will be useful in identifying if this concern is valid. Moreover, the impact of COVID–19 on student retention may warrant separate analysis of the Fall 2020 and 2021 cohorts. It is recommended to first perform analysis to determine if COVID–19 had a significant impact on retention before adding these cohorts to the data set.

Further developments are required to increase the predictive accuracy of these models. While the prediction accuracy is high, the Type I error for many of the models is comparatively large. Rectifying this may involve the inclusion of new features, such as Age, SAT/ACT, AP classes, LGBTQIA+ status, and more, most of which were not analyzed for simplification purposes. Moreover, it would be interesting

to call into question the exclusion of international students in this modeling. The original reasoning was the quantity of missing data these students have as compared to domestic students – oftentimes, international students lack CI scores and certain high school characteristics. Keeping these students, performing transformations on their missing data points, and identifying international status as a feature may be a better alternative when modeling.

One of the limitations of these models is that they only consider factors known prior to a student's college attendance in predicting their enrollment in spring. All of the machine learning techniques should be re–applied to create three new models. The first new model will predict enrollment in spring depending on factors known mid–way through the first fall semester, which requires the same label, but a different set of features. The second new model will utilize those features known at the end of a student's first fall to predict whether or not a student will re–enroll for the fall of their sophomore year. For this analysis, it is important to only include those students who were actually enrolled in spring in the data set – in other words, this data set will be slightly smaller than the previous set as the students who did not enroll for their first spring are excluded. The last model that needs development is one which uses factors known mid–way through the first spring to predict re–enrollment in the fall of sophomore year.

There are also several long–term goals for this data project. First, the data set should be fully transitioned to Python. In the past, all data pre–processing was performed in Excel, which was both time consuming and frustrating, given the size of the set. If the data set were fully integrated into Python, the pre–processing would require much less time. In addition, it would be a much better way of storing the data as all files would be in the same location as the modeling code. The second long term goal is to implement these models into a dashboard that is accessible to advisors. This

dashboard should be straightforward – it should provide the list of students who are currently at risk and some reasoning for why that student was flagged at risk (e.g., no scholarships). The dashboard is created for the layman, so that an individual with no background in machine learning can leverage the knowledge with no previous training.

REFERENCES

Ayers, L. D. G., "University experiences and women engineering student persistence", (2017).

Baker, C. N. and B. Robnett, "Race, social support and college student retention: A case study", Journal of college student development **53**, 2, 325–335 (2012).

Bean, J. P., "Dropouts and turnover: The synthesis and test of a causal model of student attrition", Research in higher education **12**, 2, 155–187 (1980).

Berger, J. B. and J. M. Braxton, "Revising tinto's interactionalist theory of student departure through theory elaboration: Examining the role of organizational attributes in the persistence process", Research in higher education **39**, 2, 103–119 (1998).

Bowman, N. A. and N. Denson, "A missing piece of the departure puzzle: Student–institution fit and intent to persist", Research in higher education **55**, 2, 123–142 (2014).

Bowman, N. A., A. Miller, S. Woosley, N. P. Maxwell and M. J. Kolze, "Understanding the link between noncognitive attributes and college retention", Research in higher education **60**, 2, 135–152 (2019).

Breiman, L., "Bagging predictors", Machine Learning **24**, 123–140 (1996).

Cabrera, A. F., *Pathways to a Four-Year Degree Determinants of Degree Completion among Socioeconomically Disadvantaged Students* (Distributed by ERIC Clearinghouse, S.l, 2003).

Chang, M. J., A. W. Astin and D. Kim, "Cross-racial interaction among undergraduates: Some consequences, causes, and patterns", Research in higher education **45**, 5, 529–553 (2004).

Chang, M. J., J. Sharkness, S. Hurtado and C. B. Newman, "What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups", Journal of Research in Science Teaching **51**, 5, 555–580 (2014).

Chen, R. and S. L. DesJardins, "Exploring the effects of financial aid on the gap in student dropout risks by income level", Research in higher education **49**, 1, 1–18 (2008).

Chen, Y., A. Johri and H. Rangwala, "Running out of stem: A comparative study across stem majors of college students at-risk of dropping out early", in "Proceedings of the 8th International Conference on Learning Analytics and Knowledge", LAK '18, p. 270–279 (Association for Computing Machinery, New York, NY, USA, 2018).

de Rooij, M. and W. Weeda, "Cross-validation: A method every psychologist should know", Advances in Methods and Practices in Psychological Science **3**, 2, 248–263 (2020).

Domingos, P. and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss", Mach. Learn. **29**, 2–3, 103–130 (1997).

Freund, Y., R. Schapire and N. Abe, "A short introduction to boosting", Journal-Japanese Society For Artificial Intelligence **14**, 771-780, 1612 (1999).

Friedman, J., "Greedy function approximation: A gradient boosting machine.", Annals of Statistics **29**, 1189–1232 (2001).

He, L., R. A. Levine, A. J. Bohonak, J. Fan and J. Stronach, "Predictive analytics machinery for stem student success studies", Applied artificial intelligence **32**, 4, 361–387 (2018).

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

Lundberg, S. and S.-I. Lee, "A unified approach to interpreting model predictions", (2017).

McKinney, L., L. McKinney, A. B. Burridge and A. B. Burridge, "Helping or hindering? the effects of loans on community college student persistence", Research in higher education **56**, 4, 299–324 (2015).

Muncie, T., "Using machine learning models to predict student retention: Building a state-wide early warning system", (2020).

Ng, A. Y., "Feature selection, l1 vs. l2 regularization, and rotational invariance", in "Proceedings of the Twenty-First International Conference on Machine Learning", ICML '04, p. 78 (Association for Computing Machinery, New York, NY, USA, 2004).

Schreiner, L. A. and D. D. Nelson, "The contribution of student satisfaction to persistence", Journal of college student retention : Research, theory practice **15**, 1, 73–111 (2013).

Seidman, A., "Where we go from here", College student retention: Formula for student success **295** (2005).

Sepulveda, T. A. C., "Development of a system architecture for the prediction of student success using machine learning techniques", (2020).

Spady, W. G., *Peer Integration and Academic Success: The Dropout Process Among Chicago Freshman*, Ph.D. thesis, University of Chicago, Department of Education (1967).

Tharwat, A., "Classification assessment methods: a detailed tutorial", (2018).

Tin Kam Ho, "Random decision forests", in "Proceedings of 3rd International Conference on Document Analysis and Recognition", vol. 1, pp. 278–282 vol.1 (1995).

Tinto, V., "Dropout from higher education: A theoretical synthesis of recent research", Review of educational research **45**, 1, 89–125 (1975).

Tinto, V., "Isaac newton and student college completion", Journal of College Student Retention: Research, Theory & Practice **15**, 1, 1–7 (2013).

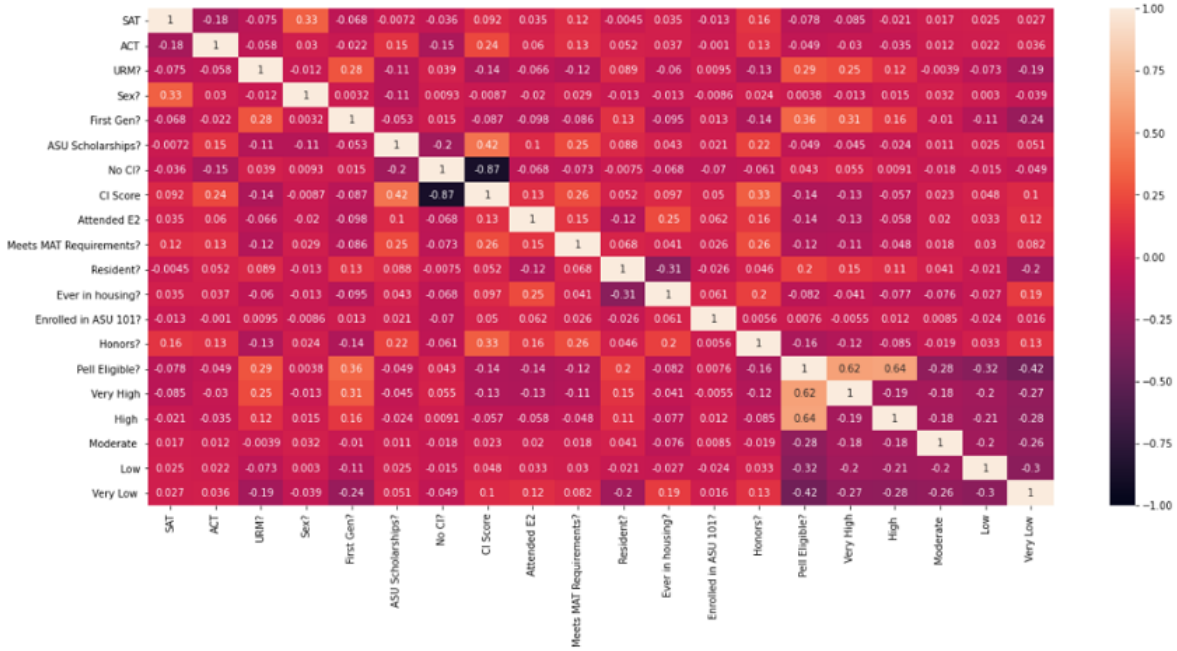
Tinto, V., "Through the eyes of students", Journal of college student retention : Research, theory practice **19**, 3, 254–269 (2017).

Wright, S. L., M. A. Jenkins-Guarnieri and J. L. Murdock, "Career development among first-year college students: College self-efficacy, student persistence, and academic success", Journal of career development **40**, 4, 292–310 (2013).

APPENDIX A

CORRELATION MATRIX

The complete correlation matrix:



APPENDIX B

FEATURE SET

The full set of features and their description:

Type	Name	Description
Nominal	Enrolled Spring?	A value of 1 (if the student did not enroll in spring) or 0 (if otherwise).
Nominal	URM?	A value of 1 (if the student is American Indian/Alaskan Native, Asian, Black or African American, Hispanic/Latino, Native Hawaiian/Pacific Islander, orTwo or More Races) or 0 (if otherwise).
Nominal	Gender?	A value of 1 (if the student identifies as Male) or 0 (if the student identifies as Female).
Nominal	First Gen?	A value of 1 (if the student is a first generation student, or the first student within immediate family to attend university for the first time) or 0 (if otherwise).
Nominal	ASU Scholarships?	A value of 1 (if the student has received any ASU scholarships) and 0 (if otherwise),
Continuous	CI Score	CI Score is a score domestic students get based on performance prior to college, such as high school GPA, test scores, etc. It is a range of 0 to 150.
Nominal	No CI	A value of 1 (if the student is missing a CI score) or 0 (if otherwise).
Nominal	Attended E2	A value of 1 (if the student attended ASU's E2 camp during the summer prior to the start of their first semester) or 0 (if otherwise).
Nominal	Meets MAT Requirements?	A value of 1 (if the student is on track with their MAT requirements) or 0 (if otherwise).
Nominal	Resident?	A value of 1 (if the student is an AZ resident) or 0 (if otherwise).
Nominal	Ever in Housing?	A value of 1 (if the student was ever in ASU housing) or 0 (if otherwise).
Nominal	Enrolled in ASU 101?	A value of 1 (if the student was enrolled in the mandatory "ASU 101" class) or 0 (if otherwise).
Nominal	Honors?	A vaue of 1 (if the student is a Barrett Honors student) or 0 (if otherwise).
Nominal	Pell Eligible?	A value of 1 (if the student is Pell Eligible due to high financial need) or 0 (if otherwise).
Nominal	Very High	A value of 1 (if the student has very high financial need) or 0 (if otherwise).
Nominal	High	A value of 1 (if the student has high financial need) or 0 (if otherwise).
Nominal	Moderate	A value of 1 (if the student has moderate financial need) or 0 (if otherwise).
Nominal	Low	A value of 1 (if the student has low financial need) or 0 (if otherwise).
Nominal	Very Low	A value of 1 (if the student has very low financial need) or 0 (if otherwise).

APPENDIX C

TOP FIVE FEATURES

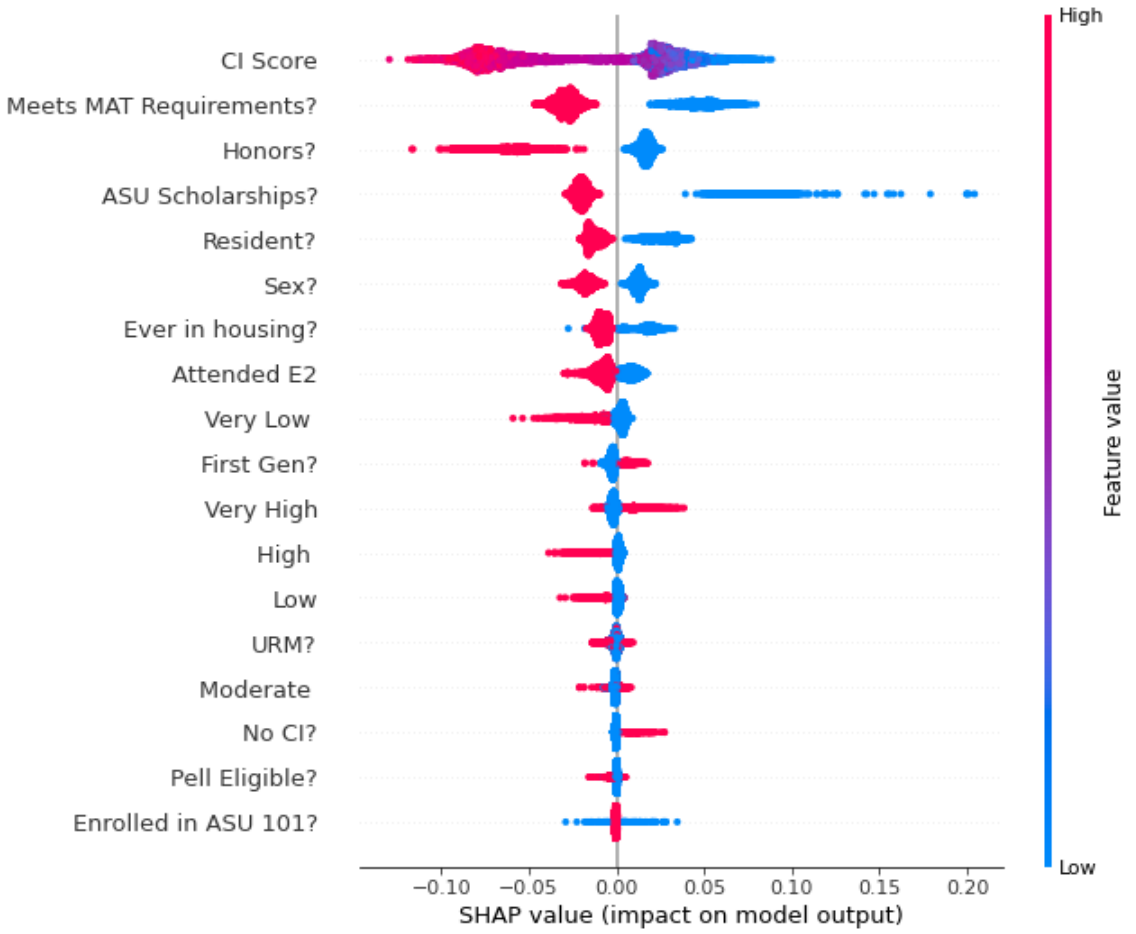
The top five features for each model, based on feature importance or coefficient size:

Technique	Top 5 Features (In Ranked Order)
Random Forest	Feature Importance: CI Score: .362 Meets MAT Requirements: .176 ASU Scholarships: .130 Honors: .053 Resident: .052
Gradient Boosting	Feature Importance: CI Score: .224 ASU Scholarships: .210 Honors: .120 Meets MAT Requirements: .105 First Gen: .065
Adaboost	Feature Importance: Resident: .172 CI: .148 Ever in housing: .148 Sex: .117 Tie - ASU Scholarship/Meets MAT Requirements: .086
Naïve Bayes	Coefficients: No CI: 5.52 Very High: 3.99 Moderate: 3.99 High: 3.91 Low: 3.74
Logistic Regression - L1	Coefficients: CI Score: -2.66 No CI: -1.69 Very High: 1.67 Pell: -1.44 High: 1.41
Logistic Regression - L2	Coefficients: CI Score: -2.66 No CI: -1.69 Very High: 1.67 Pell: -1.44 High: 1.41
Decision Trees	Feature Importance: CI: .634 Meets MAT: .135 ASU Scholarships: .126 Resident: .065 Attended E2: .028

APPENDIX D

RANDOM FOREST SHAP VISUALIZATION

SHAP visualization for Random Forest:



APPENDIX E

GRADIENT BOOSTING SHAP VISUALIZATION

SHAP visualization for Gradient Boosting: