ANALYSIS OF MTSU STUDENT RETENTION DATA

by

Danielle Baghernejad

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Mathematical Sciences

Middle Tennessee State University
May 2016

Thesis Committee:

Dr. Qiang Wu, Chair

Dr. Rebecca Calahan

Dr. Lisa Green

Dr. Cen Li

ACKNOWLEDGMENTS

I would like to thank Dr. Qiang Wu for his support and guidance throughout my research, as well Drs. Cen Li, John Wallin, and Michael Hains, without whose research this thesis would not have been possible. I am also very grateful to my fellow graduate students, whose discussions and comments helped further my analysis.

ABSTRACT

Student retention is a challenging task in higher education, since in general more students remaining in the university means better academic programs and higher revenue. Thus, improving retention rates can not only help current students achieve academic success, but help future students as well. The objective of this thesis is to employ data mining and predictive tools on student data to predict student retention among the freshman students. In particular, we aim to identify freshman students who are more likely to drop out so that preemptive actions can be taken by the university. Through data analysis, relevant variables are identified to incorporate into models for prediction. Missing values are taken into consideration, and missing value imputation methods are explored.

This thesis begins by introducing the theory behind missing value imputation and prediction methods before applying them on the student data set. For imputation, Mean Substitution and Multiple Imputation are considered, while predictive models consist of Logistic Regression and Random Forests. The final model results in the identification of several key variables, as well as areas for further study.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE PROBLEM: MTSU STUDENT RETENTION

Student retention, measured by whether a student completes a college degree in six years or less, is a useful measure of a university's overall success and student satisfaction. As of 2012, MTSU has a student retention rate of 53 percent, which is fourth-best among Tennessee's four-year public universities [3], but with only half of the students completing the degree, much improvement can be made. In order to address student retention, we must understand why students withdraw from school to be able to make any improvement. The goal of this thesis is to generate a model to predict student retention as well as identify factors contributing to students dropping out. Once an understanding of the situation is developed, the university can better assist them in completing their degree.

Student retention is a challenging task in higher education and it is reported that about one fourth of students dropped college after their first year [18]. In general, more students remaining in the university means better academic programs and higher revenue. Thus, improving retention rates can not only help current students achieve academic success, but also improve the chances of academic success for future students, improving the university overall. Recent study results show that intervention programs can have significant effects on retention, especially for the first year [12]. Thus for this thesis, focusing on the freshman class may yield the most immediate progress towards a long term goal.

While ideally we want to improve retention for all students, minority groups have often faced more difficulties while completing their degree. We will consider possible issues within minority groups as well as the general student body to see if there are different factors for specific demographics of students.

This thesis is a continuation of work done by Drs. Cen Li, John Wallin, Qiang Wu and Michael Hains at MTSU [20].

1.1 Current Research

Being an area of concern for many universities, retention research has already been pursued with some measurable success. For example, through data mining and analysis, administrators at South Texas College discovered that students who register late for a course are more likely to withdraw. Realizing that this has a negative impact, South Texas College decided to eliminate late registration [9]. Students at the University of Alabama built a predictive model of student retention and found that commuter students are more likely to drop out. Consequently, the university developed student retention strategies including requiring all freshmen to live on-campus [9]. These and other successful cases have shown that an analysis of retention can have a beneficial effect on the university and student success.

The current literature generally describes two different approaches. Some studies have focused on prediction and emphasize comparing and testing different classification models. Popular models are seen coming from both statistical and computer science approaches, with logistic regression, support vector machines, decision trees, and neural networks commonly being implemented. Other studies instead focus on identifying crucial student retention factors to develop an understanding of the underlying relationships, identifying key variables to address. For our study, an exploration of both topics can be insightful into implementing an intervention program.

While the results from different studies may vary, common variables have be found that can already be used to be model retention. Researchers have consistently found high school GPA, admissions test scores, gender, and ethnicity to be significant predictors, which should be included in any predictive model [4]. We can then expect that with MTSU students these variables will also be important in the analysis.

1.2 MTSU Data Set

The data set used was provided by the Office of Institutional Effectiveness, Planning and Research (IEPR) at MTSU. The comprehensive student data includes information from all the students enrolled at MTSU between 2007 and 2013. This data set includes over 100 variables incorporating information about academic, social, and financial aspects of the students from before they enter MTSU to the point when they graduate or leave MTSU. Variables are compiled from different sources, including FAFSA, ACT, and other university data.

This data was further broken down into target groups, yielding the following sizes for each group in our study.

Table 1: Sample Sizes By Target Group

Student Group	Total Students
African American	3122
Disabled	391
First Generation	792
Hispanic	645
All	41238

Table 2 depicts the current percentages who stayed, transferred, or dropped out for different student groups. The target groups had very different numbers compared to the entire student body, which could speak to either the size of the groups or different factors affecting for the groups. In every student group, a larger percentage of the students who did not stay at MTSU dropped out of school completely, with the highest percentage of students transferring seen in the Disabled Student group.

Table 2: Freshman Status By Student Group

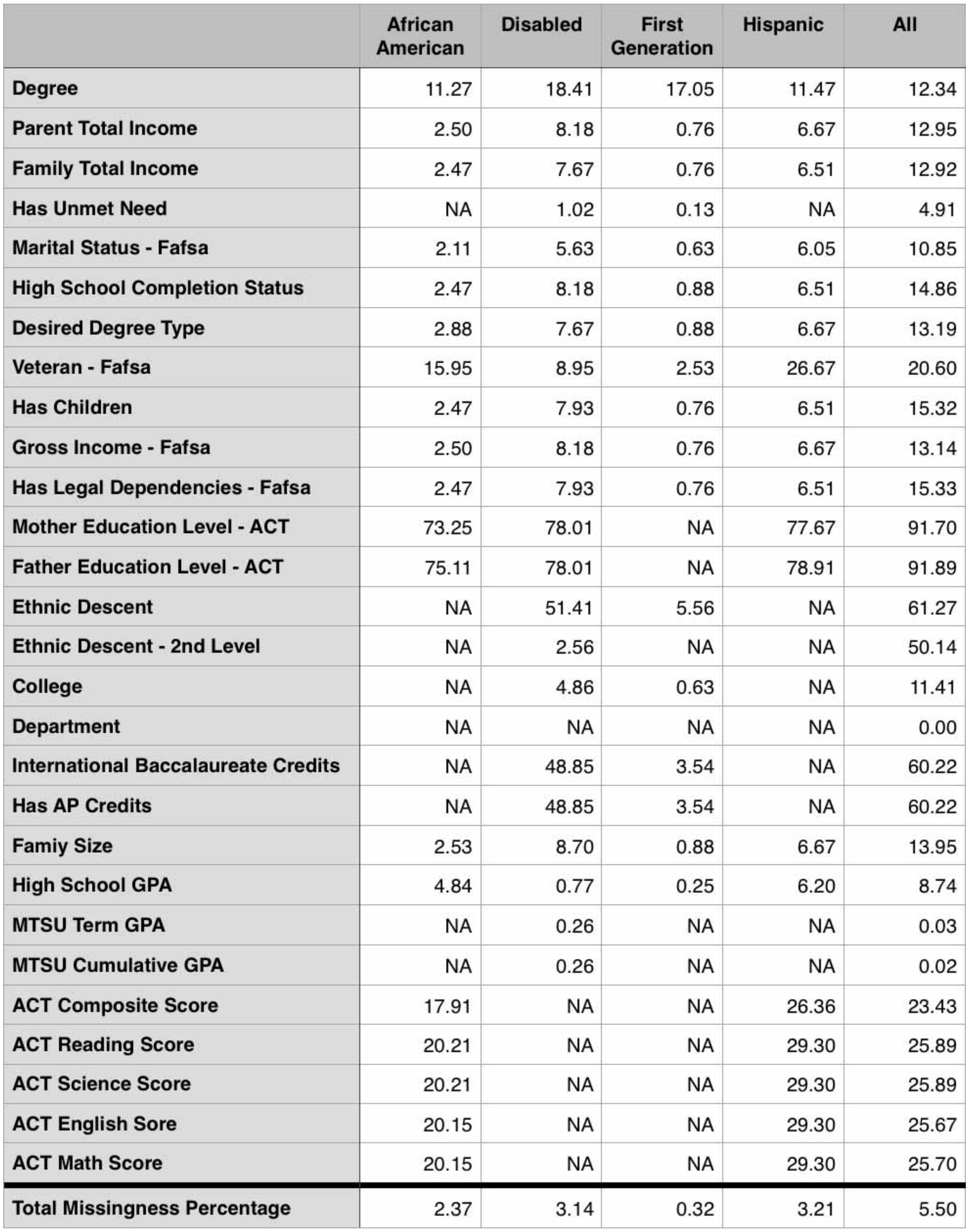
Student Group	Stayed	Transferred	Dropped
African American	0.60	0.11	0.29
Disabled	0.48	0.19	0.33
First Generation	0.70	0.08	0.22
Hispanic	0.67	0.07	0.26
All	0.57	0.11	0.32

1.3 Missing Data

While the data set itself is large, there are some entries with missing values. Accurate analysis of the data requires handling the missingness in some way. We will explore the implications of missingness in the course of our analysis to see how, if at all, missingness affects the method used for analysis. For example, if we were more likely to be missing data for students with low GPA's, our data might be biased, which would in turn bias the final model.

Most variables in the dataset are completely observed, but there are 24 variables with missing values. Table 3 below presents a list of the percentage of missingness by variable. The overall missingness in the dataset is approximately 6%, so in general the percentage of missingness is very small. However, missingness for the specific variables can be very high. Some variables, such as Term GPA with 0.03%, have almost all data observed, but others, such as Father Education Level with 91%, are very sparse. If we examine it even further at the demographic level, we see even more variation within target groups. Since there is no general rule of thumb as to what point missingness renders the data useless, it is difficult to determine at exactly what point we must throw away some data, if at all. The goal of handling missing data is to retain as much of the original data as possible in order not to lose any information within the data at hand, thus we must consider missingness when handling the data.

Table 3: Percent Missing Per Variable By Student Type



	African American	Disabled	First Generation	Hispanic	All
Degree	11.27	18.41	17.05	11.47	12.34
Parent Total Income	2.50	8.18	0.76	6.67	12.95
Family Total Income	2.47	7.67	0.76	6.51	12.92
Has Unmet Need	NA	1.02	0.13	NA	4.91
Marital Status - Fafsa	2.11	5.63	0.63	6.05	10.85
High School Completion Status	2.47	8.18	0.88	6.51	14.86
Desired Degree Type	2.88	7.67	0.88	6.67	13.19
Veteran - Fafsa	15.95	8.95	2.53	26.67	20.60
Has Children	2.47	7.93	0.76	6.51	15.32
Gross Income - Fafsa	2.50	8.18	0.76	6.67	13.14
Has Legal Dependencies - Fafsa	2.47	7.93	0.76	6.51	15.33
Mother Education Level - ACT	73.25	78.01	NA	77.67	91.70
Father Education Level - ACT	75.11	78.01	NA	78.91	91.89
Ethnic Descent	NA	51.41	5.56	NA	61.27
Ethnic Descent - 2nd Level	NA	2.56	NA	NA	50.14
College	NA	4.86	0.63	NA	11.41
Department	NA	NA	NA	NA	0.00
International Baccalaureate Credits	NA	48.85	3.54	NA	60.22
Has AP Credits	NA	48.85	3.54	NA	60.22
Famiy Size	2.53	8.70	0.88	6.67	13.95
High School GPA	4.84	0.77	0.25	6.20	8.74
MTSU Term GPA	NA	0.26	NA	NA	0.03
MTSU Cumulative GPA	NA	0.26	NA	NA	0.02
ACT Composite Score	17.91	NA	NA	26.36	23.43
ACT Reading Score	20.21	NA	NA	29.30	25.89
ACT Science Score	20.21	NA	NA	29.30	25.89
ACT English Sore	20.15	NA	NA	29.30	25.67
ACT Math Score	20.15	NA	NA	29.30	25.70
Total Missingness Percentage	2.37	3.14	0.32	3.21	5.50

1.4 Prediction

Student Retention can be viewed as a classification problem, with a set of discrete outcomes. In the classification setting, the goal is to label a new occurrence as a class, though the determining of which class is model dependent.

In our data set, we have three possible cases: the student stayed, the student transferred, or the student dropped. After initial exploration, predicting student transfer was found to be a difficult task; there are a variety of variables that come into play when a student transfers that may not be present in the data set. For example, a student might need to transfer closer to home for family reasons, or they might want to transfer to a school with a different program not available at MTSU. Including those students in the model led to very poor performance. If we take into consideration that a transfer student is still pursuing their degree, even though the student might not benefit MTSU directly, they are not in as much need of assistance to continue their education. For our purposes, we will remove those students from the data for now in order to better identify the students withdrawing from school completely.

With that in mind, we are faced with a binary classification problem. The student can be labeled either yes or no for staying in school, so the model selection needs to be adequately suited for this type of prediction. We will explore models with different assumptions to see how they perform.

1.5 Variable Selection

An issue that we face with our data set is the sheer size of the data. While more data is generally good data, including data that is not related with the outcome creates noise in the model, and it makes it more difficult to understand the interplay of different variables when the variable set is large. In order to have an accurate and interpretable model, we must perform some type of variable selection to focus on the most important variables. While the simplest approach is just selecting variables with a high correlation with our outcome, this only takes into consideration one level of correlation, ignoring possible interaction effects between sets of variables.

Some initial work in this regard has already been done. We will explore the current variable selections as well as possibilities for other variable selection processes.

1.6 Structure of This Thesis

The remaining chapters will discuss the various details of our analysis. Chapter 2 will provide an overview of missing value methods, then discuss the implementation on the MTSU data. Chapter 3 outlines several popular prediction methods for a retention model, followed by a discussion for the implementation on the MTSU data. Chapter 4 discusses the results of the final models, exploring the implications of missingness and variable selection and its effects on accuracy and interpretation.

CHAPTER 2

MISSING VALUE METHODS

The handling of missing data can be performed in a variety of ways with different underlying assumptions required to implement. In this chapter we will first properly define the missingness mechanism, identify the missingness mechanism in our data, and then explore common methods for handling missingness.

2.1 Mechanism of Missingness

To understand the behavior of missingness, a few basic notions about missing values must be defined.

2.1.1 Definitions

Let $\mathbf{D}$ denote an incomplete dataset with $\mathbf{r}$ variables $\mathbf{D} = \{\mathbf{A}_1, \mathbf{A}_2, \dots \mathbf{A}_r\}$ and $\mathbf{n}$ instances. For each variable $\mathbf{A}_j$, $j = 1, 2, \dots r$, $\mathbf{A}_j$ contains two parts, $\mathbf{A} = \{\mathbf{A}_{obs}, \mathbf{A}_{mis}\}$, where $\mathbf{A}_{obs}$ is the set of observed elements and $\mathbf{A}_{mis}$ is the set of missing elements. Similarly, the entire dataset $\mathbf{D}$ also consists of two components, $\mathbf{D} = \{\mathbf{D}_{obs}, \mathbf{D}_{mis}\}$, where $\mathbf{D}_{obs}$ is the set of observed values and $\mathbf{D}_{mis}$ is the set of missing values.

Let $\mathbf{R}$ be a response indicator matrix with the same dimensions as $\mathbf{D}$ to describe the missingness. Each element of $\mathbf{R}$ is defined as $\mathbf{r}_{ij} = 1$ if the value is missing, else $\mathbf{r}_{ij} = 0$, where $\mathbf{r}_{ij}$ corresponds to the ith instance at variable $\mathbf{A}_j$, $i = 1, 2, \dots \mathbf{n}$, $j = 1, 2, \dots \mathbf{r}$.

The aim of imputation is to fill in all the blanks of incomplete dataset $\mathbf{D}$, where $\mathbf{r}_{ij} = 1$, so that the estimated complete dataset $\mathbf{D}$ can be used for succeeding statistical analysis.

2.1.2 Missing Mechanism

Since most popular methods of imputation depend on several assumptions to hold, in order to apply any procedure to a missing data set, the underlying missingness mechanism must first be identified. The missingness mechanism determines how the missing data are generated and it is a potential factor that will affect the imputation results. Thus, a comprehensive study of the noise impact on imputation methods must take different missingness mechanisms into account. There are three types of missing data mechanisms according to Little and Rubin [13]:

Definition 2.1 *Missing Completely At Random (MCAR): If* $Pr(\mathbf{R}|\mathbf{D}_{mis}, \mathbf{D}_{obs}) = Pr(\mathbf{R})$*, then the missing mechanism is defined as MCAR, where* Pr *represents the probability.*

MCAR implies that the missingness is unrelated to both the missing and observed values in the dataset. For example, consider a survey that is being performed on a group of students and participants are asked to fill out a questionnaire. If the data is MCAR, a participant flips a coin to decide whether to complete each survey entry.

Definition 2.2 *Missing at Random (MAR): If* $Pr(\mathbf{R}|\mathbf{D}_{mis}, \mathbf{D}_{obs}) = Pr(\mathbf{R}|\mathbf{D}_{obs})$*, then the missingness mechanism is called MAR.*

MAR means the missingness depends on observed values but not on missing values. In our survey example, the data would be MAR if male participants are more likely to refuse to fill out information regarding their GPA. Whether or not they provide their GPA does not depend on the GPA itself, only their gender.

The issue with MAR is not whether gender itself can predict their GPA, but whether gender is a mechanism to explain whether or not a student will report their GPA. This mechanism can be used to identify whether a pattern of missingness exists within the data [14]. The MAR assumption is valid if only if it can be assumed that the pattern of missing variables is conditionally random, given the observed mechanism variables.

Definition 2.3 *Missing Not At Random (MNAR): If $Pr(\mathbf{R}|\mathbf{D}_{mis}, \mathbf{D}_{obs})$ is not equal to $Pr(\mathbf{R}|\mathbf{D}_{obs})$ and it depends on $\mathbf{D}_{mis}$, then the missing data is MNAR.*

MNAR implies that there is a pattern within the missing data to explain why it is missing. For example, the survey data would be MNAR if participants with low GPAs were more likely to leave out their GPA.

Since the mechanism for missingness is not random at all, MNAR data have to be handled differently if we want to use the data. Most popular approaches operate under the assumption of the data being at least MAR, and some approaches have been suggested to better meet this assumption. Indeed, in order to push the data towards MAR, it has been suggested to include auxiliary variables that are correlates of missingness and/or correlates of the variable of interest. Including the former can help to reduce bias and move the situation closer to MAR; including the latter may help to reduce variance [5].

2.2 Identifying Missing Mechanism

If the data are MAR, then we can ignore the missing data without worrying about biasing the overall data set. If, on the other hand, the data is MNAR, then the observed data is a biased sample since the missing data contains information about the response, so the missing data cannot be ignored. Unfortunately, there is not a set way to determine MAR and MNAR, and it is even more difficult to distinguish between MAR and MCAR without additional information [7]. However, most imputation methods only require MAR to hold, which is a more reasonable assumption to meet.

Most analysts operate under the assumption of MAR, and if the model results seem questionable, a reevaluation of the MAR assumption is required. If the analyst is unsure whether or not the data are MAR, it is common advice to collect information about any characteristics that might even remotely affect missingness and include those variables in the imputation model [16]. These variables can always be removed for the prediction model, but can help push the data towards MAR in the imputation process.

2.3 Missing Value Methods

There are several standard methods used to handle missing data, each with its own benefits and problems. We will first discuss classical methods, then explore more rigorous modern approaches.

2.3.1 Traditional Methods

1. Listwise (Case) Deletion

   The most trivial way to deal with missing data is simply to throw it out. Listwise Deletion entails removing instances in the data set that have a missing value within i [1]. Listwise Deletion is the most common solution to missing values, so common that it is often the default method in statistical software packages. While common, it is the most dangerous, because for it to be valid, it requires the strict assumption of MCAR to ensure the dataset is not biased, but MCAR does not always hold. However, it is generally preferred because it is seen as conservative; it does not have to create data, but generally loses $20\% - 50\%$ of the data as a trade off. Indeed, if we were to take this approach with the MTSU data set, approximately 60% of the instances would be lost.

   If MCAR is met, Listwise Deletion results in a smaller sample size, thereby inflating the standard errors and reducing the level of significance. As a result, employing listwise deletion increases the risk of a Type II error. With a larger sample this risk is reduced, but still important. The sample may still be representative, but the cost is paid in the loss of statistical power [1].

   On the other hand, if MCAR is not met, Listwise Deletion can yield biased estimates, since the removed data may leave out a part of the population of interest, thus not giving a representative sample. The estimates may not be correct, and in some instances may reverse the direction of effects, asserting negative relations that are actually positive and vice versa. The cost for employing it is paid with biased estimates [1].

In either case, the use of Listwise Deletion does not seem powerful enough to be deemed the default method, which has spurred the development of the methods discussed below.

2. Mean Substitution

   Mean Substitution is another popular method to employ, replacing missing values with the mean for that variable, or the mode if the variable is discrete [8]. Mean Substitution only requires that the data be MAR, which is a much more reasonable assumption than MCAR. This method is based on the fact that the mean is a reasonable guess of a value for a randomly selected observation from a normal distribution. If MAR does not hold, however, the mean may be a poor guess. For example, billionaires might be less likely to supply their salary, and a substitution of a mean of approximately $\$60,000$ would be a very poor guess for their true salary.

   Mean Substitution is especially problematic when there are many missing values. If 30% of the data is missing and the mean is supplied for all of them, 30% of the data then has zero variance, thus greatly attenuating the variance of the variable and thus underestimating the correlation of income with any other variable. As a result, Mean Substitution potentially distorts relationships between variables by pulling estimates of the correlation toward zero [8].

3. Regression Substitution

   Regression Substitution extends the concept of Mean Substitution, but instead of simply using the variable's mean value, a regression formula is created, using other variables in the data set as predictors to estimate a value for the missing values in the current variabl [11]. When only one variable has missing values, this is easy enough to implement, but when more than one variable has missing values, dealing with missing variables in the predictors has to be addressed. If we simply ignore the missing instances for creating the regression model, we may find the same complications seen in Listwise Deletion. If on the other hand we want to include those cases, we must create some estimate for the missing values in the predictor variables, which leads us to the original issue. Using Mean Substitution may be a practical choice, but again, the variances will be affected.

Once missing values are dealt with, there is also the issue of determining what type of regression to perform, and whether or not to include a subset of predictors in each model [11]. In a data set with a small number of variables this might be manageable, but as the size of the variables grow, so does the complexity in model building.

With all of these complications in mind, Regression Substitution does give us the benefit of a confidence interval around each missing value estimate. This can be useful in analysis, but depending on the algorithm used for prediction, inclusion of the confidence interval in the prediction model may not be possible, or may not be easily implemented. Because of these complications, Regression Substitution may not provide improved estimates over Mean Substitution, and with the added difficulty in implementation for the model building stage of analysis, Mean Substitution is still commonly preferred.

2.3.2 Modern Methods

While the methods described thus far are often used, they are far from optimal for handling missing data except under specialized circumstances. In trying to address the shortcomings addressed above, several modern methods have been developed to create more rigorous estimates. These methods have their own drawbacks, especially when the data set is large. The Curse of Dimensionality is especially noticed in methods involving integrals, which quickly become intractable as the dimensionality of the problem grows. Modifications to the algorithms have been employed with this issue in mind, but still require significantly greater computational expense.

1. Expectation Maximization (EM)

   EM is a maximum likelihood approach that replaces all missing values with maximum likelihood derived values based on the assumed distribution of the data. The approach is based on the observed relationships among all the variables and injects a degree of random error to reflect uncertainty of imputation [1]. The EM algorithm consists of iterating through cycles until convergence to a predicted value of the distribution parameters θ is met. Cycles consist of performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter estimates found on the t^{th} iteration are then used to determine the distribution of the latent variables in the $(t+1)^{th}$ iteration. Values are imputed iteratively until successive iterations are sufficiently similar.

   To express the process mathematically, the complete-data log-likelihood is $l(\theta|y)$. The expected value of the function is

   $$Q(\theta|\theta^{(t)}) = \int l(\theta|y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) dY_{mis}$$

   The goal of M step is then to find θ^{t+1}

   $$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})$$

for all θ.

Each successive iteration has more information because it utilizes the information from the preceding iteration. This iterative process is continued until the covariance matrix for the next iteration is virtually the same as that for the preceding iteration. The process generally converges quickly, but if there are many missing values and many variables, it can involve a great deal of computer time, which can be a significant weakness over traditional approaches.

The end result of the EM algorithm is one fully completed data set. Although it is an improved estimate over traditional methods, it has one inherent flaw. All the methods thus far have resulted in only one completed data set. Because this approach to imputation omits possible differences that could occur in independent runs of the algorithm, this single imputation will tend to underestimate the standard errors and thus overestimate the level of precision, giving more apparent statistical power than the data justifies. Also, we must note that the algorithm requires some initial values to begin, and may suffer from poor starting values.

2. Multiple Imputation (MI)

Multiple imputation is a way to overcome some of the inherent flaws of EM. MI allows pooling of m different parameter estimates derived from independently generated complete data sets to find an improved parameter estimate. MI imputations generally produce somewhat different solutions from each data set. If these m solutions are similar, it gives credence to the validity of the imputation. However, if they are markedly different, then it is important to incorporate this uncertainty into the standard errors [15].

MI involves a 3-step process. First, m data sets are created using data augmentation techniques. While there are different augmentation techniques that can be applied, in general we aim to draw random samples from a function of the missing values, conditional on the other variables in the data set. Schafer suggests a Markov Chain Monte Carlo (MCMC) technique to sample from the target posterior predictive distribution, which is commonly employed [15].

After the m sets have been created, a predictive model is then employed on each data set, getting m different parameters $\hat{\beta}_i$ for the desired parameter β_i. The final step of MI is then to compute pooled estimates of the parameters $\hat{\beta}_i$ and standard errors $s_1, \ldots, s_m$ [14]. Thus the point estimate for the parameters is:

$$\hat{\beta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\beta}_i$$

A final variance estimate $\mathbf{V}_\beta$ reflects variation within and between imputations:

$$\mathbf{V}_\beta = W + (1 + \frac{1}{m})B$$

where

$$W = \frac{1}{m} \sum_{i=1}^{m} s_i^2$$
$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\beta} i - \hat{\beta})^2$$

One drawback with MI is that there is not a set standard for the number of imputations to use. The general recommendation is between 3 and 10 [19], which is based on Rubin's definition of the key criterion, the fraction of missing information γ. In a univariate sample with values missing at random, γ is approximately the fraction of cases with missing values. In a multivariate sample, γ is more complicated because different variables and cases contribute different amounts of information about different parameters.

Together with the number of imputations m, the fraction of missing information γ governs the relative efficiency of an estimate, which is approximately $(1+\frac{\gamma}{m})^{-\frac{1}{2}}$ in standard error units. In other words, standard errors based on m imputations are $(1+\frac{\gamma}{m})^{\frac{1}{2}}-1$ larger than they would be with infinite imputations; the excess is close to $\frac{\gamma}{2m}$, especially if $\frac{\gamma}{m}$ is small. For example, with 40% missing information and 10 imputations, standard errors are about 2% larger than their minimum possible value. With more imputations this can approach zero, but eventually the costs in storage and computing time outweigh the marginal gains in efficiency [19].

For implementing MI, the R package MI was used. Since the MCMC approach requires a posterior distribution, the package selects appropriate posterior distributions based on the variable types being imputed. While the package has built-in support to make a guess as to the variable types, it is possible to set the variable types manually to ensure proper implementation. Table 4 lists the default models corresponding to variable types.

Table 4: MI Functions By Variable Type

Variable Types	MI Function
Binary	mi.binary
Continuous	mi.continuous
Count	mi.count
Fixed	mi.fixed
Log-continuous	mi.continuous
Nonnegative	mi.continuous
Ordered-Categorical	mi.polr
Unordered-Categorical	mi.categorical
Positive-Continuous	mi.continuous
Proportion	mi.continuous
Predictive-Mean-Matching	mi.pmm

Mi.fixed imputes values by using the observed observation since they are all the same value. Mi.categorical uses a multinomial log-linear model to impute unordered categorical variables. Mi.continuous, mi.binary, mi.count, and mi.polr fit Bayesian version of the generalized linear models (bayesglm() and bayespolr() in the arm package). The Bayesian version of the generalized linear model is different from the classical generalized linear model in that it adds a Student-t prior on the regression coefficients [17]. Once the conditional distributions have been assigned, the algorithm proceeds by starting off with reasonable starting values for the missing values, then iterates through every variable with a missing value in the data set. For each variable, it samples from the target distribution outlined in Table 4, making random draws from the distribution to fill in the missing variables. Once each variable has had values imputed, the algorithm begins again at the first variable, sampling from the target distribution to fill in the missing values again. This process continues until the change from one iteration to the next is sufficiently small [16]. This process is implemented for m different data sets to create complete data sets. A predictive model is then implemented for the variable of interest on each data set, giving m parameter estimates, which are then pooled together for one final parameter estimate.

2.4 MSTU Data Set

With a variety of methods available, one aim of this thesis is to compare the accuracy and efficiency between traditional and modern methods, especially when performing on a data set of this size. Considering the computational demands of the latter, there may be few gains to be made when the data size is sufficiently large. For this analysis, Mean substitution was employed due to its ease and relatively similar accuracy over Regression Substitution. For a modern method, MI was also implemented, given its popularity and ease of use. Implementation of each method will be discussed below.

2.4.1 Data Cleaning and Preprocessing

A difficult step in any analysis is the data preparation phase, not because of the complexity of the task, but because it can take a bit of manipulation to get the data into a usable format for a model. Much work in this regard was already completed in the initial research, resulting in data tables for each target group and documentation explaining variable values. For this thesis, a little more manipulation had to be performed, including encoding missing values as NA and normalizing continuous variables.

A few variables used for bookkeeping were dropped, such as the level code variable, which indicated undergraduate status; since we are studying undergraduate students only, the variable was by definition constant, thus held no statistical insight. Variables with missingness too high to allow for the completion of imputation were also removed, since there is little chance of them having statistical use on their own, which we discuss further later. Transfer students were removed from the data, due to the difficulty in prediction mentioned previously. Finally, financial variables were transformed to be in a similar scale be transformed.

Potential outliers needed to be addressed in some way as well, since outliers have the potential to affect the final model. While outlier detection in general is a very nontrivial task, a simple approach consists of examining normalized values of variables and labeling the instances with values farther than 2 or 3 standard deviations as outliers. In some situations, one may simply remove the outliers from the data set, but in our data, if we were to remove instances with an outlier on any of the variables, we would end up removing approximately 30% of the data. Instead, we will replace the outlier value with the maximum or minimum value of the remaining instances to retain as much of the data as possible.

After this stage, we were left with 124 variables and 41,238 instances.

2.4.2 Visualization

Before imputation, it is helpful to inspect the data in some way to see the underlying behavior of the missing data. Identifying the missing patterns in the data set is helpful in determining if there is a pattern to the missingness as well.

The R package MI has visualization functions with this in mind [17]. Figure 1 is a graphical representation of the missing data, transforming the data into a matrix and assigning black to indicate when a value is missing. Variables which are very sparse can be seen by examining the columns, for example, the ACT variables consist of the very black columns on the very right. The data has been sorted by the missing pattern of each observation, with observations of the same missing patterns grouped together. There are 81 total missing data patterns present, though some patterns only have a handful of observations. When looking at the graph, we want to see if there are any distinct horizontal bands created by grouping by missingness pattern. If there are, it leads credence to the possibility of MNAR. While there is a slight difference in the horizontal strip in the middle, overall there is not a strong visual difference.

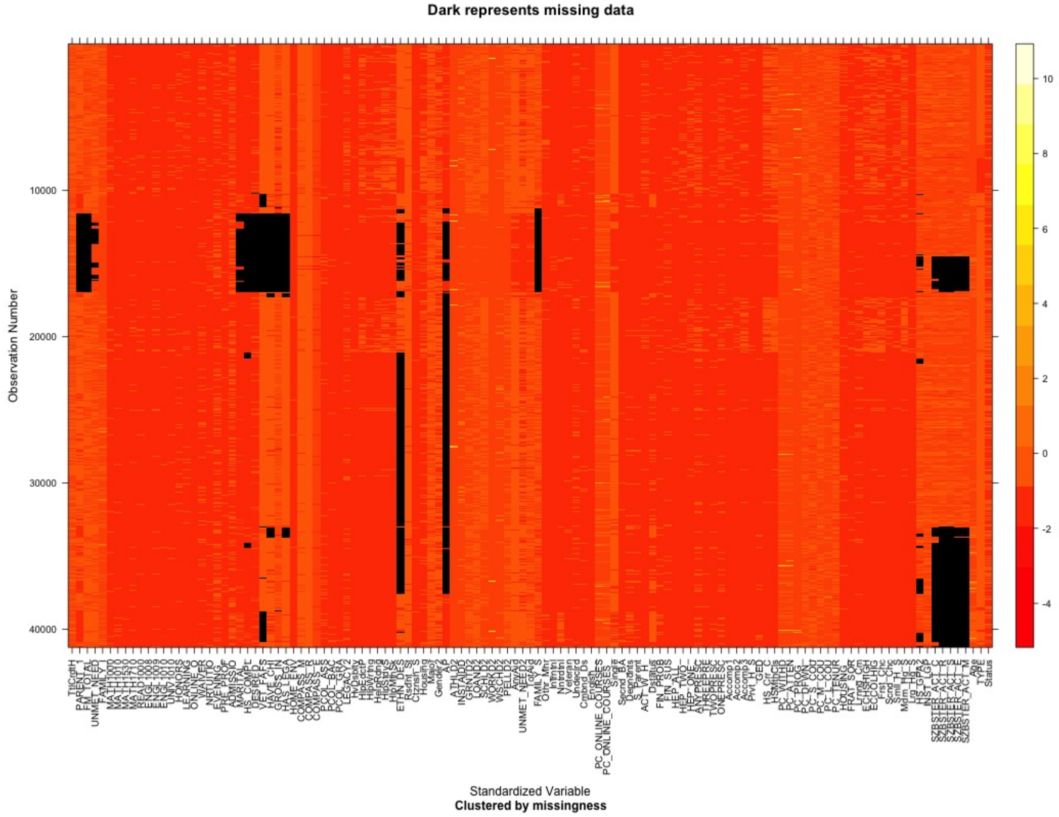


Figure 1: *Missing Data Visualization*

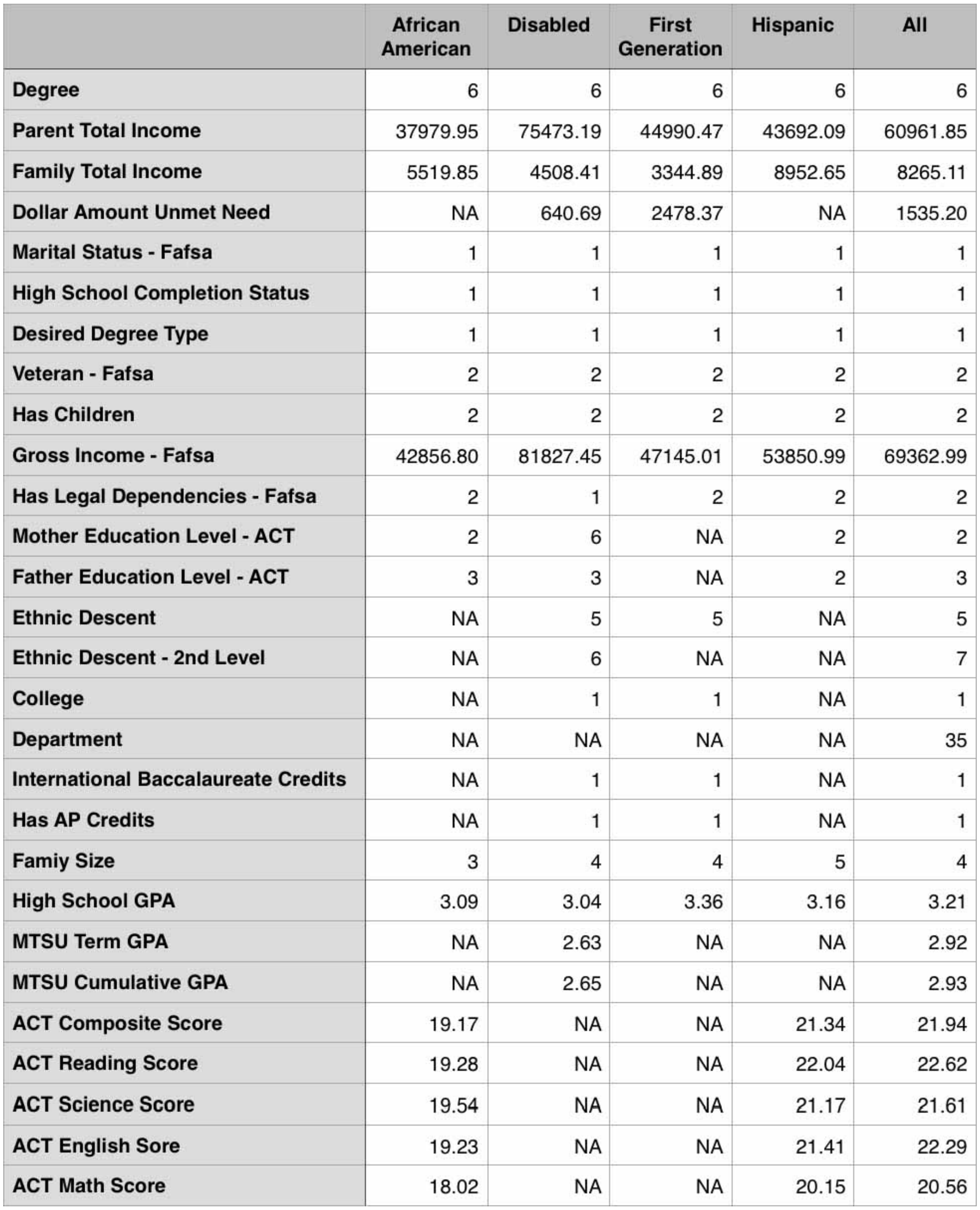
2.4.3 MTSU Data - Mean Substitution

Employing Mean Substitution for our data set is straightforward enough; we simply calculate the mean or mode for each variable and replace the missing values with the calculated one. Given that we desire to explore issues with certain student groups, different values must be calculated for each variable depending on the data set.

To begin, the data was first separated into specific target groups: African American, Disabled, First Generation, Hispanic, and All students. Then Mean Substitution for each variable was performed for each target group. Table 5 below lists the mean values calculated for each missing variable.

It is interesting to note that while some variables exhibited the same value per group, suggesting no real correlation to the student demographic, others had markedly different values. For instance, the Marital Status is always substituted as 1, which makes sense since almost all students are single, regardless of their demographic. However, we can see very different values in continuous variables, especially financial ones. Looking at the values for Parent Total Income, for example, we see a very large difference for substituted values for the groups. Hispanic students are imputed with a value of $43692.09 which is very different from the Disabled students, having a value of $75473.19. However, if we just impute the same value for all students instead, they are both imputed with a value of $60961.85. Thus we can see that filling in financial values based on which group we are imputing can yield markedly different numbers for the same student depending on how we classify them. In this situation it may be hard to tell what the true imputation value should be.

Table 5: Mean Imputation Values By Variable



	African American	Disabled	First Generation	Hispanic	All
Degree	6	6	6	6	6
Parent Total Income	37979.95	75473.19	44990.47	43692.09	60961.85
Family Total Income	5519.85	4508.41	3344.89	8952.65	8265.11
Dollar Amount Unmet Need	NA	640.69	2478.37	NA	1535.20
Marital Status - Fafsa	1	1	1	1	1
High School Completion Status	1	1	1	1	1
Desired Degree Type	1	1	1	1	1
Veteran - Fafsa	2	2	2	2	2
Has Children	2	2	2	2	2
Gross Income - Fafsa	42856.80	81827.45	47145.01	53850.99	69362.99
Has Legal Dependencies - Fafsa	2	1	2	2	2
Mother Education Level - ACT	2	6	NA	2	2
Father Education Level - ACT	3	3	NA	2	3
Ethnic Descent	NA	5	5	NA	5
Ethnic Descent - 2nd Level	NA	6	NA	NA	7
College	NA	1	1	NA	1
Department	NA	NA	NA	NA	35
International Baccalaureate Credits	NA	1	1	NA	1
Has AP Credits	NA	1	1	NA	1
Famiy Size	3	4	4	5	4
High School GPA	3.09	3.04	3.36	3.16	3.21
MTSU Term GPA	NA	2.63	NA	NA	2.92
MTSU Cumulative GPA	NA	2.65	NA	NA	2.93
ACT Composite Score	19.17	NA	NA	21.34	21.94
ACT Reading Score	19.28	NA	NA	22.04	22.62
ACT Science Score	19.54	NA	NA	21.17	21.61
ACT English Sore	19.23	NA	NA	21.41	22.29
ACT Math Score	18.02	NA	NA	20.15	20.56

2.4.4 MTSU Data - Multiple Imputation (MI)

To implement MI, we again separated the data into the specific target groups: African American, Disabled, First Generation, Hispanic, and All before passing on to the mi function. Each table was then converted into a missing data frame object, with variable types being specified according to the type of data it described. While the object provides estimates for variable types, manually specifying each variable ensures the proper function is employed during each cycle of the algorithm. The main mi function was then called, creating 5 separate data sets for each target group. While the mi package has some analysis options as well, for our purposes, the completed data sets were extracted at this point and pooled together to create one complete data set for each group.

In the initial implementation of MI, some difficulties were found. Firstly, there does seem to be a point where the rate of missingness is simply too high for even MI to overcome. What that threshold is varies from variable to variable, but some variables were not able to be filled in. For the Mother and Father Education level, having a 75% missingness rate in the African American students was too high for MI to run to completion. It is interesting to note that MI was able to finish with the Disabled students, even though the missing rate was higher at 78%, which might be attributed to the smaller group size, thus a smaller computational expense. Considering that the overall student missingness rate is 91%, it is questionable if the imputed values found are even usable. Due to this difficulty, some variables had to be removed from the variable set in order to complete MI.

Having completed MI, we can employ the same visual inspection we did previously to see how the imputed values appear in the final model. The image below shows the normalized values for each variable for all students with problematic variables removed. The top graph shows the missing data and the bottom graph shows the corresponding computed data. If there was any distinguishable difference with the sections of black data and the rest of the columns, it would suggest MI may not have properly imputed the values, or the data is MNAR. For the most part the completed data shows no significant difference. The only noticeable variation may be seen in the ACT variables, having slightly less dark red overall, and in the Veteran status variable, with more color variation in the imputed area overall.

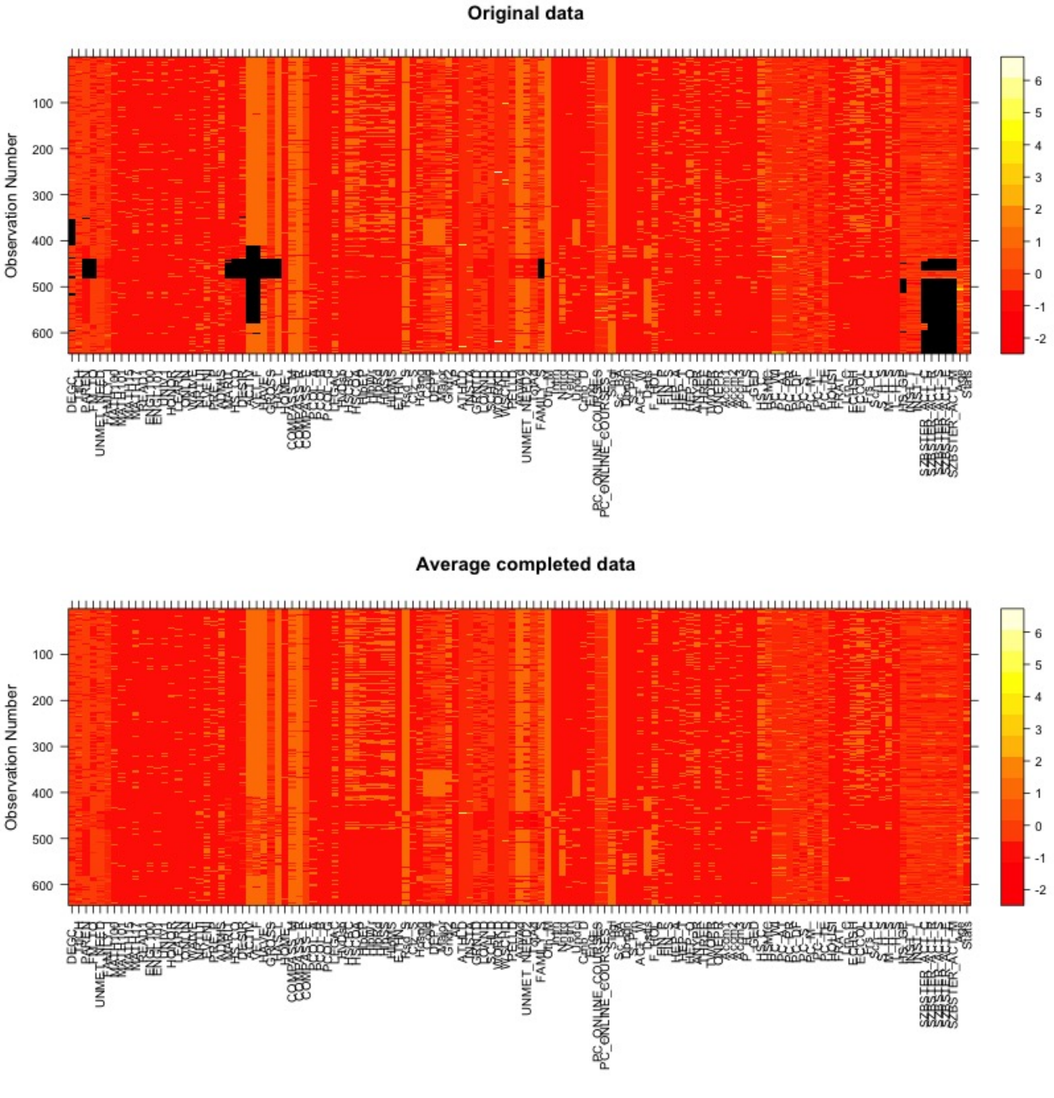
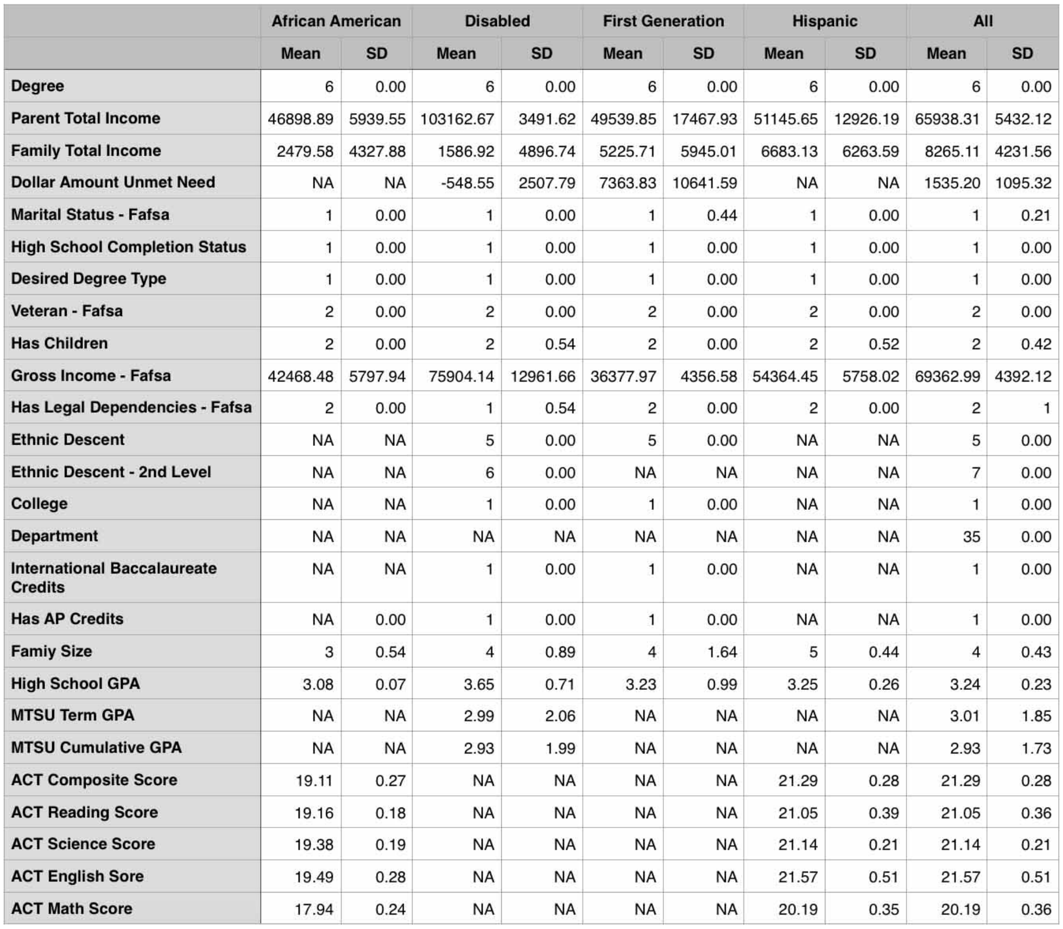


Figure 2: *Comparison of Missing Data and Completed Data in MI*

Along with a visual inspection, we can also look at the mean imputed values and standard deviation for each missing variable over all of the data sets. Table 6 below lists the statistics per variable for all imputed values. With this we can see how close the mean values are to the mean substitution value, but also view the spread of the values around that mean. The smaller the standard deviation is, the more consistently MI was able to impute the same value in each data set.

Table 6: Multiple Imputation Mean Values By Variable



	African American		Disabled		First Generation		Hispanic		All	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Degree	6	0.00	6	0.00	6	0.00	6	0.00	6	0.00
Parent Total Income	46898.89	5939.55	103162.67	3491.62	49539.85	17467.93	51145.65	12926.19	65938.31	5432.12
Family Total Income	2479.58	4327.88	1586.92	4896.74	5225.71	5945.01	6683.13	6263.59	8265.11	4231.56
Dollar Amount Unmet Need	NA	NA	-548.55	2507.79	7363.83	10641.59	NA	NA	1535.20	1095.32
Marital Status - Fafsa	1	0.00	1	0.00	1	0.44	1	0.00	1	0.21
High School Completion Status	1	0.00	1	0.00	1	0.00	1	0.00	1	0.00
Desired Degree Type	1	0.00	1	0.00	1	0.00	1	0.00	1	0.00
Veteran - Fafsa	2	0.00	2	0.00	2	0.00	2	0.00	2	0.00
Has Children	2	0.00	2	0.54	2	0.00	2	0.52	2	0.42
Gross Income - Fafsa	42468.48	5797.94	75904.14	12961.66	36377.97	4356.58	54364.45	5758.02	69362.99	4392.12
Has Legal Dependencies - Fafsa	2	0.00	1	0.54	2	0.00	2	0.00	2	1
Ethnic Descent	NA	NA	5	0.00	5	0.00	NA	NA	5	0.00
Ethnic Descent - 2nd Level	NA	NA	6	0.00	NA	NA	NA	NA	7	0.00
College	NA	NA	1	0.00	1	0.00	NA	NA	1	0.00
Department	NA	NA	NA	NA	NA	NA	NA	NA	35	0.00
International Baccalaureate Credits	NA	NA	1	0.00	1	0.00	NA	NA	1	0.00
Has AP Credits	NA	0.00	1	0.00	1	0.00	NA	NA	1	0.00
Famiy Size	3	0.54	4	0.89	4	1.64	5	0.44	4	0.43
High School GPA	3.08	0.07	3.65	0.71	3.23	0.99	3.25	0.26	3.24	0.23
MTSU Term GPA	NA	NA	2.99	2.06	NA	NA	NA	NA	3.01	1.85
MTSU Cumulative GPA	NA	NA	2.93	1.99	NA	NA	NA	NA	2.93	1.73
ACT Composite Score	19.11	0.27	NA	NA	NA	NA	21.29	0.28	21.29	0.28
ACT Reading Score	19.16	0.18	NA	NA	NA	NA	21.05	0.39	21.05	0.36
ACT Science Score	19.38	0.19	NA	NA	NA	NA	21.14	0.21	21.14	0.21
ACT English Sore	19.49	0.28	NA	NA	NA	NA	21.57	0.51	21.57	0.51
ACT Math Score	17.94	0.24	NA	NA	NA	NA	20.19	0.35	20.19	0.36

We can see from the table that for a lot of the categorical data, the procedure found no difference between the 5 data sets, yielding a standard deviation of 0. For the ones that had a standard deviation greater than 0, it was still small, suggesting that the values were mostly the same. If we look at the values for the categorical data between MI and mean imputation, the mean values are the same, suggesting no impact on the imputation for those values.

After looking at each model separately, we can also compare the values imputed with each method to see how they compare. For the most part the mean imputation values are very close to the MI mean values. Most categorical variables are found to be the same, and most continuous variables are within the range of the standard deviation. Other variables are a bit off from the mean counterparts, most noticeably in the Disabled students, having a Family Total Income of $75,473 in mean imputation and $103,162 in MI. It is in these discrepancies that the effects of imputation might be seen in the predictive model. While overall both methods seem to be comparable for this data set, the variables with larger standard deviations might bring more variation into the final model.

CHAPTER 3

VARIABLE SELECTION AND PREDICTION

We will now discuss variable selection techniques and prediction methods aimed towards classification. Variable selection is an important step in this process, since the size of our data set is too large for us to feasibly understand the relationships between all the variables. If we can reduce the size of the variable set but still maintain as much information within the data set as possible, no essential information will be lost, but the overall accuracy within the subsequent prediction will still be maintained.

3.1 Variable Selection

The goal of this thesis is not only to predict student retention, but also to understand the issues involving retention to improve student success. Recall that the student data set is quite large, having 138 variables. We discussed previously that before analysis was performed, data preparation and cleaning was employed, reducing the variable set to 124. While smaller, this is still significantly large.

In order to better understand the problem, variable selection must be performed in order to more feasibly see how the variables relate to the target variable. Not only does variable reduction help with interpretation, but including additional variables in the model may only provide noise, reducing the overall accuracy. The aim of variable selection is to reduce the data set down to the most significant 30 or so variables, though how to measure the variable significance can vary. It is also not clear if the imputation process will affect the significance measure.

Some progress in this regard has already been achieved in previous research by MTSU faculty [20], which we will examine first before discussing other approaches.

3.1.1 Statistical Bootstrapping

The first approach taken previously was statistical in nature using a bootstrapped t-test method [20]. To ensure that data is well sampled and balanced, a bootstrapping procedure is applied where multiple runs of t-tests were performed on a subset of data extracted randomly from the original data. During each run, 500 data points were used. The t-test significance values over different runs are then averaged to produce the final ranking of the features. The set of features deemed significant against the target variable using this approach is listed in Table A.1 in Appendix A.

3.1.2 Data Mining

The other approach previously taken was to rank order the variables/features in terms of their importance in prediction [20]. The approach aimed to identify individual feature ranking and selection, treating all variables as independent features. Feature predictiveness is computed between the single feature and the target variable. Four feature ranking methods have been applied:

1. Info Gain - Evaluates the worth of a feature by measuring the information gain with respect to the target variable;

2. Gain Ratio - Evaluates the worth of a feature by measuring the gain ratio with respect to the target variable;

3. Chi Squared - Evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the target variable;

4. Correlation - Evaluates the worth of a feature by measuring the Pearson's correlation between it and the target variable.

The feature rankings from each of the four methods were combined through a weighted sum for a final ranking, and the top variables were selected for each group as a final variable set. The final variable sets can be seen in Table A.2 Appendix A.

3.1.3 Model Dependent Methods

For other variable selection methods explored in this thesis, the approach will be implemented within the given model itself. Since the prediction methods we will discuss below are very popular choices, modifications in the general approach have been created in order to incorporate variable selection into the model. We will discuss below the details of the variable selection process below.

3.2 Prediction Methods

Once the initial variable selection has been performed, we can move on to the main goal of this thesis, predicting student retention with a given data set. For comparison of variable sets, each model was run on each variable subset listed above as well as the whole data set to see if any improvements can be made. In addition, each model was run on each student subset and each imputation method, resulting in 52 different models, giving a good selection of models to compare. In taking this approach, differences between variable sets and imputation methods may be discerned.

3.2.1 Notation

To understand the details of the algorithms below, let us define some notation. Unless otherwise specified, a capital letter represents a variable and a lowercase letter represents an instance of the variable.

Let $\mathbf{D}$ denote the completed realization of the dataset defined above, where $\mathbf{D} = \{\mathbf{A}_1, \mathbf{A}_2, \ldots \mathbf{A}_r\}$ with $\mathbf{n}$ instances. For prediction, we must divide the $\mathbf{r}$ variables into two sets, the target variable and predictor variables. Thus for some $j \in 1, 2, \ldots, r$, let $\mathbf{Y} = \mathbf{A}_j$ be the target variable. Then for all $i \in 1, 2, \ldots r, i \neq j$, let $X_i = \mathbf{A}_i$ be a predictor variable and $\mathbf{X} = \{\mathbf{X}_i, i \in 1, 2, \ldots r, i \neq j\}$.

3.2.2 Random Forests (RF)

The first prediction model implemented follows a non parametric approach based on the concept of a tree. Tree based methods are populars choices for prediction based on their simplicity and ease of interpretation; no assumptions about the data need to be made, and its easily adapted to suit a variety of fields. The most basic method, the Decision Tree, consists of making splits in the data on different variables, resulting in leaves, or subgroups. After enough splits are performed, classification can be performed by simply labeling each student depending on what leaf they fall into on the tree [21]. We predict that the student falling into each leaf belongs to the most commonly occurring class in the leaf. In other words, if a student falls into a leaf with 5 dropped observations and 9 stayed observations, then we predict that student will stay.

To build a Decision Tree, a recursive binary splitting scheme is implemented depending on a measure of variance. While RSS is a common metric in regression Decision Trees, for classification problems like the one we face, the Gini Index is often used, being defined as:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

where $\hat{p}_{mk}$ represents the proportion of training observations in the mth region from the kth class.

We can see that the Gini Index is a measure of node purity, since it takes on a small value if all the $\hat{p}_{mk}$ are close to 0 or 1.

To begin the algorithm, we calculate the Gini Index for each variable using the set of all predictors $\mathbf{X}$, then split the data into two leaves on the X_i for which the Gini Index is the lowest. We then recursively split on each leaf in the same manner until no improvements in the Gini Index can be made or a minimum threshold in the number of observations in the leaf is reached. The final result is one tree which can be used to predict future observations.

While it is a good starting point, Decision Trees suffer from one weakness: a split early on can greatly affect the shape of the tree down the road, thus potentially leading to very different predictions depending on the splits made. Random Forests have been proposed as a way reduce these effects and decorrelate the tree [10].

To implement Random Forests, we bootstrap B new data sets from our training data sets, usually 300-500, and create a Decision Tree with each bootstrapped set. Then for each observation, the average of the predictions over all the trees is calculated to get our final prediction $\hat{y}_i$ for each observation. But, instead of splitting on the variable that has the lowest Gini Index every time, we restrict our choices by only considering $m \approx \sqrt{r}$ possible variables at each split, thus using a subset of $\mathbf{X}$ for each split. The m variables are chosen randomly at each split, and thus are uncorrelated with previous splits. In proceeding in this way, the algorithm is not even allowed to consider a large number of the available predictors in each step.

The efficacy of only considering m predictors stems from the idea that if there exists one very strong predictor in the data set, as well as several moderately strong predictors, we want to ensure the moderately strong predictors have a fair chance of being used. If we consider all possible variables at every split, then a large proportion of our bootstrapped trees will all have the strong predictor at the top split, and generally will look quite similar to each other and the predictions per tree will be highly correlated. If instead we restrict our options to only m predictors, we give the moderately strong predictors a chance to appear earlier on, and when averaging the results to get our final prediction, we have a less variable result [10].

3.2.3 RF Variable Selection

The Gini Index, which helps to determine on which variable to split, is also useful for variable selection [10]. While interpreting a single Decision Tree is relatively straightforward, aggregating the trees together might yield an improvement in accuracy, but results in added difficulty in interpretation, since we cannot easily see how one variable performs in several hundred trees together. However, we can still get a measure of variable importance by looking at the Gini Index on each split. If we add up the total amount that the Gini Index is decreased by splits on a given predictor, then average that sum over all the trees, we can calculate a general measure of variable importance within the model. In mathematical notation [10],

$$I = \frac{1}{B}\sum_{k=1}^{B} G_k$$

In this way, we can identify variables that have a high variable importance value as another potential variable subset. Note that we cannot tell the effects of this variable, whether it is positively or negatively correlated or the magnitude of the effect, only that the variable does have some effect.

The final model of a Random Forest results in the incorporation of all variables in the original data set and provides a way to rank the variables to identify weak predictors. It is sometimes preferred to implement Random Forests once to identify the significant variables, then use only the variables with $I \geq \epsilon$ for some specified ϵ in a final model to serve as a variable selection process.

3.2.4 RF Cross Validation

The key component of Random Forests is the number of variables to consider at each split. While $m \approx \sqrt{r}$, the approximation leaves some room for flexibility. In order to select the optimal m to maximize accuracy, cross validation can be employed.

The idea behind cross validation is to use resampling to estimate the accuracy of the model. In cross validation, one begins by splitting the training data into two parts, a training set and a validation set. The model is fit on the training set and used to predict the responses for the validation set. The resulting validation set error rate provides an estimate of the test error rate.

Typical use of cross validation involves splitting the data into k folds, called k-fold cross validation. In this approach, we divide the observations into k equally sized folds. The first fold is treated as the validation set, and the method is fit on the remaining k-1 folds and an error rate calculated. This procedure is repeated on each fold in turn, until we have k error rates. Our final estimate for the error is computed by averaging the k values. Values of 5 or 10 are most common, having been shown to be sufficiently accurate for most purposes.

In our situation, we are not only trying to estimate the test error rate, but also identify the value of m that yields the lowest error. To find m, we must employ a nested cross validation approach. To begin, we execute random forests with all the predictors and use cross validation to estimate the error. Then, we reduce a fixed proportion of variables from the model and run cross validation to estimate the error again. We proceed in this manner until all variables have been removed, then we can look at the range of estimates to find what the optimal m is. While $m \approx \sqrt{r}$, we may find that in some data sets m is slightly above, and others where m is slightly below.

Once the optimal m is found, we can build a final model with the m selected, and employ the variable importance measure to isolate the most important variables.

3.2.5 Logistic Regression (LR)

While trees provide an ease of interpretation, they lack the statistical properties provided by a parametric approach. A popular parametric method is regression, which can easily be adapted for the classification setting by utilizing the logit function.

For Logistic Regression, consider that each observation $y_i|x_i$ has an outcome of either stayed or dropped, or rather 0 or 1. Being a binary outcome, we can consider $y_i|x_i$ as following a Bernoulli distribution with probability p_i. The distribution function is then $f(y_i|x_i) = p_i^{y_i}(1-p_i)^{1-y_i}$, where p_i is a yet to be determined probability.

To apply regression, we need a linear function. While it is tempting to model $p_i = \theta^T x_i$, where $\theta^T x_i = \sum_{j=1}^{r} \theta_j x_{ij}$, this will not work, since we need p_i to be a probability between 0 and 1, and a general linear function can yield probabilities outside this range.

Instead, consider the logit function, $\ln(\frac{p_i}{1-p_i}) = \theta^T x_i$. Notice that in solving for p_i, we find

$$\begin{aligned}
\ln\left(\frac{p_i}{1-p_i}\right) &= \theta^T x_i \\
\frac{p_i}{1-p_i} &= \exp\left(\theta^T x_i\right) \\
p_i &= (1-p_i)\exp\left(\theta^T x_i\right) \\
p_i &= \frac{1}{1+\exp\left(\theta^T x_i\right)}
\end{aligned}$$

Which is clearly a positive value bounded between 0 and 1. Similarly, we can see that $1 - p_i$ also fits the requirements, since

$$1 - p_i = \frac{\exp\left(\theta^T x_i\right)}{1+\exp\left(\theta^T x_i\right)}$$

Now that p_i can be expressed, the goal of regression is to maximize this function with respect to the parameters θ in order to develop a prediction model utilizing the function. This is achieved through maximum likelihood estimation, wherein we maximize the likelihood function.

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{N} f(y_i|x_i) f(x_i) \\
l(\theta) &\propto \sum_{i=1}^{N} \ln f(y_i|x_i)
\end{aligned}$$

Where $l(\theta)$ is the log-likelihood function. In simplifying, we see that the function to maximize is

$$\begin{aligned}
l(\theta) &\propto \sum_{i=1}^{N} \ln f(y_i|x_i) \\
&= \sum_{i=1}^{N} \ln(p_i^{y_i}(1-p_i)^{1-y_i}) \\
&= \sum_{y_i=0} \ln(1-p_i) + \sum_{y_i=1} \ln(p_i) \\
&= \sum_{y_i=0} \ln\left(\frac{\exp(\theta^T x_i)}{1+\exp(\theta^T x_i)}\right) + \sum_{y_i=1} \ln\left(\frac{1}{1+\exp(\theta^T x_i)}\right) \\
&= \sum_{y_i=0} \theta^T x_i - \sum_{y_i=0} \ln(1+\exp(\theta^T x_i)) - \sum_{y_i=1} \ln(1+\exp(\theta^T x_i)) \\
&= \sum_{y_i=0} \theta^T x_i - \sum_{i=1}^{N} \ln(1+\exp(\theta^T x_i))
\end{aligned} \tag{1}$$

Thus logistic regression consists of maximizing above function to find the parameter $\hat{\theta}$ and setting our prediction function to

$$p_i = \frac{1}{1+\exp(\hat{\theta}^T x_i)}$$

Then for new observations, we classify y_i as 1 if $p_i > 0.5$ and 0 otherwise.

3.2.6 LR Variable Selection

In the traditional approach, logistic regression includes all variables in the model, thus calculating a regression coefficient for every variable. As described previously, this is not always desirable, since added variables may provide added noise and reduce accuracy.

The lasso approach was developed with the aim of filtering out variables uncorrelated with $\mathbf{Y}$. The lasso consists of adding an l_1 penalty term to equation (1), maximizing the new function

$$l(\theta) = \sum_{y_i=0} \theta^T x_i - \sum_{i=1}^{N} \ln(1 + \exp(\theta^T x_i)) - \lambda \sum_{j=1}^{p} |\theta_j|$$

By adding the λ penalty term, the lasso shrinks the coefficient estimates towards zero, and in addition forces some of the estimates to be exactly equal to zero when the tuning parameter λ is large enough [10]. Thus, the lasso yields a sparser model than regular logistic regression, giving more interpretable results with a smaller variable set.

In implementing the lasso, setting $\lambda = 0$ yields the traditional logistic regression. Setting λ high enough, however, forces all the coefficients to be zero, yielding no model. The key component in implementing the lasso is setting λ to some intermediate value that maximizes the accuracy, which can be achieved through cross validation.

3.2.7 LR Cross Validation

As in Random Forests, the key component of employing lasso is the tuning parameter λ. Cross validation can be employed in the same manner we saw in Random Forests to find the optimal λ value [10]. But in this case we do not just need the optimal λ, but the specific subset of variables to accompany it.

In designating a training and testing set, some variation may occur between models due to differences in the training data. To remedy this, a nested cross validation approach can be implemented to get a better estimate. To begin, we divide the data into a training and testing set, then perform regular k-fold cross validation on the training set with lasso applied, determining the error rate at successively larger values of λ. We then choose the λ that yields the smallest error rate, and retain those variables included in the model. We then repeat this process, again dividing the data into a new training and testing set and finding a new variable set corresponding to the λ yielding the smallest error. After a sufficient number of iterations, we select the variables that are present in a high enough percentage of instances for the variables to include in the final model. The last step is then to run a single k-fold cross validation on the final variable set to determine the optimal λ for the final model.

CHAPTER 4

MODEL ANALYSIS AND DISCUSSION

This chapter will discuss the results of our analysis, both in terms of imputation and prediction with regards to target groups. The results of the models discussed can be found in Appendix B, with tables ordered by student group.

4.1 Missing Value and Model Analysis

4.1.1 Missing Value Imputation

In the initial implementation of MI, some difficulties were found. Firstly, there is a point where the rate of missingess is simply too high for even MI to overcome. What that threshold is varies from variable to variable, but some variables were not able to be filled in. The rate of missingness varied between the student groups, so while some variables were fine with certain groups, in other groups MI was not able to run to completion. Due to this difficulty, these variables had to be removed from the variable set for certain groups in order to complete MI. This is a disadvantage over mean imputation which could always find a value for all values. However, it also calls to question the validity of the imputed value for mean imputation, and MI might do a better job in identifying that the variable had little statistical insight.

Once the imputation process was complete, some counter intuitive results were found in subsequent analysis for both imputation methods; in performing logistic regression, a positive weight was found for each ACT score variable for both the MI and Mean data sets. MI performed slightly better than Mean imputation in the sense that a smaller positive weight was found for each variable, but even MI was not able to completely eliminate the positive trend. In general, one would assume that a higher ACT score corresponds to a more academically prepared student, and so a higher probability of the student completing their degree. For other GPA related variables, a negative weight was found, supporting the idea that a higher GPA increases their chances of staying, so a further examination of the ACT variables is required.

There are two possible issues to explain this: either the data are not MAR, implying that the methods used are not appropriate, or the rate of missingness is too high for the methods to feasibly impute the values correctly. Considering that the variables have a missingness rate of 25 percent, either or both issues are quite likely, but considering that the visual inspection of the imputed data sets for MI showed a distinguishable pattern in the imputed area, it is likely to be MNAR. With these considerations in mind, little insight may be gained from them that is not already present in other GPA variables.

Indeed, if we just consider Mean imputation, when imputing values for all students, we assign an ACT Composite Score of 21.93. Considering that the 2015 ACT national average score is 21.0 [2], we are artificially assuming the average MTSU student has a higher average than the national average, and the differences become even larger in the subject scores and student groups. MI does a better job preventing this misrepresentation, but cannot completely overcome it. In this instance we must wonder what is the appropriate mean to impute. If the national average is already known, this might be a better value to impute than just using the estimates found within the data set. However, even if this change was made, it still may not overcome the amount of missingness, which warrants removing the variable from the model.

While issues with the ACT variables were apparent within the model, other contradictions were not as clear. Considering that there were other variables with missingness higher than the ACT variables, it is possible that those variables are also too sparse to be recovered. However, the only other variable with a missing percentage higher than 25% that showed up in the final models was the Ethnic Descent, and since that variable was already used in creating the target groups, the other variables with missing data do not present as great an issue, since the prediction model seems to be able to filter them out on its own.

While it does seem that in general both methods are comparable, at least on our data set, the fact that MI was able to find a smaller coefficient with the ACT variables points to the possibility of more accurate results for the remaining variables. Especially as the rate of missingness increases, MI's ability to spread the data better throughout the variable's distribution yields a better chance of representing the true distribution instead of creating an exaggerated peak around the mean.

In examining the effects of imputation on the models themselves, there does not seem to be a significant difference. In both Random Forests and Logistic Regression, the relationships between the variables seem similar. There is slight variation in the smaller groups, given that LR found some coefficients to be zero with one imputation method and nonzero for the other, but given the size of the smaller groups, this is to be expected.

Although MI does require considerably more computational expense, especially as the percentage of missingness increases, the ability to represent the true variable distribution is essential for proper model implementation. Thus, for our final model selection, we will implement MI with the ACT variables removed to ensure the most accurate results.

4.1.2 Variable Selection

In comparing the accuracy of the models with respect to the variable subset, there is a slight difference present in both models. In every target group, we see that the machine learning variable subset was less accurate, sometimes by a considerable amount, whereas the bootstrapping variables and the entire dataset were pretty consistent with each other. It may be that the statistical approach taken for subset selection did a better job at removing the noise in extraneous variables. Although the subsets themselves already reduced the data set, even on the subsets of variables, both methods employed were able to reduce the variable subset further through built in variable selection in the model, removing extra potential noise.

This further reduction may be attributed to the nature of the student groups themselves. Especially when the student group is small, there may be too many variables to find more results. Considering that the Disabled group had 391 students, when looking at all variables, the number of observations may not be big enough to require over 100 variables to explain the relationship to retention. It may be sufficient to use only a handful of variables when only considering a handful of students, since the characteristics that define that group may be less diverse.

While there is some variation in the variables present between student groups, there are a few common variables present in most models. In looking at the Random Forests tables, in general the same variables appear at the top of the list every time. While the degree of importance varies, at least one if not all of the GPA and financial related variables are always ranked high, followed by class percentage variables and aid variables. While the Logistic Regression models found fewer variables overall, there are still a few common variables, which suggests that while the student groups are different, the main factors affecting them may be very much the same. It may be worth considering the few variables that make them unique instead of the many variables that they share to better characterize each group and identify opportunities for intervention.

4.1.3 Prediction Models

In general there was no distinguishable difference between both models in terms of accuracy. Comparisons of the accuracy for both models on each different student group and variable subset yielded approximately the same values, with minor differences which can be contributed to the variation in the model building process itself.

The real difference between the models comes not from accuracy, but interpretability of the results. To understand which model is better, we have to consider what the final goal is; both are useful, but for different purposes. On the one hand, Random Forests provide a way of measuring the overall importance of a variable, whereas Logistic Regression gives us a measure of the actual effect. Random Forests gives us a better measure of how the variable affect the whole student body, but does not tell us what the effect actually is. Logistic Regression may find variables that have large weights and strong correlations, but whose values may apply to only a handful of students. Also, Random Forests only consider a variable at a time when being shaped, so it cannot incorporate correlations between variables, whereas Logistic Regression may filter out variables correlated to each other.

For example, in looking at table B.9, the Random Forests results for all students, we see that the Total Family Income variable finds a variable importance of 475.29 on the MI data set, one of the largest values in the set. We can tell it is important, but we don't know how. If we look at the Logistic Regression model in table B.10, however, we find that it does not even appear on the list. This seems rather odd, but if we consider that the variable may be correlated with other variables in the data set, only using a handful of financial variables is required. Thus it may be that looking at other variables in the model, possibly the dollar amounts of various scholarships and demographic variables, we can already account for whatever information the Total Family Income provides. Random Forests is able to determine that it is important, but by incorporating it into a parametric model, Logistic Regression recognized the correlation with other variables as well.

4.2 Final Model - Ensemble Method

To conclude the analysis, a final model was created to incorporate the benefits found in both models. MI imputation method was used, and our final variable set was pulled from Random Forests using an importance threshold for each group to use only the 30 or so most important variables in the model. To better understand the relationship with the target variable, Logistic Regression was then employed, resulting in the final coefficients in Table B.11. The table is sorted by Variable Importance For All Students, with the corresponding values for the other groups listed alongside.

In general, the target groups have similar variable importance, and while there are slight variations in rankings for other groups, the degree of the differences between are very small, implying that in general the variables have the same effect overall for the student population. It is also worth noting that the most significant variables were found to be significant for all groups, and only farther down the list does some variation occur. Considering the Disabled and Hispanic groups are also the smallest groups, the fact that they did not find some variable important, like the Financial Probation, might be attributed to the smaller group size.

While it is still a complex model, intervention may be most successful by considering two different goals. For intervention at the individual student level, the weights of each variable can be considered in the advising process. Variables with positive weights can be seen as potential risks, and the larger the weight is on the variable, the stronger the effect towards increasing the probability of dropping out. If a student has a high value for one of those variables, care should be taken to ensure the student has proper support to be successful. For example, the largest positive weight seen for all students is 1.107 for Financial Probation. If a student falls in this category, we can assume that their probability of dropping out is increased a lot more by that variable than their Percentage of Large Courses, which only has a weight of 0.001. Advising for this student should take this into consideration to help them prioritize getting off Financial Probation.

If however, we want to identify factors overall that affect all students, we can instead focus on the variables with higher Variable Importance. We can use those to try to improve retention for students overall. For example, we can infer that the GPA variables, being first in the list, affect the entire student body the most, so promoting ways to increase their GPA will go a long way towards increasing retention for the entire student body.

To further explore factors affecting all students, let us look closer at a few interesting variables.

4.2.1 Financial Variables

One adjustment that was made with the variable selection found from the Random Forests results was the removal of the Family Total Income and Gross Family Income variables. As noted previously, while deemed to be important by Random Forests, including these variables into the model led to unusual behavior within the student groups. Both positive and negative weights were found depending on the student group, and even the two variables themselves had both positive and negative weights within one group, even though they would seem to be measuring a similar thing. If we consider that the actual weights found for these variables were rather small, their role in the model is small in comparison to other variables. In fact, in removing these variables, no accuracy in the model was lost, which points to the variables already being explained by other variables in the group. Indeed, since student funding might require this information when determining which students qualify for what aid, the information within these variables are already seen in the data.

Part of the difficulty in these variables is interpreting exactly what they represent. When talking about the income for the whole family, this may or may not include extra sources of income from other members in the family, including the student. Breaking apart the variable into specific sources may lead to a more insightful variable, since it would be easier to see exactly how the income is defined for each student, but just looking at the aggregate level does not give a clear picture across the whole student body. Attempting to piece apart the variables by using the difference between Parent Income and Total Income as a measure for the student's income did not make the distinction any more clear.

From this we can see that a student's financial situation does not strongly affect their chances of academic success. Having a large income does not automatically increase their chances of graduating, nor does having a small income increase the chances of dropping. There is already a lot of support that goes into the aid granting process, so much so that the students needing support are able to find it. At this time, we are not able to incorporate the knowledge behind the aid granting process, so the models suffer from that weakness. Looking at the picture from the aggregate level as we are, we cannot clearly see the trend. In fact, only including the Parent Total Income still achieved the same accuracy with a clearer picture as to how the variable effects retention; when strictly looking at the parents and not the family as whole, we can conclude that generally the more income the parents make, the more likely the student is to complete their degree.

4.2.2 Scholarship Variables

The same issues seen with the income variables are also seen with the various scholarship variables for similar reasons. In looking at the different types of aid, there does not seem to be a very clear picture as to how these variables affect retention. With positive weights for some students and negative weights for others, it may be that the factors that go into determining which students qualify for different funding is too complicated to explain with an aggregate picture such as this. To gain more insight and predictive power from these variables, further variables may be required to incorporate the insight that went into assigning funding to the student.

Some of the difficulty in understanding these variables can be seen by examining the box plots by target group. For most variables in the study, the box plots had the same general shape over all student groups, which points to a similar distribution for each variable among the different groups. The only variables with markedly different shapes for one or more target groups are the Dollar Amount of Loans and Dollar Amount of Scholarships.

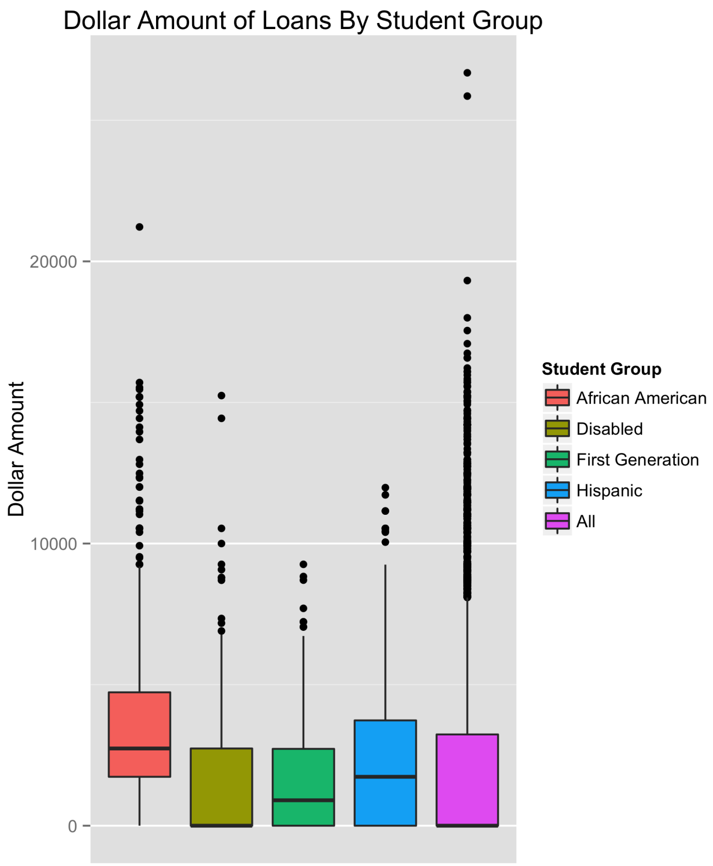


Figure 3: *Loan Distribution By Target Group*

In looking at Figure 3 above depicting student loans, it is very clear that African American students are more likely to have loans than the other groups. With the median falling at about $2500 and the box representing the middle 50 percent being above the axis, we can see that a large percentage of the group have student loans. When considering that the lower whisker of the plot extends to 0, while it does imply that some students do not have loans, it does mean that the majority do, with only a few being at the tail end of the distribution. In looking at the other groups, while the First Generation and Hispanic students have a similar mean value, both of their boxes spread to 0, and for the other groups, the median value is 0, which tells us that the middle 50 percent of the other groups do not have any student loans. If we recall that this study is only focusing on freshman students, the fact that African American students are more likely to have student loans from the beginning of their college career is a very insightful observation.

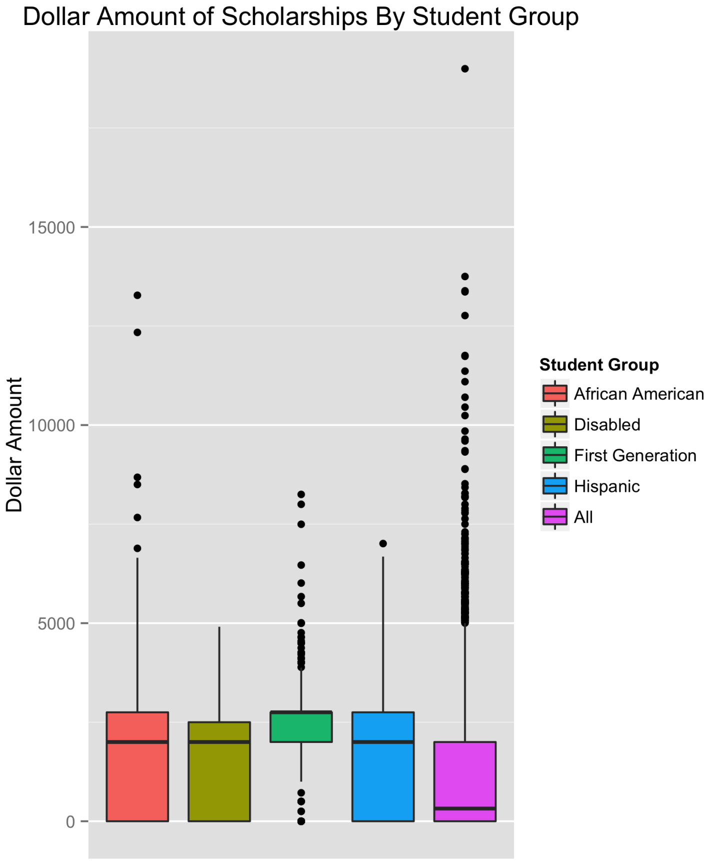


Figure 4: *Scholarship Distribution By Target Group*

Figure 4 above depicts the box plots for scholarships by student group, and here we see that the First Generation group is very different from the remaining groups. If we notice that the median value for all minority groups is around $2500, while the value for the non target group students is close to 0, we can see that in general minority students are more likely to have a scholarship. The spread around that median value, though, is very distinct for the First Generation students. In fact, the box and whiskers are even more narrow than in the previous box plot, suggesting that almost all First Generation students have a scholarship in the range, and even the outlier values are for the most part higher in value than the other groups.

While it is neither good nor bad that these differences in aid exist, it is worth noting because it paints a very different situation for those students. Especially with regard to loans, there is more at risk for the student to take out a loan, since they are investing in their education with the expectation of a career in the future to facilitate paying back the loan. If the student drops out of school, they have a financial burden without the anticipated reward, thereby putting themselves in financial strain with no clear solution as to paying it back. The that fact African American students are more likely to have loans gives a different understanding to the effect of that variable in the model, since these students might view loans differently than other groups. They might be more grateful for the loans than other students, so we have to keep in mind that the context by which the aid was received. The fact that African American students have a negative coefficient for the dollar amount of loans tells us that in general, having loans increases their chances of staying. It may be that these students are motivated to pay back those loans, thereby increasing their likelihood of success.

In examining these variables within the model, we have to keep in mind that the process by which aid is determined for each student is very complicated, and at this time we can only incorporate the aggregate picture into the model. More accurate prediction may be achievable with further exploration of the relationships behind these variables.

4.2.3 Age Variable

In general, we tend to assume that the older a person is, the more prepared they are to handle the demands college brings. The older, nontraditional students generally have more responsibilities, having their own families and jobs, so there is a different intent to pursuing their degree. The Logistic Regression weight for all students reflects this assumption, however, in the target groups, the same behavior is not seen.

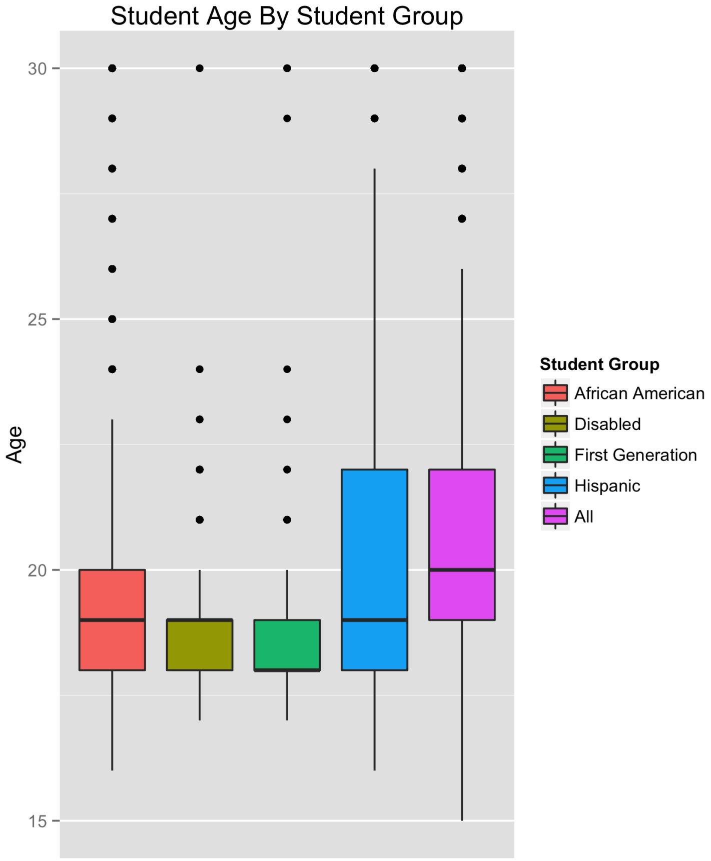


Figure 5: *Age Distribution of Students By Target Group*

If we look at Figure 5, the extreme values may show some insight into this behavior. In looking at the median value, we can see that the different groups generally have the same age, with the First Generation being slightly younger. However, the spread of values is much narrower in the target groups. Especially for the Disabled and First Generation students, the box representing the middle 50 percent is very narrow, and the upper and lower whiskers are also quite small. The remaining dots representing the outliers, even after being adjusted as mentioned in Chapter 2, show a very far spread from the rest of the group. Even though the Hispanic group is closest to representing the remaining students, the Hispanic group is also significantly smaller. It may be that these outliers are affecting the overall model, and removing them completely may reverse the effect. Whether or not they should be removed is something to consider, but at least in general, we can feel pretty confident that for most students, age has a positive effect towards finishing their degree.

4.2.4 Housing Variables

One surprising result from this model is that there is no clear gain from the student's living status. As mentioned in previous research in Chapter 1 [9], research done at other universities found a positive correlation between living on campus and finishing the degree, and indeed some universities require out of town freshman to live on campus their first year. This effect was not found with the MTSU student body, which may point to a different campus culture at MTSU than other universities. While a positive effect was found for all students living with parents, it was also found for living on campus. The weight for living on campus was smaller, but still positive. Living on campus actually had the opposite effect for First Generation students, but it did not even appear in the other target groups.

If we look at the initial models, we do see a negative weight for the target groups when Living On Campus appears in the model, thus it is possible that living on campus for the target groups does improve retention; however for the general student the effect does not seem to help significantly with retention. If MTSU were to implement this change to improve retention, the same results as other universities may not be seen, or at least not to the same degree of success. This may be contributed to the location of the university; it is possible that universities in bigger urban areas have different effects and strains on students, so the typical student living off campus at other universities may have other factors to consider.

4.2.5 Compass Variables

Looking at the Compass Variables, English has a consistent negative weight across the groups, suggesting that the Compass test is going a good job for successful class placement. The Math score, however, has variation within the groups. While it is negative for most, the African American students have a positive weight. Considering that this test figures in to class placement, this might explain why the Lasso results in the previous models identified math courses for these students and not the others. In fact, this may be why Math 1710 is present for African American students in the final model, and why it had a positive weight.

If we keep in mind that Random Forests assigns a variable importance for every variable, although we only selected the 30 or so most important variables for each group, had we adjusted the cut off threshold, the Compass Variables may appear for the other groups, though they are slightly more significant for the groups currently found. Including these variables in the model yielded negative weights for all groups with Compass English, but positive weights for Compass Math with the African American and Disabled students. If we examine it further and include Math 1710 in the model for all students, we see a positive weight for every student group, which was the only course to get consistent behavior across demographics.

Interpreting these results is difficult, but especially in regards to Math 1710, while we cannot say that the course is causing students to drop out of school, there is a positive effect between the variable and retention. What this relationship is exactly is not easy to determine; it could be a form of selection bias, where students in certain majors are more likely to take the course, or students less comfortable in math may choose to take the course over other courses, or perhaps students just need a stronger foundation before being ready for the material in the course. It may also be that the Compass test could use further refinement to better measure a student's preparedness for Math 1710. Whatever the case may be, further investigation of the course and course placement may help improve student success.

Another factor that may be considered in the effects of the math courses is the material itself. While the other math courses do not show up as often in the models, when they appear, we also see positive weights for 1010 and negative weights for 1530. Math 1710 covers College Algebra, 1010 covers Math for General Studies, and 1530 covers Applied Statistics. Since 1530 has a different effect than the other courses, it could be that the course material attracts certain students to take it, or during the advising process, the students that would find it useful are being encouraged to take it. Considering that 1710 covers material they have seen in high school, the material in 1530 is new to many students, which may make the class more stimulating. Examining the effects of advising and course selection may lead more insight as to why students are taking one course over another.

4.3 Concluding Remarks

The goal of this thesis is to develop a better understanding of student retention in the hopes of identifying factors contributing to dropping out of school. By exploring various predictive models and variable selection, we now have a better understanding of variables related to student retention. Both Logistic Regression and Random Forests resulted in models with similar accuracy, but both models had different benefits in regards to understanding. The final model incorporated the benefits from both models to create the most insightful results. The impact of missing values were taken into consideration, and MI imputation was found to be better in recovering the missing data.

While in general it is a very complex problem, by analyzing these models, a few variables point to areas that may help improve retention for students. By analyzing variables with large Variable Importance, we can identify factors affecting a large portion of students. By analyzing variables with large coefficients, we can identify factors affecting a particular student. Further exploration of the variables mentioned above may lead to the most impaction program for imploring MTSU student retention.

BIBLIOGRAPHY

[1] Acock, A. C. "Working With Missing Values." *J. Marriage & Family* 67 (2005): 1012-1028.

[2] ACT. "Average Scores by State." *2015 ACT National and State Scores.* Web. www.act.org/newsroom/data/2015. 16 Sept. 2015.

[3] Associated Press. "MTSU Making Efforts to Increase Graduation Rate." *Diverse Issues in Education.* 12 June 2012. Web. www.diverseeducation.com. 19 Dec. 2015.

[4] Astin, A. "How Good is Your Institutions Retention Rate?" *Research in Higher Education* 38 (1997): 647-658.

[5] Collins, L. M., Schafer, J. L. , and Kam, C. M. "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures." *Psychological Methods* 6 (2001): 330-51.

[6] Enders, C. K. "A Primer on Maximum Likelihood Algorithms Available for Use With Missing Data." *Structural Equation Modeling: A Multidisciplinary Journal* 8 (2001): 128-41.

[7] Gelman, A., and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* New York NY: Cambridge UP, 2007.

[8] Graham, J. W. *Missing Data Analysis and Design.* New York NY: Springer, 2012.

[9] Hanover Research. "How Data Mining Helped 11 Universities Improve Student Retention Strategies." *Hanover Research.* Web. www.hanoverresearch.com/insights/how-11-universities-will-improve-student-retention/?i=higher-education. 19 Dec. 2015.

[10] Hastie, T., Tibshirani, R. and Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.

[11] Kaiser, J. “The Robustness of Regression and Substitution by Mean Methods in Handling Missing Values.” Annual Islamic Conference on Statistical Sciences. Johor Bahru, Malaysia. 26 Aug. 1990.

[12] Pan, W., Guo, S., Alikonis, C. and Bai, H. “Do Intervention Programs Assist Students to Succeed in College?: A Multilevel Longitudinal Study.” *College Student Journal* 42 (2008): 90-98.

[13] Rubin, D. B. “Inference and Missing Data.” *Biometrika* 63 (1976): 581-92.

[14] Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons, 1987.

[15] Schafer, J. L. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, 1997.

[16] Sinharay, S., Stern, H. S., and Russell, D. “The Use of Multiple Imputation for the Analysis of Missing Data.” *Psychological Methods* 6 (2001): 317-29.

[17] Su, Y. S., and Gelman, A. “Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box.” *Journal of Statistical Software* 45 (2011): 1-31.

[18] Tinto, V. *Leaving College: Rethinking the Causes and Cures of Student Attrition*, Chicago, IL: U of Chicago, 1993.

[19] Von Hippel, P. T. “How Many Imputations Are Needed? A Comment on Hershberger and Fisher.” *Structural Equation Modeling* 12 (2005): 334-35.

[20] Wallin, J., Wu, Q., Hains, M., and Li, C. “Improving Minority Student Success through Data Driven Analysis.” Technical Report, MTSU, 2015.

[21] Witten, I. H., and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam: Morgan Kaufman, 2005.

[22] Zhu, B., He, C. and Liatsis, P. “A Robust Missing Value Imputation Method For Noisy Data.” *Appl Intell* 36 (2012): 61-74.

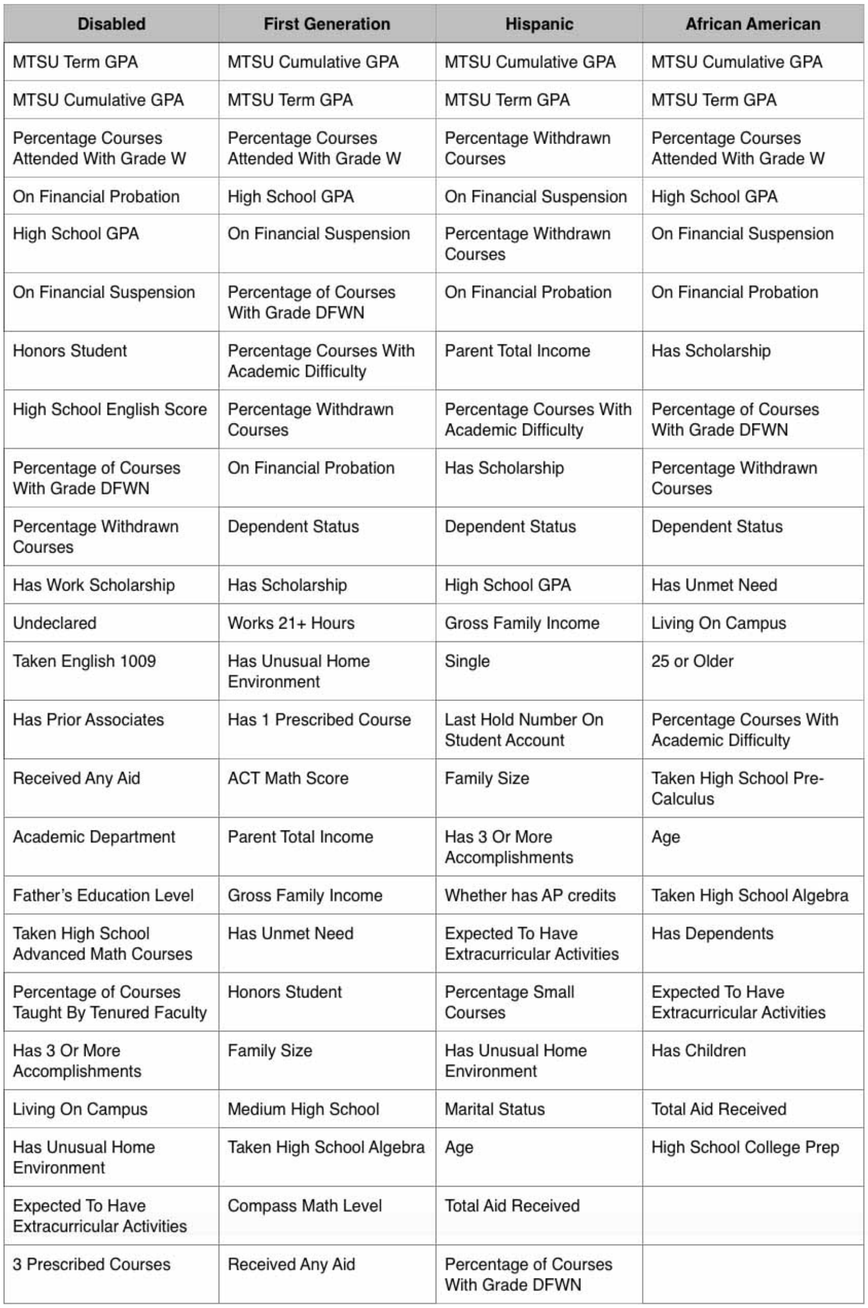
APPENDICES

APPENDIX A
VARIABLE SUBSETS

The following tables contain information found in the previous research [20]. Table A.1 and A.2 list the variable subsets found, as mentioned in chapter 3. Table A.1 shows the variable subsets found from a bootstrapped approach, performing multiple runs of t-tests on a subset of data extracted randomly from the original data. Table A.2 shows the results from a ranking approach, aiming to identify individual feature ranking and selection, treating all the features as independent features. Feature predictiveness is computed between the single feature and the target variable based on several different ranking methods.

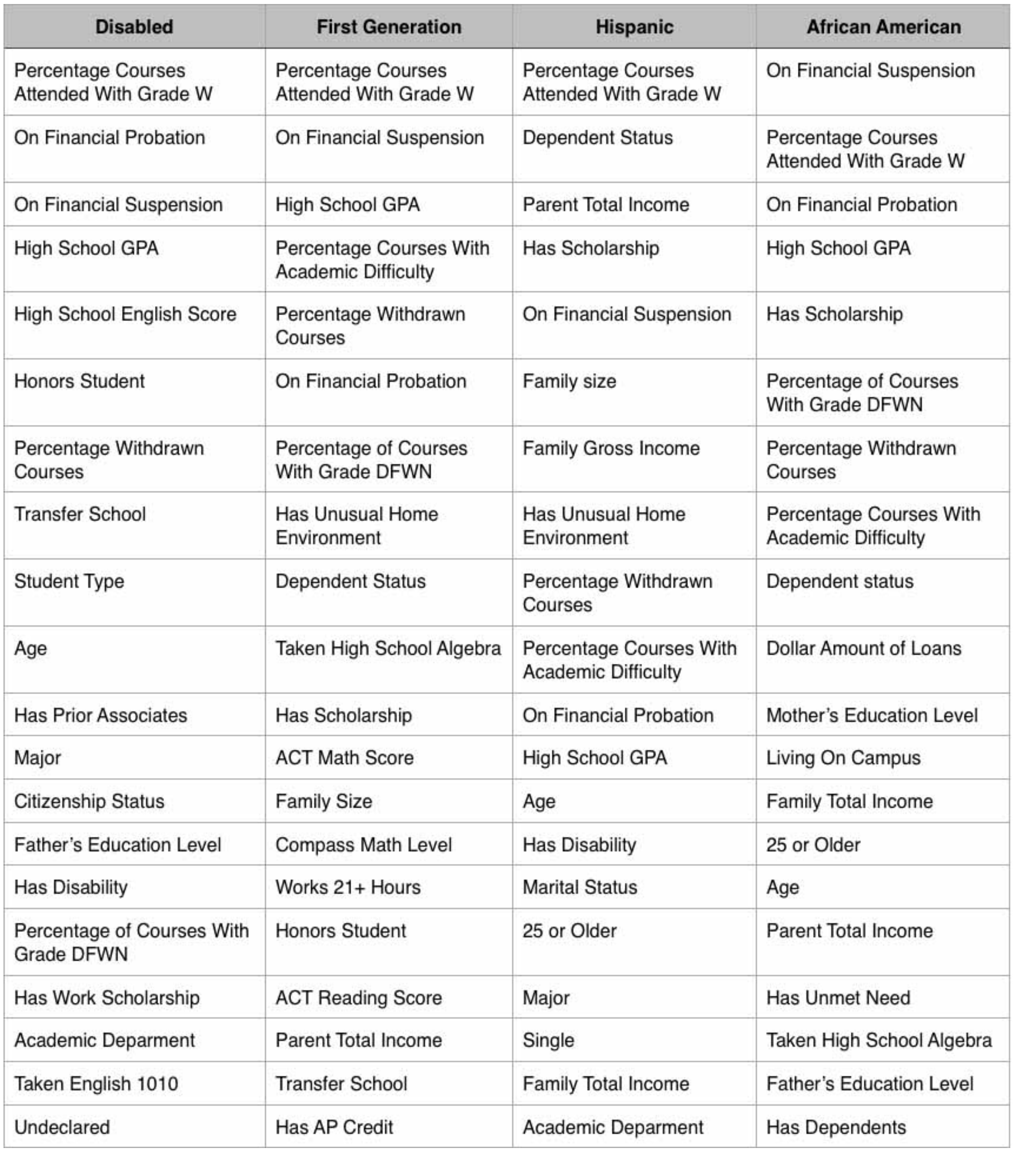
Variables are ranked in order of importance within the group. These subsets are used in subsequent analysis, as outlined in chapter 3.

Table A.1: Statistical Bootstrapping Variable Selection



Disabled	First Generation	Hispanic	African American
MTSU Term GPA	MTSU Cumulative GPA	MTSU Cumulative GPA	MTSU Cumulative GPA
MTSU Cumulative GPA	MTSU Term GPA	MTSU Term GPA	MTSU Term GPA
Percentage Courses Attended With Grade W	Percentage Courses Attended With Grade W	Percentage Withdrawn Courses	Percentage Courses Attended With Grade W
On Financial Probation	High School GPA	On Financial Suspension	High School GPA
High School GPA	On Financial Suspension	Percentage Withdrawn Courses	On Financial Suspension
On Financial Suspension	Percentage of Courses With Grade DFWN	On Financial Probation	On Financial Probation
Honors Student	Percentage Courses With Academic Difficulty	Parent Total Income	Has Scholarship
High School English Score	Percentage Withdrawn Courses	Percentage Courses With Academic Difficulty	Percentage of Courses With Grade DFWN
Percentage of Courses With Grade DFWN	On Financial Probation	Has Scholarship	Percentage Withdrawn Courses
Percentage Withdrawn Courses	Dependent Status	Dependent Status	Dependent Status
Has Work Scholarship	Has Scholarship	High School GPA	Has Unmet Need
Undeclared	Works 21+ Hours	Gross Family Income	Living On Campus
Taken English 1009	Has Unusual Home Environment	Single	25 or Older
Has Prior Associates	Has 1 Prescribed Course	Last Hold Number On Student Account	Percentage Courses With Academic Difficulty
Received Any Aid	ACT Math Score	Family Size	Taken High School Pre-Calculus
Academic Department	Parent Total Income	Has 3 Or More Accomplishments	Age
Father's Education Level	Gross Family Income	Whether has AP credits	Taken High School Algebra
Taken High School Advanced Math Courses	Has Unmet Need	Expected To Have Extracurricular Activities	Has Dependents
Percentage of Courses Taught By Tenured Faculty	Honors Student	Percentage Small Courses	Expected To Have Extracurricular Activities
Has 3 Or More Accomplishments	Family Size	Has Unusual Home Environment	Has Children
Living On Campus	Medium High School	Marital Status	Total Aid Received
Has Unusual Home Environment	Taken High School Algebra	Age	High School College Prep
Expected To Have Extracurricular Activities	Compass Math Level	Total Aid Received	
3 Prescribed Courses	Received Any Aid	Percentage of Courses With Grade DFWN	

Table A.2: Data Mining Variable Selection



Disabled	First Generation	Hispanic	African American
Percentage Courses Attended With Grade W	Percentage Courses Attended With Grade W	Percentage Courses Attended With Grade W	On Financial Suspension
On Financial Probation	On Financial Suspension	Dependent Status	Percentage Courses Attended With Grade W
On Financial Suspension	High School GPA	Parent Total Income	On Financial Probation
High School GPA	Percentage Courses With Academic Difficulty	Has Scholarship	High School GPA
High School English Score	Percentage Withdrawn Courses	On Financial Suspension	Has Scholarship
Honors Student	On Financial Probation	Family size	Percentage of Courses With Grade DFWN
Percentage Withdrawn Courses	Percentage of Courses With Grade DFWN	Family Gross Income	Percentage Withdrawn Courses
Transfer School	Has Unusual Home Environment	Has Unusual Home Environment	Percentage Courses With Academic Difficulty
Student Type	Dependent Status	Percentage Withdrawn Courses	Dependent status
Age	Taken High School Algebra	Percentage Courses With Academic Difficulty	Dollar Amount of Loans
Has Prior Associates	Has Scholarship	On Financial Probation	Mother’s Education Level
Major	ACT Math Score	High School GPA	Living On Campus
Citizenship Status	Family Size	Age	Family Total Income
Father’s Education Level	Compass Math Level	Has Disability	25 or Older
Has Disability	Works 21+ Hours	Marital Status	Age
Percentage of Courses With Grade DFWN	Honors Student	25 or Older	Parent Total Income
Has Work Scholarship	ACT Reading Score	Major	Has Unmet Need
Academic Deparment	Parent Total Income	Single	Taken High School Algebra
Taken English 1010	Transfer School	Family Total Income	Father’s Education Level
Undeclared	Has AP Credit	Academic Deparment	Has Dependents

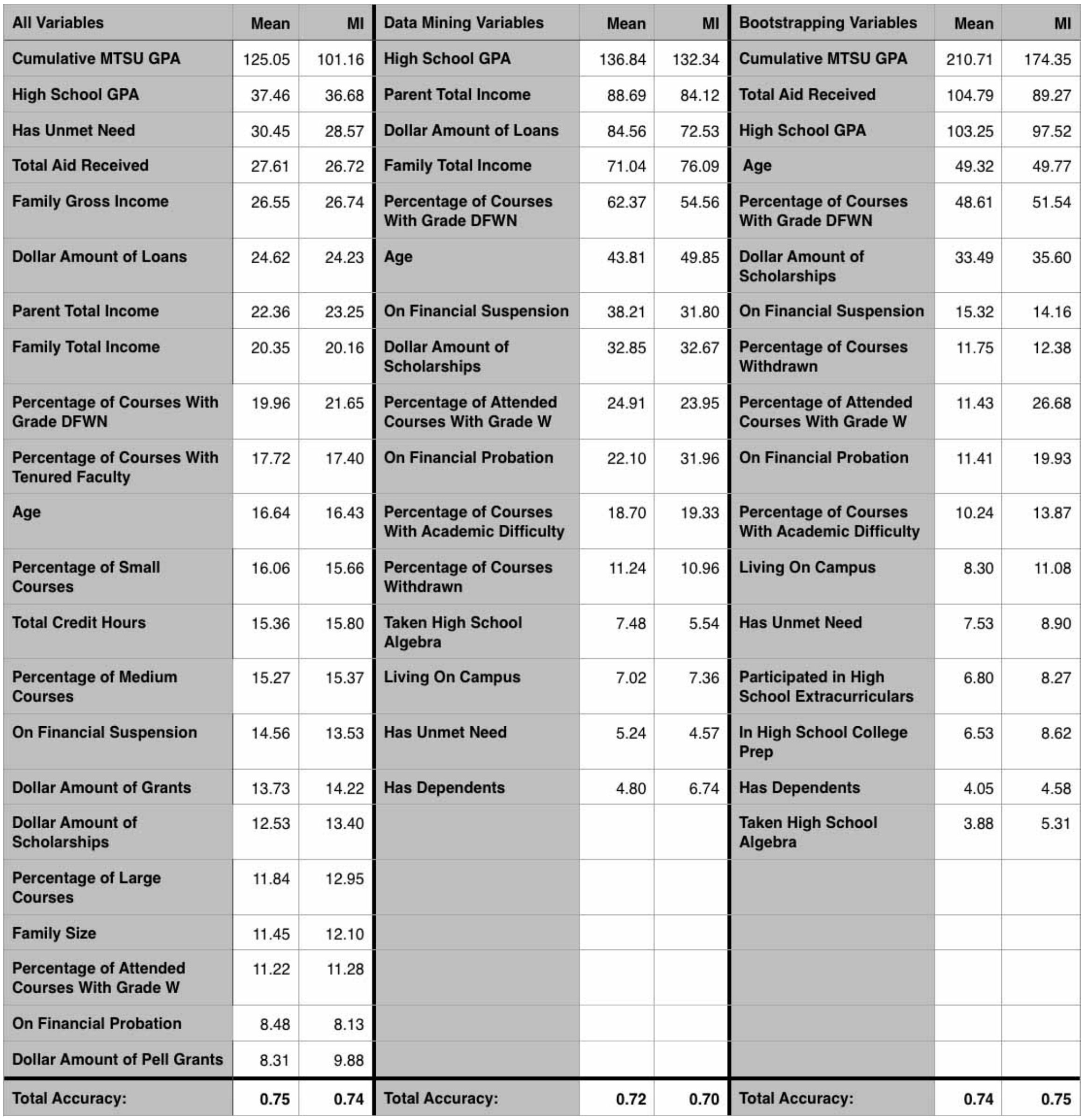
APPENDIX B
MODEL RESULTS

The following tables list the results from each model. The tables are grouped by student type, then broken down into variable subset and imputation type. For the Random Forests tables, variables are listed in order of variable importance above a certain threshold which varied by student group. The variables listed are large enough in value to be noticeably different from the remaining variables. The Logistic Regression variables are listed in no specific order.

Table B.11 lists the results of the final model. For this model, MI imputation was used for missing values and Random Forests was used for variable selection to find the 30-35 most important variables per group. The model used for prediction was Logistic Regression. The table is sorted by variable importance according to All Students, with the corresponding values for the other groups listed alongside.

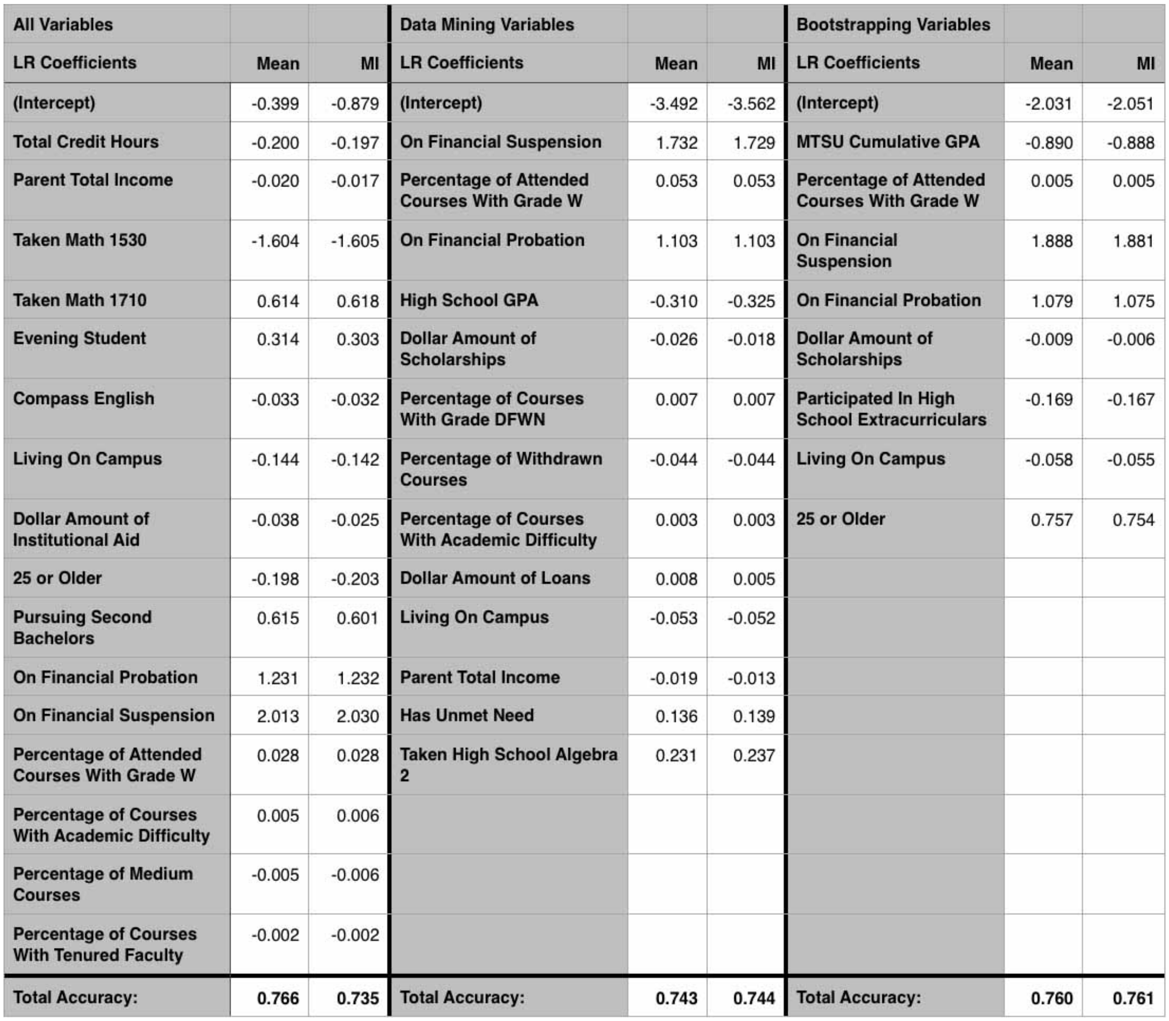
Table B.12 groups the Logistic Regression weights in the final model by their size and effect. Variables with weights larger than 0.1 were labeled strong, between 0.01 and 0.1 as moderate, and less than 0.01 as weak.

Table B.1: Random Forest Variable Importance For African American Students



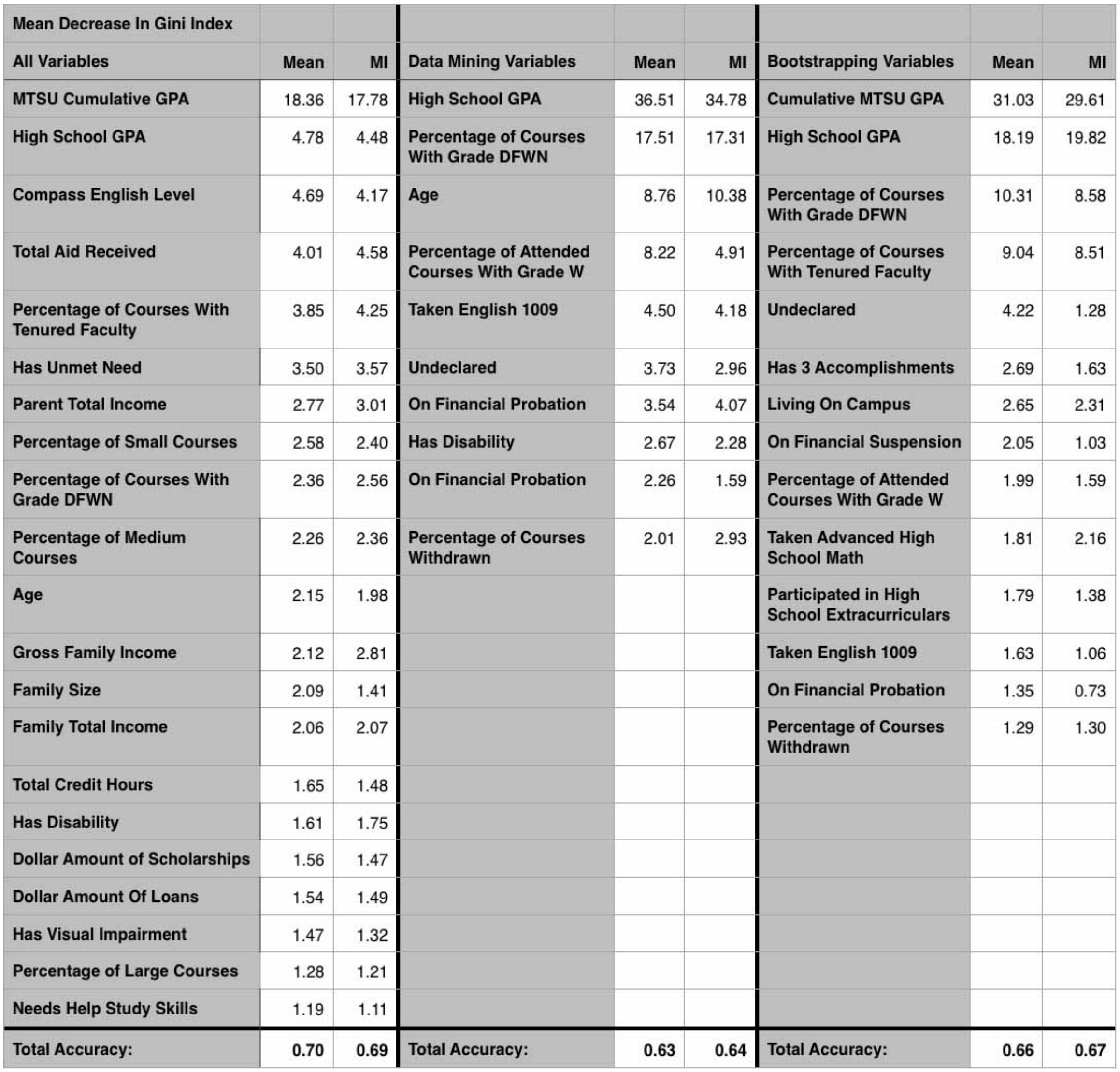
All Variables	Mean	MI	Data Mining Variables	Mean	MI	Bootstrapping Variables	Mean	MI
Cumulative MTSU GPA	125.05	101.16	High School GPA	136.84	132.34	Cumulative MTSU GPA	210.71	174.35
High School GPA	37.46	36.68	Parent Total Income	88.69	84.12	Total Aid Received	104.79	89.27
Has Unmet Need	30.45	28.57	Dollar Amount of Loans	84.56	72.53	High School GPA	103.25	97.52
Total Aid Received	27.61	26.72	Family Total Income	71.04	76.09	Age	49.32	49.77
Family Gross Income	26.55	26.74	Percentage of Courses With Grade DFWN	62.37	54.56	Percentage of Courses With Grade DFWN	48.61	51.54
Dollar Amount of Loans	24.62	24.23	Age	43.81	49.85	Dollar Amount of Scholarships	33.49	35.60
Parent Total Income	22.36	23.25	On Financial Suspension	38.21	31.80	On Financial Suspension	15.32	14.16
Family Total Income	20.35	20.16	Dollar Amount of Scholarships	32.85	32.67	Percentage of Courses Withdrawn	11.75	12.38
Percentage of Courses With Grade DFWN	19.96	21.65	Percentage of Attended Courses With Grade W	24.91	23.95	Percentage of Attended Courses With Grade W	11.43	26.68
Percentage of Courses With Tenured Faculty	17.72	17.40	On Financial Probation	22.10	31.96	On Financial Probation	11.41	19.93
Age	16.64	16.43	Percentage of Courses With Academic Difficulty	18.70	19.33	Percentage of Courses With Academic Difficulty	10.24	13.87
Percentage of Small Courses	16.06	15.66	Percentage of Courses Withdrawn	11.24	10.96	Living On Campus	8.30	11.08
Total Credit Hours	15.36	15.80	Taken High School Algebra	7.48	5.54	Has Unmet Need	7.53	8.90
Percentage of Medium Courses	15.27	15.37	Living On Campus	7.02	7.36	Participated in High School Extracurriculars	6.80	8.27
On Financial Suspension	14.56	13.53	Has Unmet Need	5.24	4.57	In High School College Prep	6.53	8.62
Dollar Amount of Grants	13.73	14.22	Has Dependents	4.80	6.74	Has Dependents	4.05	4.58
Dollar Amount of Scholarships	12.53	13.40				Taken High School Algebra	3.88	5.31
Percentage of Large Courses	11.84	12.95						
Family Size	11.45	12.10						
Percentage of Attended Courses With Grade W	11.22	11.28						
On Financial Probation	8.48	8.13						
Dollar Amount of Pell Grants	8.31	9.88						
Total Accuracy:	0.75	0.74	Total Accuracy:	0.72	0.70	Total Accuracy:	0.74	0.75

Table B.2: Logistic Regression Coefficients For African American Students



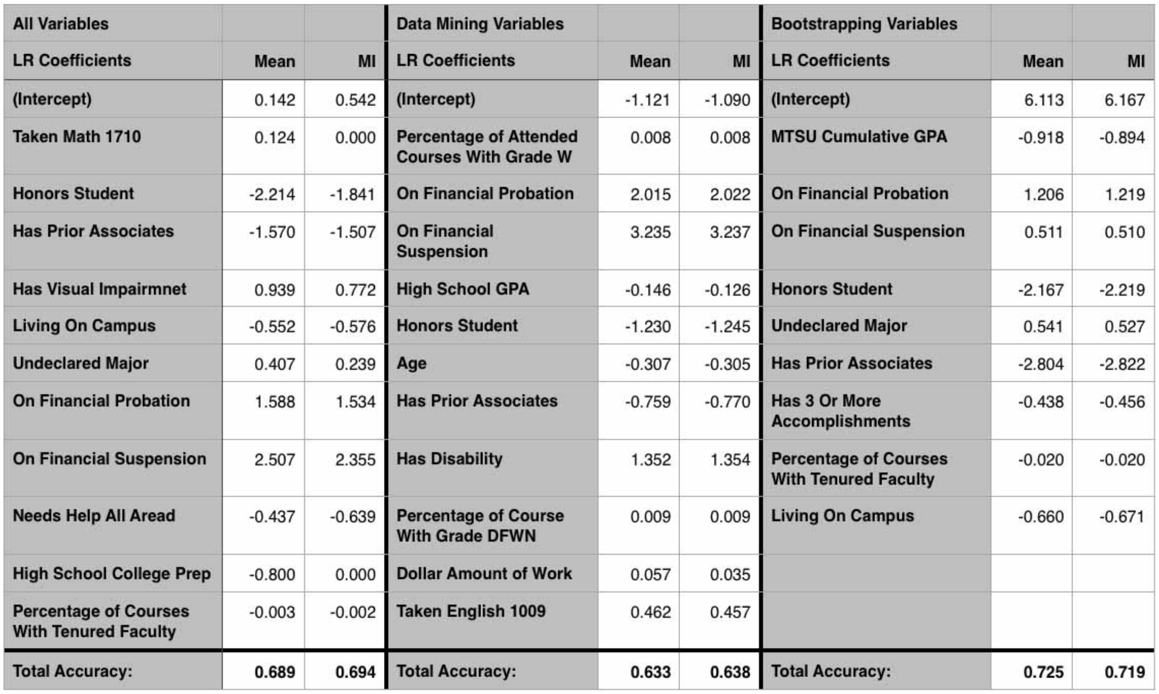
All Variables			Data Mining Variables			Bootstrapping Variables		
LR Coefficients	Mean	MI	LR Coefficients	Mean	MI	LR Coefficients	Mean	MI
(Intercept)	-0.399	-0.879	(Intercept)	-3.492	-3.562	(Intercept)	-2.031	-2.051
Total Credit Hours	-0.200	-0.197	On Financial Suspension	1.732	1.729	MTSU Cumulative GPA	-0.890	-0.888
Parent Total Income	-0.020	-0.017	Percentage of Attended Courses With Grade W	0.053	0.053	Percentage of Attended Courses With Grade W	0.005	0.005
Taken Math 1530	-1.604	-1.605	On Financial Probation	1.103	1.103	On Financial Suspension	1.888	1.881
Taken Math 1710	0.614	0.618	High School GPA	-0.310	-0.325	On Financial Probation	1.079	1.075
Evening Student	0.314	0.303	Dollar Amount of Scholarships	-0.026	-0.018	Dollar Amount of Scholarships	-0.009	-0.006
Compass English	-0.033	-0.032	Percentage of Courses With Grade DFWN	0.007	0.007	Participated In High School Extracurriculars	-0.169	-0.167
Living On Campus	-0.144	-0.142	Percentage of Withdrawn Courses	-0.044	-0.044	Living On Campus	-0.058	-0.055
Dollar Amount of Institutional Aid	-0.038	-0.025	Percentage of Courses With Academic Difficulty	0.003	0.003	25 or Older	0.757	0.754
25 or Older	-0.198	-0.203	Dollar Amount of Loans	0.008	0.005			
Pursuing Second Bachelors	0.615	0.601	Living On Campus	-0.053	-0.052			
On Financial Probation	1.231	1.232	Parent Total Income	-0.019	-0.013			
On Financial Suspension	2.013	2.030	Has Unmet Need	0.136	0.139			
Percentage of Attended Courses With Grade W	0.028	0.028	Taken High School Algebra 2	0.231	0.237			
Percentage of Courses With Academic Difficulty	0.005	0.006						
Percentage of Medium Courses	-0.005	-0.006						
Percentage of Courses With Tenured Faculty	-0.002	-0.002						
Total Accuracy:	**0.766**	**0.735**	**Total Accuracy:**	**0.743**	**0.744**	**Total Accuracy:**	**0.760**	**0.761**

Table B.3: Random Forest Variable Importance For Disabled Students



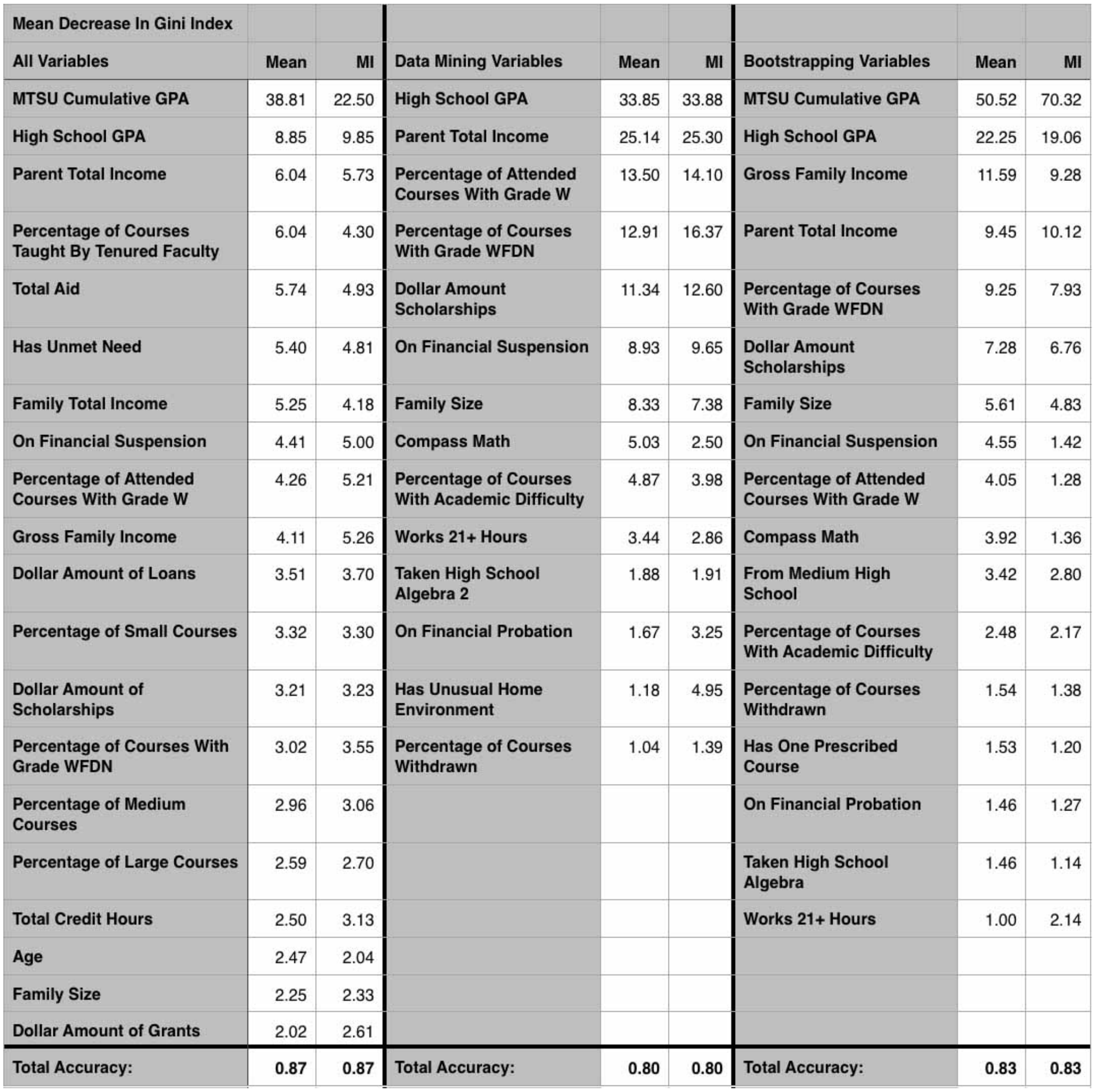
Mean Decrease In Gini Index								
All Variables	**Mean**	**MI**	**Data Mining Variables**	**Mean**	**MI**	**Bootstrapping Variables**	**Mean**	**MI**
MTSU Cumulative GPA	18.36	17.78	High School GPA	36.51	34.78	Cumulative MTSU GPA	31.03	29.61
High School GPA	4.78	4.48	Percentage of Courses With Grade DFWN	17.51	17.31	High School GPA	18.19	19.82
Compass English Level	4.69	4.17	Age	8.76	10.38	Percentage of Courses With Grade DFWN	10.31	8.58
Total Aid Received	4.01	4.58	Percentage of Attended Courses With Grade W	8.22	4.91	Percentage of Courses With Tenured Faculty	9.04	8.51
Percentage of Courses With Tenured Faculty	3.85	4.25	Taken English 1009	4.50	4.18	Undeclared	4.22	1.28
Has Unmet Need	3.50	3.57	Undeclared	3.73	2.96	Has 3 Accomplishments	2.69	1.63
Parent Total Income	2.77	3.01	On Financial Probation	3.54	4.07	Living On Campus	2.65	2.31
Percentage of Small Courses	2.58	2.40	Has Disability	2.67	2.28	On Financial Suspension	2.05	1.03
Percentage of Courses With Grade DFWN	2.36	2.56	On Financial Probation	2.26	1.59	Percentage of Attended Courses With Grade W	1.99	1.59
Percentage of Medium Courses	2.26	2.36	Percentage of Courses Withdrawn	2.01	2.93	Taken Advanced High School Math	1.81	2.16
Age	2.15	1.98				Participated in High School Extracurriculars	1.79	1.38
Gross Family Income	2.12	2.81				Taken English 1009	1.63	1.06
Family Size	2.09	1.41				On Financial Probation	1.35	0.73
Family Total Income	2.06	2.07				Percentage of Courses Withdrawn	1.29	1.30
Total Credit Hours	1.65	1.48						
Has Disability	1.61	1.75						
Dollar Amount of Scholarships	1.56	1.47						
Dollar Amount Of Loans	1.54	1.49						
Has Visual Impairment	1.47	1.32						
Percentage of Large Courses	1.28	1.21						
Needs Help Study Skills	1.19	1.11						
Total Accuracy:	**0.70**	**0.69**	Total Accuracy:	**0.63**	**0.64**	Total Accuracy:	**0.66**	**0.67**

Table B.4: Logistic Regression Coefficients For Disabled Students



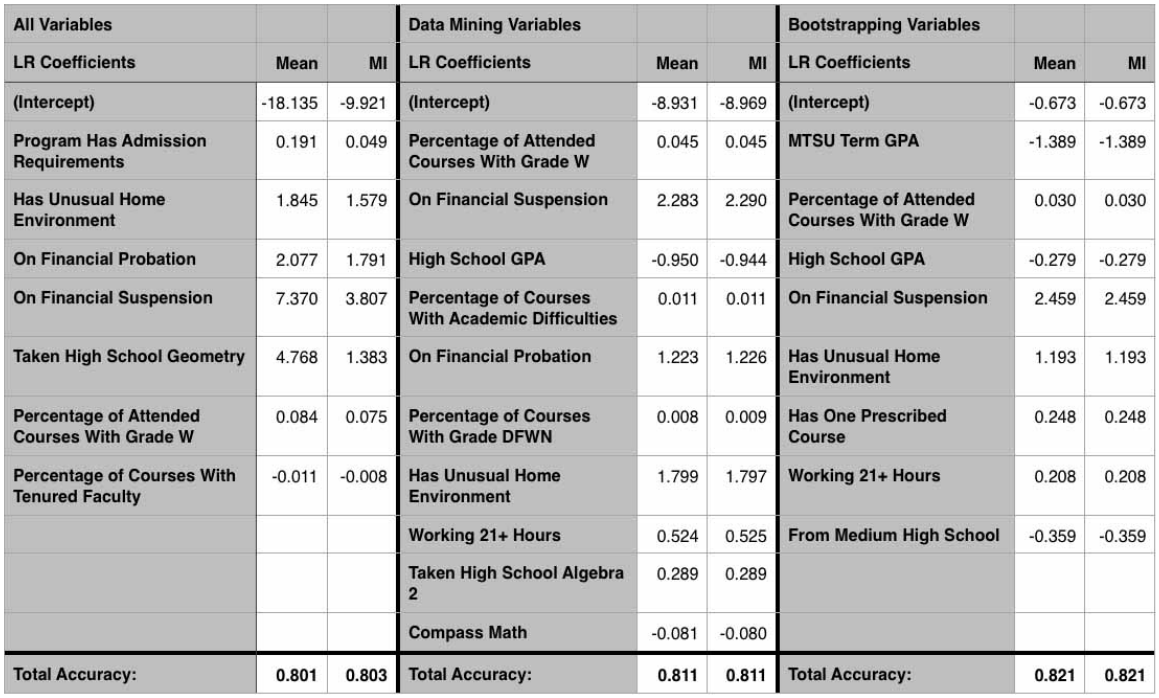
All Variables			Data Mining Variables			Bootstrapping Variables		
LR Coefficients	Mean	MI	LR Coefficients	Mean	MI	LR Coefficients	Mean	MI
(Intercept)	0.142	0.542	(Intercept)	-1.121	-1.090	(Intercept)	6.113	6.167
Taken Math 1710	0.124	0.000	Percentage of Attended Courses With Grade W	0.008	0.008	MTSU Cumulative GPA	-0.918	-0.894
Honors Student	-2.214	-1.841	On Financial Probation	2.015	2.022	On Financial Probation	1.206	1.219
Has Prior Associates	-1.570	-1.507	On Financial Suspension	3.235	3.237	On Financial Suspension	0.511	0.510
Has Visual Impairmnet	0.939	0.772	High School GPA	-0.146	-0.126	Honors Student	-2.167	-2.219
Living On Campus	-0.552	-0.576	Honors Student	-1.230	-1.245	Undeclared Major	0.541	0.527
Undeclared Major	0.407	0.239	Age	-0.307	-0.305	Has Prior Associates	-2.804	-2.822
On Financial Probation	1.588	1.534	Has Prior Associates	-0.759	-0.770	Has 3 Or More Accomplishments	-0.438	-0.456
On Financial Suspension	2.507	2.355	Has Disability	1.352	1.354	Percentage of Courses With Tenured Faculty	-0.020	-0.020
Needs Help All Aread	-0.437	-0.639	Percentage of Course With Grade DFWN	0.009	0.009	Living On Campus	-0.660	-0.671
High School College Prep	-0.800	0.000	Dollar Amount of Work	0.057	0.035			
Percentage of Courses With Tenured Faculty	-0.003	-0.002	Taken English 1009	0.462	0.457			
Total Accuracy:	**0.689**	**0.694**	Total Accuracy:	**0.633**	**0.638**	Total Accuracy:	**0.725**	**0.719**

Table B.5: Random Forest Variable Importance For First Generation Students



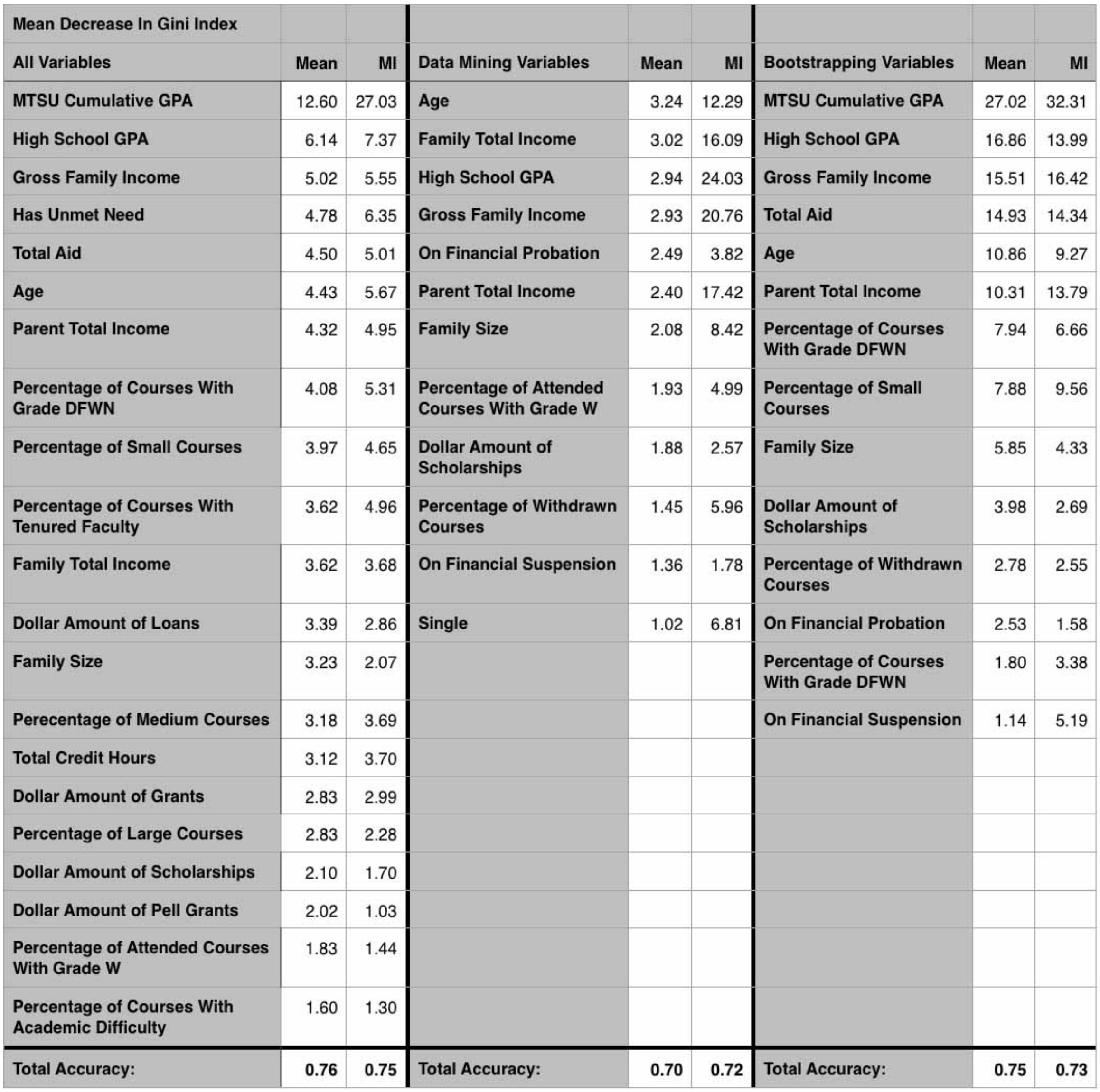
Mean Decrease In Gini Index								
All Variables	**Mean**	**MI**	**Data Mining Variables**	**Mean**	**MI**	**Bootstrapping Variables**	**Mean**	**MI**
MTSU Cumulative GPA	38.81	22.50	High School GPA	33.85	33.88	MTSU Cumulative GPA	50.52	70.32
High School GPA	8.85	9.85	Parent Total Income	25.14	25.30	High School GPA	22.25	19.06
Parent Total Income	6.04	5.73	Percentage of Attended Courses With Grade W	13.50	14.10	Gross Family Income	11.59	9.28
Percentage of Courses Taught By Tenured Faculty	6.04	4.30	Percentage of Courses With Grade WFDN	12.91	16.37	Parent Total Income	9.45	10.12
Total Aid	5.74	4.93	Dollar Amount Scholarships	11.34	12.60	Percentage of Courses With Grade WFDN	9.25	7.93
Has Unmet Need	5.40	4.81	On Financial Suspension	8.93	9.65	Dollar Amount Scholarships	7.28	6.76
Family Total Income	5.25	4.18	Family Size	8.33	7.38	Family Size	5.61	4.83
On Financial Suspension	4.41	5.00	Compass Math	5.03	2.50	On Financial Suspension	4.55	1.42
Percentage of Attended Courses With Grade W	4.26	5.21	Percentage of Courses With Academic Difficulty	4.87	3.98	Percentage of Attended Courses With Grade W	4.05	1.28
Gross Family Income	4.11	5.26	Works 21+ Hours	3.44	2.86	Compass Math	3.92	1.36
Dollar Amount of Loans	3.51	3.70	Taken High School Algebra 2	1.88	1.91	From Medium High School	3.42	2.80
Percentage of Small Courses	3.32	3.30	On Financial Probation	1.67	3.25	Percentage of Courses With Academic Difficulty	2.48	2.17
Dollar Amount of Scholarships	3.21	3.23	Has Unusual Home Environment	1.18	4.95	Percentage of Courses Withdrawn	1.54	1.38
Percentage of Courses With Grade WFDN	3.02	3.55	Percentage of Courses Withdrawn	1.04	1.39	Has One Prescribed Course	1.53	1.20
Percentage of Medium Courses	2.96	3.06				On Financial Probation	1.46	1.27
Percentage of Large Courses	2.59	2.70				Taken High School Algebra	1.46	1.14
Total Credit Hours	2.50	3.13				Works 21+ Hours	1.00	2.14
Age	2.47	2.04						
Family Size	2.25	2.33						
Dollar Amount of Grants	2.02	2.61						
Total Accuracy:	**0.87**	**0.87**	Total Accuracy:	**0.80**	**0.80**	Total Accuracy:	**0.83**	**0.83**

Table B.6: Logistic Regression Coefficients For First Generation Students



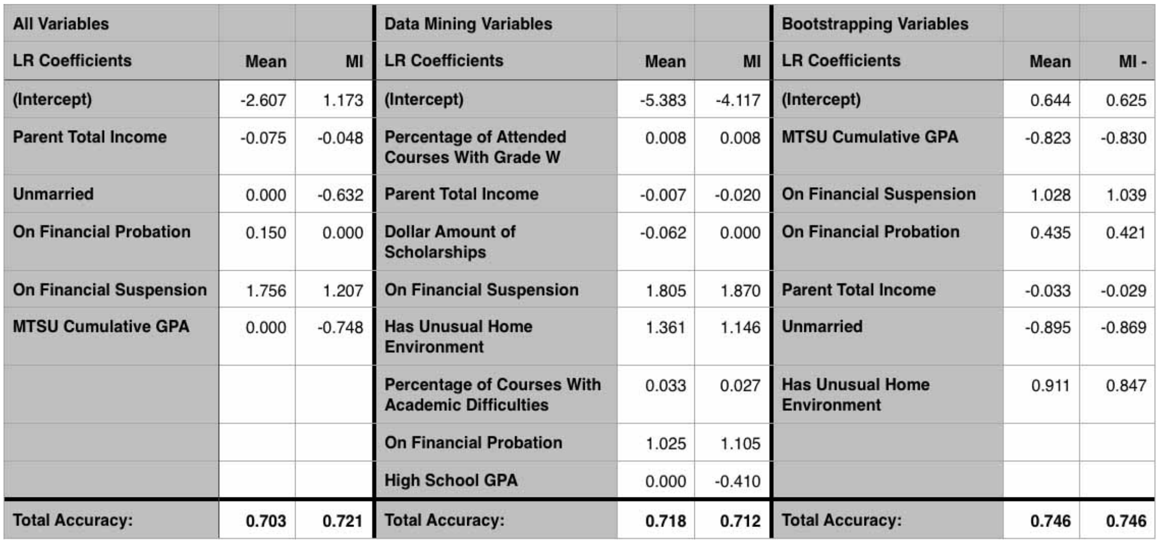
All Variables			Data Mining Variables			Bootstrapping Variables		
LR Coefficients	Mean	MI	LR Coefficients	Mean	MI	LR Coefficients	Mean	MI
(Intercept)	-18.135	-9.921	(Intercept)	-8.931	-8.969	(Intercept)	-0.673	-0.673
Program Has Admission Requirements	0.191	0.049	Percentage of Attended Courses With Grade W	0.045	0.045	MTSU Term GPA	-1.389	-1.389
Has Unusual Home Environment	1.845	1.579	On Financial Suspension	2.283	2.290	Percentage of Attended Courses With Grade W	0.030	0.030
On Financial Probation	2.077	1.791	High School GPA	-0.950	-0.944	High School GPA	-0.279	-0.279
On Financial Suspension	7.370	3.807	Percentage of Courses With Academic Difficulties	0.011	0.011	On Financial Suspension	2.459	2.459
Taken High School Geometry	4.768	1.383	On Financial Probation	1.223	1.226	Has Unusual Home Environment	1.193	1.193
Percentage of Attended Courses With Grade W	0.084	0.075	Percentage of Courses With Grade DFWN	0.008	0.009	Has One Prescribed Course	0.248	0.248
Percentage of Courses With Tenured Faculty	-0.011	-0.008	Has Unusual Home Environment	1.799	1.797	Working 21+ Hours	0.208	0.208
			Working 21+ Hours	0.524	0.525	From Medium High School	-0.359	-0.359
			Taken High School Algebra 2	0.289	0.289			
			Compass Math	-0.081	-0.080			
Total Accuracy:	0.801	0.803	Total Accuracy:	0.811	0.811	Total Accuracy:	0.821	0.821

Table B.7: Random Forest Variable Importance For Hispanic Students



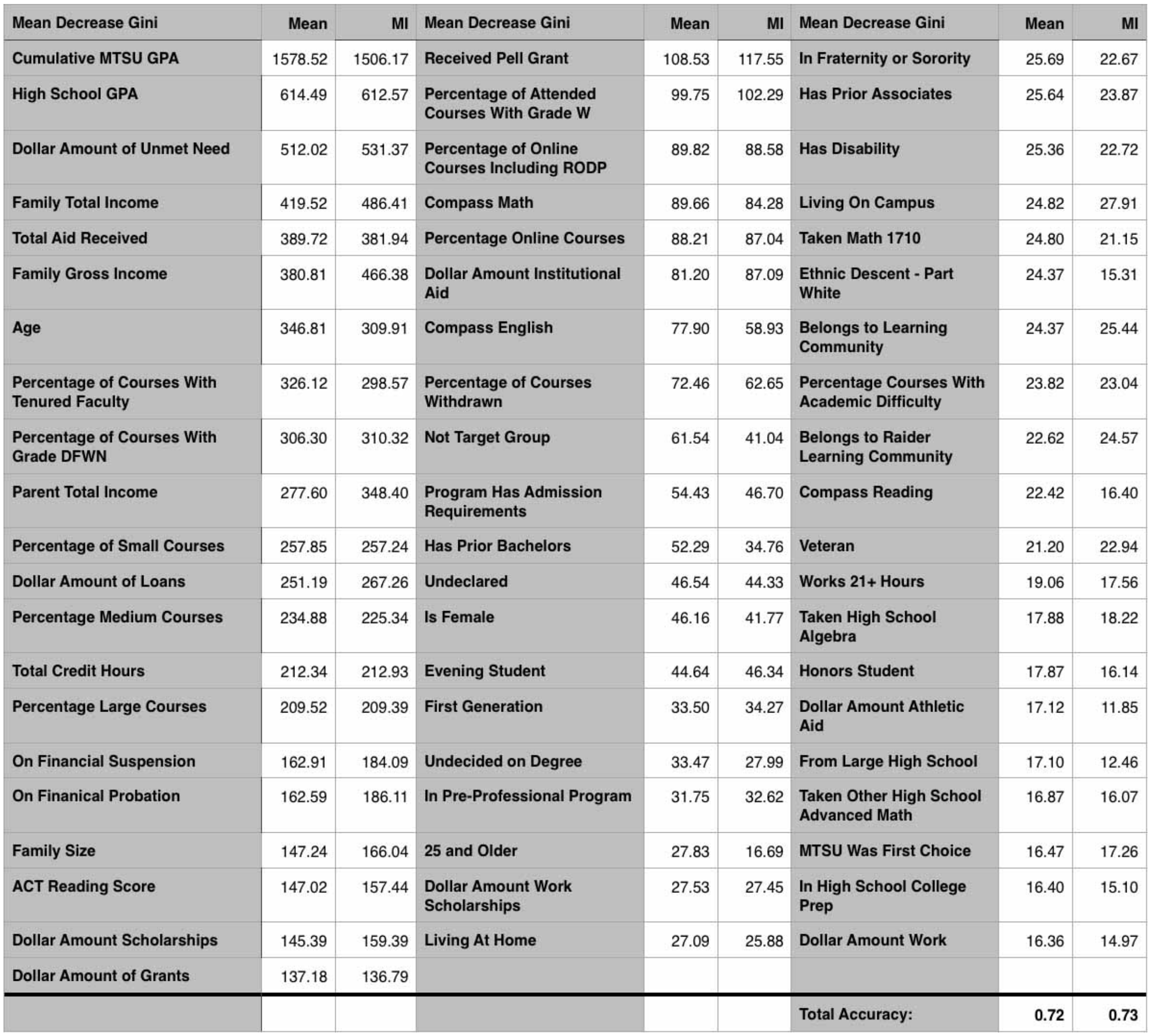
Mean Decrease In Gini Index								
All Variables	Mean	MI	Data Mining Variables	Mean	MI	Bootstrapping Variables	Mean	MI
MTSU Cumulative GPA	12.60	27.03	Age	3.24	12.29	MTSU Cumulative GPA	27.02	32.31
High School GPA	6.14	7.37	Family Total Income	3.02	16.09	High School GPA	16.86	13.99
Gross Family Income	5.02	5.55	High School GPA	2.94	24.03	Gross Family Income	15.51	16.42
Has Unmet Need	4.78	6.35	Gross Family Income	2.93	20.76	Total Aid	14.93	14.34
Total Aid	4.50	5.01	On Financial Probation	2.49	3.82	Age	10.86	9.27
Age	4.43	5.67	Parent Total Income	2.40	17.42	Parent Total Income	10.31	13.79
Parent Total Income	4.32	4.95	Family Size	2.08	8.42	Percentage of Courses With Grade DFWN	7.94	6.66
Percentage of Courses With Grade DFWN	4.08	5.31	Percentage of Attended Courses With Grade W	1.93	4.99	Percentage of Small Courses	7.88	9.56
Percentage of Small Courses	3.97	4.65	Dollar Amount of Scholarships	1.88	2.57	Family Size	5.85	4.33
Percentage of Courses With Tenured Faculty	3.62	4.96	Percentage of Withdrawn Courses	1.45	5.96	Dollar Amount of Scholarships	3.98	2.69
Family Total Income	3.62	3.68	On Financial Suspension	1.36	1.78	Percentage of Withdrawn Courses	2.78	2.55
Dollar Amount of Loans	3.39	2.86	Single	1.02	6.81	On Financial Probation	2.53	1.58
Family Size	3.23	2.07				Percentage of Courses With Grade DFWN	1.80	3.38
Perecentage of Medium Courses	3.18	3.69				On Financial Suspension	1.14	5.19
Total Credit Hours	3.12	3.70						
Dollar Amount of Grants	2.83	2.99						
Percentage of Large Courses	2.83	2.28						
Dollar Amount of Scholarships	2.10	1.70						
Dollar Amount of Pell Grants	2.02	1.03						
Percentage of Attended Courses With Grade W	1.83	1.44						
Percentage of Courses With Academic Difficulty	1.60	1.30						
Total Accuracy:	0.76	0.75	Total Accuracy:	0.70	0.72	Total Accuracy:	0.75	0.73

Table B.8: Logistic Regression Coefficients For Hispanic Students



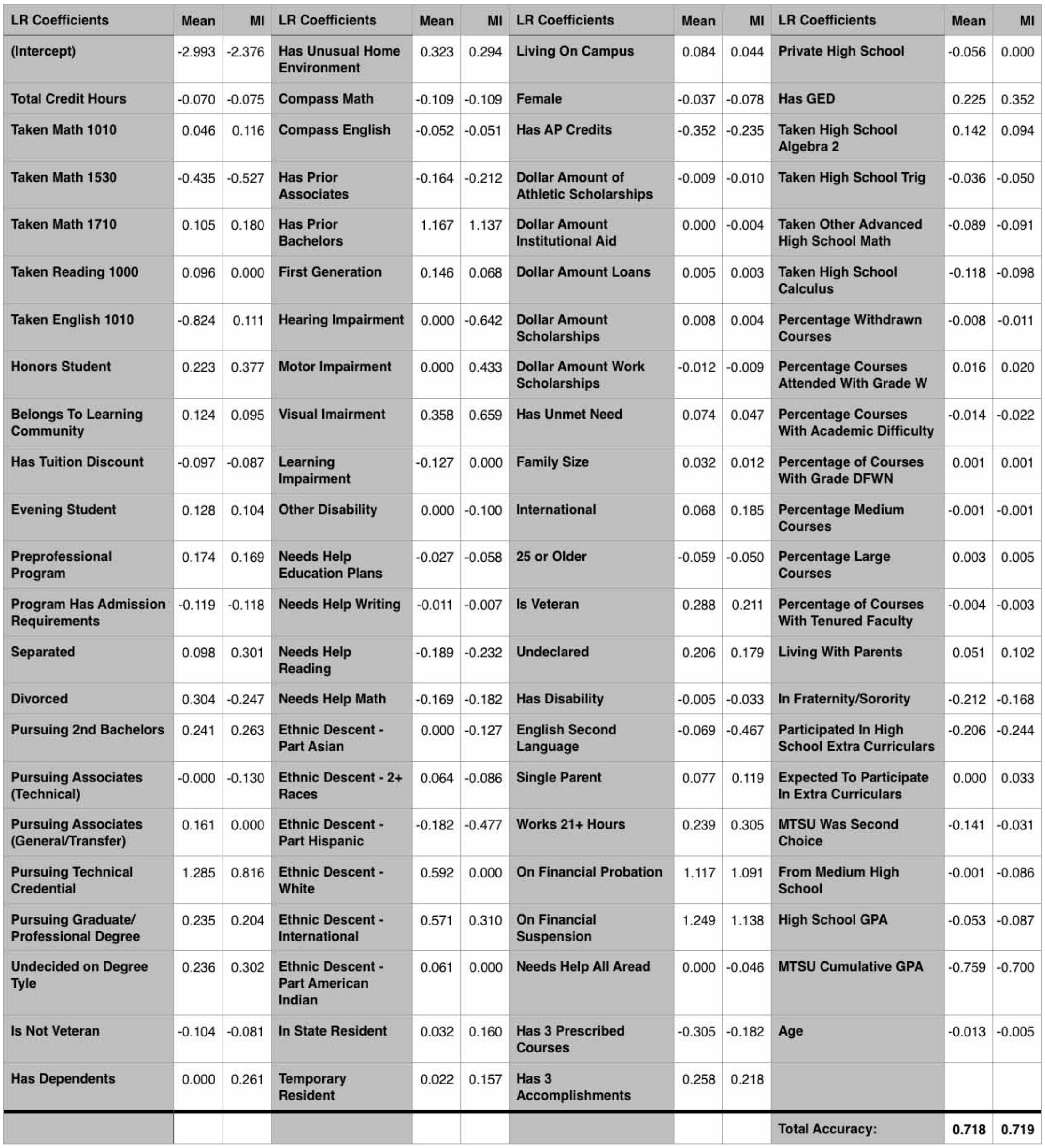
All Variables			Data Mining Variables			Bootstrapping Variables		
LR Coefficients	Mean	MI	LR Coefficients	Mean	MI	LR Coefficients	Mean	MI -
(Intercept)	-2.607	1.173	(Intercept)	-5.383	-4.117	(Intercept)	0.644	0.625
Parent Total Income	-0.075	-0.048	Percentage of Attended Courses With Grade W	0.008	0.008	MTSU Cumulative GPA	-0.823	-0.830
Unmarried	0.000	-0.632	Parent Total Income	-0.007	-0.020	On Financial Suspension	1.028	1.039
On Financial Probation	0.150	0.000	Dollar Amount of Scholarships	-0.062	0.000	On Financial Probation	0.435	0.421
On Financial Suspension	1.756	1.207	On Financial Suspension	1.805	1.870	Parent Total Income	-0.033	-0.029
MTSU Cumulative GPA	0.000	-0.748	Has Unusual Home Environment	1.361	1.146	Unmarried	-0.895	-0.869
			Percentage of Courses With Academic Difficulties	0.033	0.027	Has Unusual Home Environment	0.911	0.847
			On Financial Probation	1.025	1.105			
			High School GPA	0.000	-0.410			
Total Accuracy:	**0.703**	**0.721**	Total Accuracy:	**0.718**	**0.712**	Total Accuracy:	**0.746**	**0.746**

Table B.9: Random Forest Variable Importance For All Students



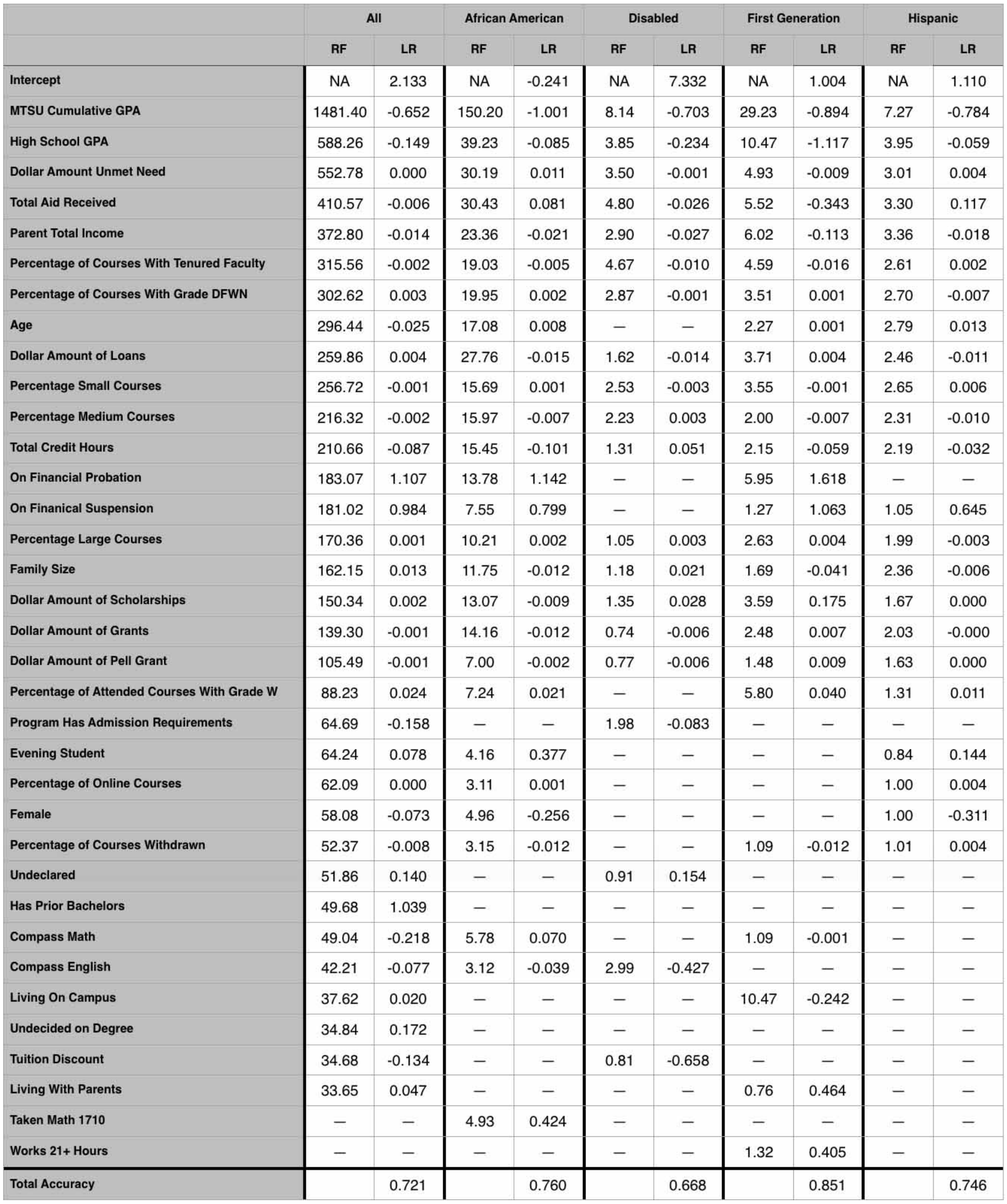
Mean Decrease Gini	Mean	MI	Mean Decrease Gini	Mean	MI	Mean Decrease Gini	Mean	MI
Cumulative MTSU GPA	1578.52	1506.17	Received Pell Grant	108.53	117.55	In Fraternity or Sorority	25.69	22.67
High School GPA	614.49	612.57	Percentage of Attended Courses With Grade W	99.75	102.29	Has Prior Associates	25.64	23.87
Dollar Amount of Unmet Need	512.02	531.37	Percentage of Online Courses Including RODP	89.82	88.58	Has Disability	25.36	22.72
Family Total Income	419.52	486.41	Compass Math	89.66	84.28	Living On Campus	24.82	27.91
Total Aid Received	389.72	381.94	Percentage Online Courses	88.21	87.04	Taken Math 1710	24.80	21.15
Family Gross Income	380.81	466.38	Dollar Amount Institutional Aid	81.20	87.09	Ethnic Descent - Part White	24.37	15.31
Age	346.81	309.91	Compass English	77.90	58.93	Belongs to Learning Community	24.37	25.44
Percentage of Courses With Tenured Faculty	326.12	298.57	Percentage of Courses Withdrawn	72.46	62.65	Percentage Courses With Academic Difficulty	23.82	23.04
Percentage of Courses With Grade DFWN	306.30	310.32	Not Target Group	61.54	41.04	Belongs to Raider Learning Community	22.62	24.57
Parent Total Income	277.60	348.40	Program Has Admission Requirements	54.43	46.70	Compass Reading	22.42	16.40
Percentage of Small Courses	257.85	257.24	Has Prior Bachelors	52.29	34.76	Veteran	21.20	22.94
Dollar Amount of Loans	251.19	267.26	Undeclared	46.54	44.33	Works 21+ Hours	19.06	17.56
Percentage Medium Courses	234.88	225.34	Is Female	46.16	41.77	Taken High School Algebra	17.88	18.22
Total Credit Hours	212.34	212.93	Evening Student	44.64	46.34	Honors Student	17.87	16.14
Percentage Large Courses	209.52	209.39	First Generation	33.50	34.27	Dollar Amount Athletic Aid	17.12	11.85
On Financial Suspension	162.91	184.09	Undecided on Degree	33.47	27.99	From Large High School	17.10	12.46
On Finanical Probation	162.59	186.11	In Pre-Professional Program	31.75	32.62	Taken Other High School Advanced Math	16.87	16.07
Family Size	147.24	166.04	25 and Older	27.83	16.69	MTSU Was First Choice	16.47	17.26
ACT Reading Score	147.02	157.44	Dollar Amount Work Scholarships	27.53	27.45	In High School College Prep	16.40	15.10
Dollar Amount Scholarships	145.39	159.39	Living At Home	27.09	25.88	Dollar Amount Work	16.36	14.97
Dollar Amount of Grants	137.18	136.79						
						Total Accuracy:	0.72	0.73

Table B.10: Logistic Regression Coefficients For All Students



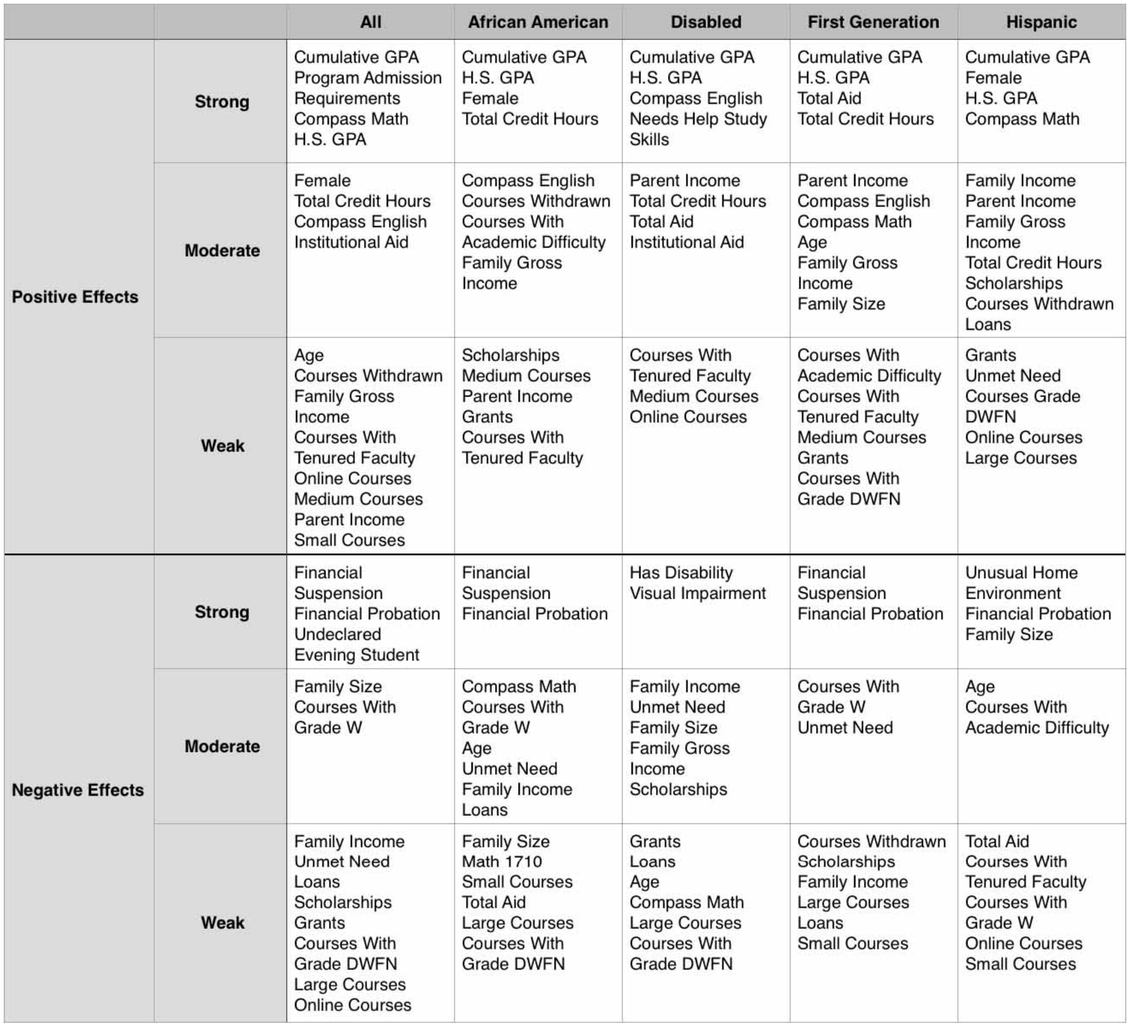
LR Coefficients	Mean	MI	LR Coefficients	Mean	MI	LR Coefficients	Mean	MI	LR Coefficients	Mean	MI
(Intercept)	-2.993	-2.376	Has Unusual Home Environment	0.323	0.294	Living On Campus	0.084	0.044	Private High School	-0.056	0.000
Total Credit Hours	-0.070	-0.075	Compass Math	-0.109	-0.109	Female	-0.037	-0.078	Has GED	0.225	0.352
Taken Math 1010	0.046	0.116	Compass English	-0.052	-0.051	Has AP Credits	-0.352	-0.235	Taken High School Algebra 2	0.142	0.094
Taken Math 1530	-0.435	-0.527	Has Prior Associates	-0.164	-0.212	Dollar Amount of Athletic Scholarships	-0.009	-0.010	Taken High School Trig	-0.036	-0.050
Taken Math 1710	0.105	0.180	Has Prior Bachelors	1.167	1.137	Dollar Amount Institutional Aid	0.000	-0.004	Taken Other Advanced High School Math	-0.089	-0.091
Taken Reading 1000	0.096	0.000	First Generation	0.146	0.068	Dollar Amount Loans	0.005	0.003	Taken High School Calculus	-0.118	-0.098
Taken English 1010	-0.824	0.111	Hearing Impairment	0.000	-0.642	Dollar Amount Scholarships	0.008	0.004	Percentage Withdrawn Courses	-0.008	-0.011
Honors Student	0.223	0.377	Motor Impairment	0.000	0.433	Dollar Amount Work Scholarships	-0.012	-0.009	Percentage Courses Attended With Grade W	0.016	0.020
Belongs To Learning Community	0.124	0.095	Visual Imairment	0.358	0.659	Has Unmet Need	0.074	0.047	Percentage Courses With Academic Difficulty	-0.014	-0.022
Has Tuition Discount	-0.097	-0.087	Learning Impairment	-0.127	0.000	Family Size	0.032	0.012	Percentage of Courses With Grade DFWN	0.001	0.001
Evening Student	0.128	0.104	Other Disability	0.000	-0.100	International	0.068	0.185	Percentage Medium Courses	-0.001	-0.001
Preprofessional Program	0.174	0.169	Needs Help Education Plans	-0.027	-0.058	25 or Older	-0.059	-0.050	Percentage Large Courses	0.003	0.005
Program Has Admission Requirements	-0.119	-0.118	Needs Help Writing	-0.011	-0.007	Is Veteran	0.288	0.211	Percentage of Courses With Tenured Faculty	-0.004	-0.003
Separated	0.098	0.301	Needs Help Reading	-0.189	-0.232	Undeclared	0.206	0.179	Living With Parents	0.051	0.102
Divorced	0.304	-0.247	Needs Help Math	-0.169	-0.182	Has Disability	-0.005	-0.033	In Fraternity/Sorority	-0.212	-0.168
Pursuing 2nd Bachelors	0.241	0.263	Ethnic Descent - Part Asian	0.000	-0.127	English Second Language	-0.069	-0.467	Participated In High School Extra Curriculars	-0.206	-0.244
Pursuing Associates (Technical)	-0.000	-0.130	Ethnic Descent - 2+ Races	0.064	-0.086	Single Parent	0.077	0.119	Expected To Participate In Extra Curriculars	0.000	0.033
Pursuing Associates (General/Transfer)	0.161	0.000	Ethnic Descent - Part Hispanic	-0.182	-0.477	Works 21+ Hours	0.239	0.305	MTSU Was Second Choice	-0.141	-0.031
Pursuing Technical Credential	1.285	0.816	Ethnic Descent - White	0.592	0.000	On Financial Probation	1.117	1.091	From Medium High School	-0.001	-0.086
Pursuing Graduate/ Professional Degree	0.235	0.204	Ethnic Descent - International	0.571	0.310	On Financial Suspension	1.249	1.138	High School GPA	-0.053	-0.087
Undecided on Degree Tyle	0.236	0.302	Ethnic Descent - Part American Indian	0.061	0.000	Needs Help All Aread	0.000	-0.046	MTSU Cumulative GPA	-0.759	-0.700
Is Not Veteran	-0.104	-0.081	In State Resident	0.032	0.160	Has 3 Prescribed Courses	-0.305	-0.182	Age	-0.013	-0.005
Has Dependents	0.000	0.261	Temporary Resident	0.022	0.157	Has 3 Accomplishments	0.258	0.218			
									Total Accuracy:	0.718	0.719

Table B.11: Final Model - Logistic Regression With Random Forest Selection



	All		African American		Disabled		First Generation		Hispanic	
	RF	LR	RF	LR	RF	LR	RF	LR	RF	LR
Intercept	NA	2.133	NA	-0.241	NA	7.332	NA	1.004	NA	1.110
MTSU Cumulative GPA	1481.40	-0.652	150.20	-1.001	8.14	-0.703	29.23	-0.894	7.27	-0.784
High School GPA	588.26	-0.149	39.23	-0.085	3.85	-0.234	10.47	-1.117	3.95	-0.059
Dollar Amount Unmet Need	552.78	0.000	30.19	0.011	3.50	-0.001	4.93	-0.009	3.01	0.004
Total Aid Received	410.57	-0.006	30.43	0.081	4.80	-0.026	5.52	-0.343	3.30	0.117
Parent Total Income	372.80	-0.014	23.36	-0.021	2.90	-0.027	6.02	-0.113	3.36	-0.018
Percentage of Courses With Tenured Faculty	315.56	-0.002	19.03	-0.005	4.67	-0.010	4.59	-0.016	2.61	0.002
Percentage of Courses With Grade DFWN	302.62	0.003	19.95	0.002	2.87	-0.001	3.51	0.001	2.70	-0.007
Age	296.44	-0.025	17.08	0.008	—	—	2.27	0.001	2.79	0.013
Dollar Amount of Loans	259.86	0.004	27.76	-0.015	1.62	-0.014	3.71	0.004	2.46	-0.011
Percentage Small Courses	256.72	-0.001	15.69	0.001	2.53	-0.003	3.55	-0.001	2.65	0.006
Percentage Medium Courses	216.32	-0.002	15.97	-0.007	2.23	0.003	2.00	-0.007	2.31	-0.010
Total Credit Hours	210.66	-0.087	15.45	-0.101	1.31	0.051	2.15	-0.059	2.19	-0.032
On Financial Probation	183.07	1.107	13.78	1.142	—	—	5.95	1.618	—	—
On Finanical Suspension	181.02	0.984	7.55	0.799	—	—	1.27	1.063	1.05	0.645
Percentage Large Courses	170.36	0.001	10.21	0.002	1.05	0.003	2.63	0.004	1.99	-0.003
Family Size	162.15	0.013	11.75	-0.012	1.18	0.021	1.69	-0.041	2.36	-0.006
Dollar Amount of Scholarships	150.34	0.002	13.07	-0.009	1.35	0.028	3.59	0.175	1.67	0.000
Dollar Amount of Grants	139.30	-0.001	14.16	-0.012	0.74	-0.006	2.48	0.007	2.03	-0.000
Dollar Amount of Pell Grant	105.49	-0.001	7.00	-0.002	0.77	-0.006	1.48	0.009	1.63	0.000
Percentage of Attended Courses With Grade W	88.23	0.024	7.24	0.021	—	—	5.80	0.040	1.31	0.011
Program Has Admission Requirements	64.69	-0.158	—	—	1.98	-0.083	—	—	—	—
Evening Student	64.24	0.078	4.16	0.377	—	—	—	—	0.84	0.144
Percentage of Online Courses	62.09	0.000	3.11	0.001	—	—	—	—	1.00	0.004
Female	58.08	-0.073	4.96	-0.256	—	—	—	—	1.00	-0.311
Percentage of Courses Withdrawn	52.37	-0.008	3.15	-0.012	—	—	1.09	-0.012	1.01	0.004
Undeclared	51.86	0.140	—	—	0.91	0.154	—	—	—	—
Has Prior Bachelors	49.68	1.039	—	—	—	—	—	—	—	—
Compass Math	49.04	-0.218	5.78	0.070	—	—	1.09	-0.001	—	—
Compass English	42.21	-0.077	3.12	-0.039	2.99	-0.427	—	—	—	—
Living On Campus	37.62	0.020	—	—	—	—	10.47	-0.242	—	—
Undecided on Degree	34.84	0.172	—	—	—	—	—	—	—	—
Tuition Discount	34.68	-0.134	—	—	0.81	-0.658	—	—	—	—
Living With Parents	33.65	0.047	—	—	—	—	0.76	0.464	—	—
Taken Math 1710	—	—	4.93	0.424	—	—	—	—	—	—
Works 21+ Hours	—	—	—	—	—	—	1.32	0.405	—	—
Total Accuracy		0.721		0.760		0.668		0.851		0.746

Table B.12: Final Model - Variable Effects Sorted By Weights



		All	African American	Disabled	First Generation	Hispanic
Positive Effects	Strong	Cumulative GPA Program Admission Requirements Compass Math H.S. GPA	Cumulative GPA H.S. GPA Female Total Credit Hours	Cumulative GPA H.S. GPA Compass English Needs Help Study Skills	Cumulative GPA H.S. GPA Total Aid Total Credit Hours	Cumulative GPA Female H.S. GPA Compass Math
	Moderate	Female Total Credit Hours Compass English Institutional Aid	Compass English Courses Withdrawn Courses With Academic Difficulty Family Gross Income	Parent Income Total Credit Hours Total Aid Institutional Aid	Parent Income Compass English Compass Math Age Family Gross Income Family Size	Family Income Parent Income Family Gross Income Total Credit Hours Scholarships Courses Withdrawn Loans
	Weak	Age Courses Withdrawn Family Gross Income Courses With Tenured Faculty Online Courses Medium Courses Parent Income Small Courses	Scholarships Medium Courses Parent Income Grants Courses With Tenured Faculty	Courses With Tenured Faculty Medium Courses Online Courses	Courses With Academic Difficulty Courses With Tenured Faculty Medium Courses Grants Courses With Grade DWFN	Grants Unmet Need Courses Grade DWFN Online Courses Large Courses
Negative Effects	Strong	Financial Suspension Financial Probation Undeclared Evening Student	Financial Suspension Financial Probation	Has Disability Visual Impairment	Financial Suspension Financial Probation	Unusual Home Environment Financial Probation Family Size
	Moderate	Family Size Courses With Grade W	Compass Math Courses With Grade W Age Unmet Need Family Income Loans	Family Income Unmet Need Family Size Family Gross Income Scholarships	Courses With Grade W Unmet Need	Age Courses With Academic Difficulty
	Weak	Family Income Unmet Need Loans Scholarships Grants Courses With Grade DWFN Large Courses Online Courses	Family Size Math 1710 Small Courses Total Aid Large Courses Courses With Grade DWFN	Grants Loans Age Compass Math Large Courses Courses With Grade DWFN	Courses Withdrawn Scholarships Family Income Large Courses Loans Small Courses	Total Aid Courses With Tenured Faculty Courses With Grade W Online Courses Small Courses