

# Mining Educational Data to Create a Model to Predict Student Retention

by  
Kailey McDonald

A thesis presented to the Honors College of Middle Tennessee State University in partial fulfillment of the requirements for graduation from the University Honors College

Fall 2015

Mining Educational Data to Create a Model to Predict Student Retention

by  
Kailey McDonald

APPROVED:

---

Dr. Cen Li  
Project Advisor  
Computer Science Department

---

Dr. Chrisila Pettey  
Computer Science Department Chair

---

Dr. John Pennington  
Psychology Department  
Honors Council Representative

---

Dr. Philip E. Phillips, Associate Dean  
University Honors College

## **Abstract**

Student retention is a widespread issue in higher education. This study applies data mining methods to analyze student data at MTSU, specifically focusing on minority student groups. The hope of this study is to determine a set of attributes that are highly predictive of student retention and to develop models to predict the retention status of future students within these target groups. Decision tree classification models will be created for each target minority group in order to predict the retention status.

We found that a student's GPA and financial factors are the most predictive on the student's retention status. The developed models were able to successfully classify students at rates as high as 68.7%. We hope that this data can help to provide the university with a way to identify students with a high risk of failing to remain enrolled and to improve retention rates of minority students.

# Contents

<b>List of Figures</b>	<b>iv</b>
<b>List of Tables</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Importance of Studying Student Retention .....	1
1.2 Data Mining .....	2
1.3 Classification Approach .....	3
1.4 Common Classification Techniques.....	5
1.5 Previous Studies of Student Retention .....	7
<b>2 Methodology</b>	<b>9</b>
2.1 Overview .....	9
2.2 General Explanation of Data Set.....	10
2.3 Initial Analysis of Data .....	11
2.4 Attribute Selection.....	12
2.5 Creation of the Decision Tree Model.....	13
2.6 Validation of the Model .....	14
<b>3 Experimental Results</b>	<b>16</b>
3.1 Initial Analysis .....	16
3.2 Attribute Selection.....	24
3.3 Decision Tree Model.....	30
3.4 Validation .....	35
3.4.1 Cross-Validation of Existing Models.....	35
3.4.2 Creating a New Model .....	42
<b>4 Discussion</b>	<b>44</b>
4.1 MTSU's Current Initiatives.....	44
4.2 Conclusions .....	45
<b>Appendices</b>	<b>48</b>
<b>Bibliography</b>	<b>69</b>

# List of Figures

1.	Visualization of model creation & prediction for classification approach.....	5
2.	Outline of the steps to create a predictive model for student retention.....	9
3.	Overview of the results of retention classifications for each minority group ....	17
4.	Portion of Decision Tree Model for African American Students .....	32
5.	Portion of Decision Tree Model for first generation Students .....	33
6.	Portion of Decision Tree Model for disabled Students .....	34
7.	Portion of Decision Tree Model for Hispanic Students .....	35
8.	Overview of accuracy for each minority group model .....	36
9.	Detailed accuracy for the African American student model .....	38
10.	Detailed accuracy for the first generation student model.....	39
11.	Detailed accuracy for the disabled student model.....	40
12.	Detailed accuracy for the Hispanic student model.....	42

# List of Tables

1. Count and Percentage of all freshman students' status .....	18
2. Count and Percentage of all freshman students' status with GPA $\geq 2.0$ .....	19
3. Count and Percentage of all freshman students' status with GPA $< 2.0$ .....	20
4. Count and Percentage of all freshman students' status with GPA $\geq 2.4$ .....	21
5. Count and Percentage of all freshman students' status with GPA $< 2.4$ .....	22
6. Count and Percentage of all freshman students' status with GPA $\geq 2.75$ .....	23
7. Count and Percentage of all freshman students' status with GPA $< 2.75$ .....	24
8. Selected Attributes in descending order.....	27
9. Selected Attributes in alphabetical order .....	29
10. Performance of the decision tree model for the African American group .....	37
11. Results of the decision tree model for the African American group.....	37
12. Performance of the decision tree model for the first generation group.....	39
13. Results of the decision tree model for the first generation group .....	39
14. Performance of the decision tree model for the disabled group.....	40
15. Results of the decision tree model for the disabled group .....	40
16. Performance of the decision tree model for the Hispanic group.....	41
17. Results of the decision tree model for the Hispanic group .....	41
18. Performance of the decision tree model for the reduced Hispanic group .....	43
19. Results of the decision tree model for the reduced Hispanic group.....	43

# Chapter 1

## Introduction

### 1.1 Importance of Studying Student Retention

Student retention is a dynamic obstacle plaguing universities across the country, and Middle Tennessee State University is not estranged from this issue. According to Federal Student Aid (FAFSA), a university's retention rate is defined as the percentage of first-time, first-year undergraduate students who continue at the school for the next year (2010). In the United States, the national retention rate for public universities is about 79% (National Center for Education Statistics, 2014). MTSU is slightly below this average, retaining about 70% of our students after their initial year of enrollment (Middle Tennessee State University, 2013). There are many motives behind analyzing our school's retention rates. The first is to help our students earn their respective degrees and be successful. It is additionally important to investigate the causes behind student retention because of performance-based funding that is critical to the budget of our institution. Failing to retain students and keep them on track to graduation could cause this funding to decline.

This study is subset of on-going research by Dr. Cen Li, Dr. Qiang Wu, Dr. John Wallin, and Dr. Michael Hein. They have been studying retention of students at MTSU and completing different analyses of the acquired educational data that I will also be

using in this study. By evaluating student data and finding an accurate model for identifying students at risk for attrition, we can improve our process for interventions and help these students before it is too late.

## **1.2 Data Mining**

The purpose of data mining is to uncover interesting data patterns hidden in large sets of data (Han & Kamber, 2006). This field has grown rapidly and gained much interest in the recent years. Businesses and institutions have collected and stored huge amounts of data over the years. However, this data does no good if it is merely stored away in a database. This scenario has been described as a “data rich, but information poor” situation (Slotnik & Orland, 2010). There is an imminent need to turn this data into useful knowledge – hence the growth of data mining.

There are many different approaches to data mining with specific purposes for each approach. The goal of mining data can be to find interesting patterns or classes (cluster analysis), detect unusual instances in the data (anomaly detection), analyze association relations among data, or find patterns that can be used to make predictions about specific types of data (classification) (Han & Kamber, 2006). All methods involve finding interesting patterns in data sets but use this information in unique ways.

Cluster analysis is used to generate class labels from a data set that is not already divided into specific groups. This approach will begin with information about a set of objects and will result in the organization of this data set into relatively homogenous groups (Aldenderfer & Blachfield, 1984). An example of when this method would be

used is if a company would like to create a unique marketing strategy for specific target groups.

Anomaly detection is a data mining method used to detect objects that do not subscribe to the expected pattern of the data set (Chandola, Banerjee, & Kumar, 2009). This approach could be used by a bank or credit card company to analyze data for fraud detection.

The association analysis method is used to find association and correlation relationships in extremely large data sets (Han & Kamber, 2006). This method could be helpful if a company is trying to determine a subset of items that a customer typically buys at the same time.

Classification is a type of predictive analysis on a set of data. This method is used to create classification models that distinguish specific classes from one another with the purpose of using the model to predict the class of an object whose label is previously unknown (Han & Kamber, 2006). This is the method that is used in this study.

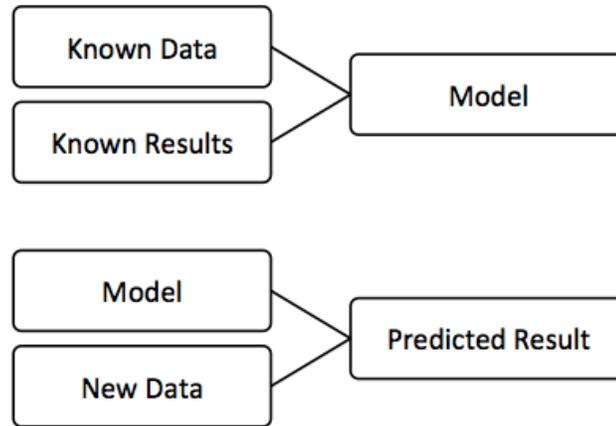
### **1.3 Classification Approach**

Our goal was to use the classification data mining approach to create a model that can predict the retention status of students based on their educational and demographic data. The typical steps of the classification approach include preprocessing, application of the algorithm, and validation of results (De Bruyne & Plastria, 2010). Each of these steps also includes different tasks.

Data preprocessing in the classification approach includes both data cleaning and attribute selection. Data cleaning typically involves handling missing data values.

Missing data values in large data sets is unavoidable and a familiar issue for projects in the field of data mining. These missing values can cause issues when building the predictive model and can also be problematic when running new data through the model (Han & Kamber, 2006). Typical solutions include removing records with missing fields or filling the missing values with estimates (De Bruyne & Plastria, 2010). Another preprocessing step is attribute selection. Attribute selection aims to narrow down the number of attributes in a data set to the most relevant attributes to predict the target variable (De Bruyne & Plastria, 2010). This helps to reduce the size of the large data set and avoid overfitting, which can occur for overly complex data models. Overfitting occurs when the data is overly analyzed to reduce error for the training set, which in turn increases the error of the model when used to predict future instances (Han & Kamber, 2006).

The next step is the application of the classification algorithm. This step is where the actual predictive model is built. The process of building a predictive classification model can be referred to as supervised learning. Supervised learning is the task of building a model by learning from a given set of data with known results (Mohri, Rostamizadeh, & Talwalkar, 2012). A model of this task can be seen in Figure 1.



**Figure 1: Visualization of model creation and prediction for the classification approach.**

The final step of classification is the validation of the derived model. This step includes testing the model against never before seen data and analyzing the results of the model. The accuracy of the model is calculated based on the number of instances that the model classifies correctly (Han & Kamber, 2006). Once the model is validated and determined acceptably accurate, it can be used to predict outcomes of new data.

## **1.4 Common Classification Techniques**

Much like there are different approaches to Data Mining as a whole, there are also many different techniques within the classification method. Some of the most common methods include: decision tree method, nearest neighbor method, naïve Bayes method, regression method, and support vector machine method. In each of these approaches, a classification model is learned from an existing set of data, which can then be used to classify instances of a new data set into specific classes.

A decision tree model is a recursively built tree that is essentially a set of if-then-statements (Li, Wu, Wallin, & Hains, 2015). Each node of the tree represents a test on an attribute, each branch represents the result of that test, and each terminal node, referred to as a leaf, denotes the classification of the object (Han & Kamber, 2006). To classify an object with an unknown class, the attribute values of this item are tested by the decision tree. It will follow the nodes and branches until it eventually reaches a leaf node which holds the objects class prediction.

*k*-Nearest-neighbor methods classify an object based on the closeness of the object to other objects in the data set (Hastie, Tibshirani, & Friedman, 2008). This method finds the *k* number of objects that are most similar to the tested object, and classifies the object with the majority label from its neighborhood (Li, Wu, Wallin, & Hains, 2015).

The naïve Bayes method is a type of statistical classifier. This method predicts the probability that a given object belongs to a specific class (Han & Kamber, 2006). This method is considered “naïve” because it assumes that the effect of an attribute on the class value is independent of the values of all other attributes (Hastie, et al., 2008).

Using the logistic regression method, the class is predicted based on its relationship to predictor variables. A separate logistic regression model is built for each class of the target variable (Li, Wu, Wallin, & Hains, 2015).

The support vector machine (SVM) method searches for the optimum “decision boundary” that separates class objects from one another (Han & Kamber, 2006). This model determines the separating margin of the values of attributes that is an indicator of a specific class. It also can determine which attributes are more predictive on the class

result than other attributes (Bruyne & Plastria, 2010). Typically, the SVM approach is used as a “linear classifier”, which means that it is used on data sets that have only 2 distinct classes (Bruyne & Plastria, 2010).

For this study, we will be using the decision tree classification method. Decision tree classification is widely used and very efficient at classifying data objects (Han & Kamber, 2006).

## **1.5 Previous Studies of Student Retention**

Student retention rates have been a long standing issue and a hot-spot for data analysis in universities. Many studies have been conducted to predict student retention using empirical data. These studies, coupled with the creation of analytic models, have been found to be beneficial to student success and retention (Arnold, 2010).

The foundation for this type of analysis was built with a report discussing a Theoretical Model of Dropout (Tinto, 1975). Tinto elaborates on the complexities of what goes into a student’s decision to leave a university and the causes of a student being forcibly removed from a university. Most tend to believe that a student’s intellect and grades will ultimately determine his or her success in higher-education, but Tinto argues that “it is the individual's integration into the college environment which most directly relates to continuance in college.” This integration with the university can be affected by any number of attributes relating to the student’s family background, former educational experiences, and motivation to succeed. More recent models also use many different attributes to predict student retention, such as high school information, financial status, and early college information (Herzog, 2005; Ronco & Cahill, 2006). Other models put a

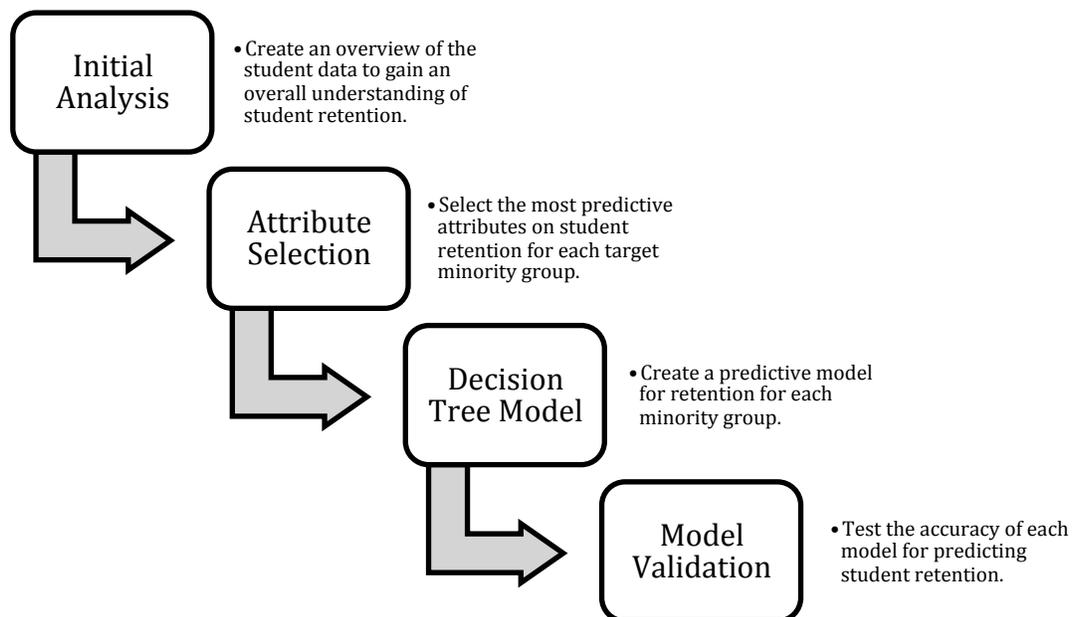
specific focus on financial factors when attempting to predict student retention (Bresciani & Carson, 2002; Chen & DesJardins, 2010). These models provide an initial foundation for the creation of a model specifically relevant for students at Middle Tennessee State University.

# Chapter 2

## Methodology

### 2.1 Overview

The goal of this study was to create a model to predict student retention. We created separate models for unique minority groups at MTSU. I designed 4 experiments to complete this task, which are outlined in Figure 2 below. The 4 steps include an initial analysis of the student data set, attribute selection of predictive attributes on student retention for each minority group, creation of the decision tree models for each group, and validation of the models.



**Figure 2: Outline of the steps to create a predictive model for student retention.**

## 2.2 General Explanation of Data Set

This study analyzes a compilation of student data that has been collected over the past 7 years at MTSU. The Office of Institutional Effectiveness, Planning and Research (IEPR) at MTSU provided that data being used in this study. The comprehensive data set includes information from all students enrolled at MTSU between 2007 and 2013. This study concentrates on analyzing only data from undergraduate students after their first year of enrollment and their retention status at the end of their college career. Only students who were enrolled full time for at least two semesters in a row with their first semester being the fall semester are included in the data set.

The data set contains 157 variables for each student. These attributes cover many different types of information, which can be broken down into the following categories:

- 1) demographic information (e.g., family income, ethnicity, gender, etc.);
- 2) high school information (e.g., GPA, ACT scores, classes taken, etc.);
- 3) college information after their first year (e.g., GPA, courses taken, major, on-campus living, etc.); and
- 4) financial information (e.g., total amount of financial aid, student loans, etc.).

In order to have a complete understanding of the data set, it is important to understand each attribute included in the set, such as: the definition of the attribute name, the domain of its potential values, and what each value actually represents. The most important attribute to be aware of with regards to this study is the student's "status" at the end of his or her college career. A student's status can be one of three values: (1) Stayed, (2) Transferred, or (3) Dropped. The information regarding attributes in the data set has been compiled into a dictionary and is included in Appendix A.

The comprehensive data set has been broken down into four distinct minority groups including:

- 1) African American Students
- 2) First Generation Students
- 3) Disabled Students
- 4) Hispanic Students

These minority groups were selected by the faculty members working on the project because of the large number of students in each of these groups at MTSU and the university's interest in their retention rates. Each data set was analyzed separately to determine if there were different models to predict a student's retention depending on his or her affiliation with a specific minority group. Splitting the data into specific minority groups allowed the data to be more distinctly analyzed for a specific student, hopefully allowing MTSU to use this information to improve student retention within minority groups.

## **2.3 Initial Analysis of Data**

Before a full analysis of the data began, a high level investigation of the student data was completed regarding the target variable (student retention status). To determine if there was an overall difference in retention rates between each target minority group, I created an overview that looked at each target minority group and found the percentage of students who stayed, transferred, and dropped from the university. I then split up each

group based on GPA and created the same type of overview. I split the GPA in 3 separate ways:

- 1)  $\text{GPA} < 2.0, \text{GPA} \geq 2.0$
- 2)  $\text{GPA} < 2.4, \text{GPA} \geq 2.4$
- 3)  $\text{GPA} < 2.75, \text{GPA} \geq 2.75$

Next, I generated an overview for each split. The first GPA split (2.0) was based on if the student had a C-average or higher, which is required for many courses. The last split (2.75) was chosen because many scholarships, including the Tennessee HOPE Lottery Scholarship, require a 2.75 GPA to be renewed. The second split was chosen as an intermediate split between the other two. This general analysis gives an abstract summary of the retention data and generated many questions regarding relationships between financial aid, student grades, and student retention. This overview allows for the development of a general understanding of the data.

## **2.4 Attribute Selection**

The next step of this study was attribute selection. The initial data set included 157 different attributes. The goal of attribute selection is to perform statistical analysis on these attributes in order to narrow down the collection to only the top most predictive attributes for student retention. This selection was completed separately for each target minority group: African American, disabled, first generation, and Hispanic students. The top predictive attributes were expected to vary between the different minority populations.

To perform the data analysis, I used the Weka data mining tool. Weka is an open source data mining software that provides machine learning algorithms for use on large sets of data. Weka provides many different methods of machine learning and analysis of statistical data.

For this feature selection, I completed a chi-square analysis of each attribute to find the highest predictive correlation to student retention. This test treats each attribute as an independent variable and computes its predictiveness of the target variable. The hypotheses for the test include:

$H_0$ : The selected attribute and student's status are independent.

$H_a$ : The selected attribute and student's status are not independent.

The chi-square test for the independence of attributes will determine whether each attribute is a possible candidate for predicting the result of a student's retention status at the end of their first year of enrollment. Specifically, it will determine if the attribute independent of the students status. If the attribute and student's status are not independent, the variable is a good candidate for prediction of the student's status. After determining which attributes are candidates for the prediction of a student's status, we can choose the top most predictive attributes. The selected attributes will then be used for the development of a predictive model for student retention.

## **2.5 Creation of the Decision Tree Model**

The next step of this study was to create the classification model that is used to predict student retention. For this study, we developed a decision tree model. A decision tree is a set of if-then-statements built recursively from the root to the leaves. All of the

data begins in the root of the tree. The leaves contain the end result for our prediction of student retention. The tree grows by recursively branching at each node, beginning at the root. The algorithm used to build the tree will compute the Boolean result of an *if statement* at each node, using the most predictive remaining attribute as the criteria for the expression. The decision tree is completed when one of the two situations occur:

- 1) Each end node (leaf) contains a collection of data values that have the same result for the target variable (in our case this would be the student's status).
- 2) Every attribute has already been used.

Once the model was created, the tree was pruned to prevent over-fitting, which occurs when the model is too specific to the given data set. Ideally, each leaf is a homogeneous mixture of dropped, transferred, or stayed students.

This classification model was built using the Weka Data Mining software that was introduced earlier. To build the model, the program used a subset of the data set as training data to learn the classification rules. We used 90% of the data set as the training data. After the rules were learned (the tree was built), it was used to classify new, unseen student data. These classifications are into categories of stayed, transferred, or dropped. The remaining 10% of the data set was used to test the model. This led us to the final step of this study: validation of the model.

## **2.5 Validation of the Model**

To determine how well the model performed when classifying new, never before seen data, it was important to test the model on data that was not already used for the

construction of the model. This is why we constructed the decision tree using only 90% of the data set as the training set.

The method used to test the effectiveness of a classification model is called ten fold cross validation. This method randomized the data and broke it into 10 equal sets . It then used 9 of the 10 sets to construct the decision tree model, and ran the 10<sup>th</sup> set as test data (Han & Kamber, 2006). The accuracy of the classification was calculated. This process was completed 10 separate times, with the classification performance scores averaged at the end of the 10 iterations. These scores told us how effective the model was at predicting whether a student will stay, transfer, or drop from the university for each of the 4 minority groups.

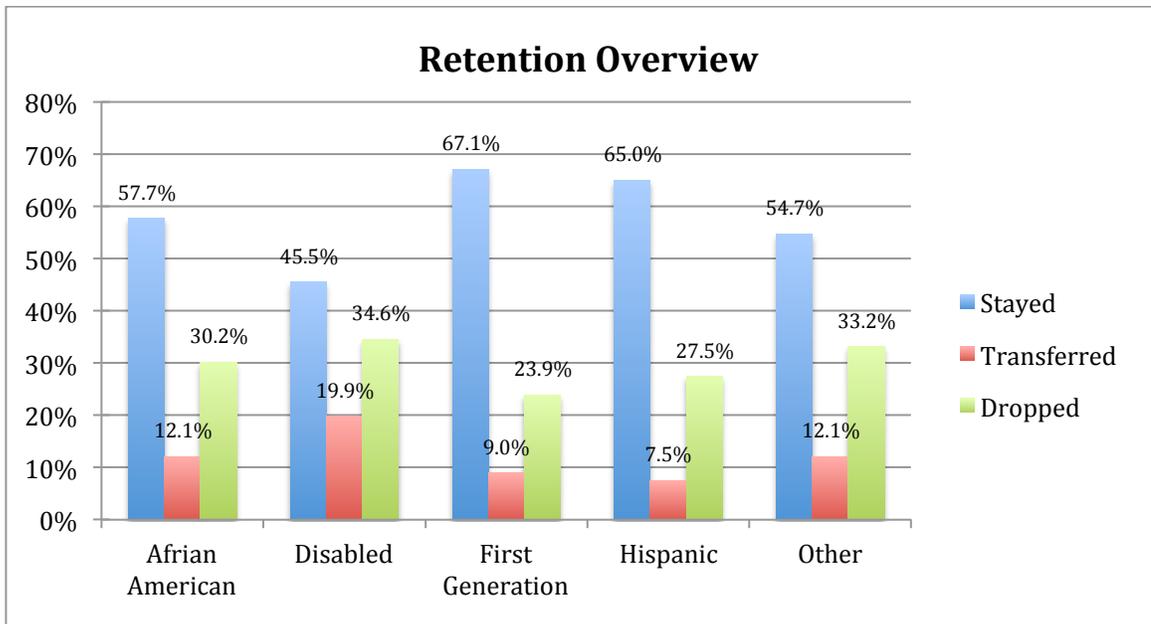
If a model was not as successful as we would like, we rebuilt the model and tested its accuracy with another round of ten fold cross validation.

# Chapter 3

## Experimental Results

### 3.1 Initial Analysis

The results of the high level analysis of the student data gave us an initial understanding of the data with regards to the status target variable. I calculated the percentage of students who stayed, transferred, and dropped for each of the 4 minority categories: African American, disabled, first generation, and Hispanic. I also completed analysis on a subset of all other students, which includes all students that are not in one of the four defined minority groups. A complete record of these results can be found in the tables at the end of each section, and an overview of all of the results can be found in Figure 3 below.



**Figure 3: Overview of the results of student retention classifications for each minority group.**

Looking at all of the freshman students, I found that the average amount of retained students ranged anywhere from 45.5% to 67.1%, with disabled students having the smallest percentage of students with a “stayed” status after the first year of enrollment, and first generation students having the highest percentage of students continue to be enrolled after their first year. African American students stayed at a rate of about 57.7%. Hispanic students came in with the second highest retention rate of the minority groups with 65% of students being retained. All other students had a retention rate of 54.7%, right in the middle of the range of retention rates. Transferring students within the minority groups ranged from 7.5% (Hispanic) to 19.9% (disabled). African American students transferred at a rate of 12.1%, and 9.0% of first generation students transferred from MTSU after their first year. Other students transferred from the university at a rate of 12.1%.

**Table 1: Count and percentage of all freshman students who Stayed, Transferred, and Dropped at MTSU.**

	Total	Stayed	Transferred	Dropped
African American	7337	4237	886	2214
Disabled	1017	463	202	352
First Generation	1801	1209	162	430
Hispanic	1435	933	108	394
Other	85857	47004	10377	28476

**(a) Count**

	Total	Stayed	Transferred	Dropped
African American	1	57.7%	12.1%	30.2%
Disabled	1	45.5%	19.9%	34.6%
First Generation	1	67.1%	9.0%	23.9%
Hispanic	1	65.0%	7.5%	27.5%
Other	1	54.7%	12.1%	33.2%

**(b) Percentage**

We hypothesized that the retention rates would strongly differ based on the GPA of the student. In order to test this hypothesis, I completed the same analysis as before, but on different subsets of the data based on GPA.

The first split was at a 2.0 GPA. This split is at a C average, which is required for many pre-requisite courses to move on to the next level and for major courses. The category of other students had a retention rate of 61.5% for students above a 2.0 GPA and 11.9% for students with a GPA lower than a 2.0. The subset of all other students with a GPA above 2.0 transferred and dropped with rates of 9.5% and 29.0%, respectively.

In the subset of students with a GPA higher 2.0, the retention rate ranged from 58.1% (disabled) to as high as 78.8% (first generation). African American students had a 69.9% rate and Hispanic students had a 73.0% success rate. Students transferred out of the university at rates of 14.2% (disabled), 9.5% (African American), 7.1% (first generation), and 6.3% (Hispanic). Dropout rates for this subset of students includes rates

of 14.1% (first generation), 20.6% (African American), 20.6% (Hispanic), and 27.6% (disabled).

**Table 2: Count and percentage of all freshman students having first year of college GPA  $\geq$  2.0 who Stayed, Transferred, and Dropped at MTSU.**

	Total	Stayed	Transferred	Dropped
African American	5425	3794	513	1118
Disabled	743	432	106	205
First Generation	1444	1138	103	203
Hispanic	1202	878	76	248
Other	74193	45616	7038	21539

**(a) Count**

	Total	Stayed	Transferred	Dropped
African American	1	69.9%	9.5%	20.6%
Disabled	1	58.1%	14.3%	27.6%
First Generation	1	78.8%	7.1%	14.1%
Hispanic	1	73.0%	6.3%	20.6%
Other	1	61.5%	9.5%	29.0%

**(b) Percentage**

The results from the subset of students with a GPA less than 2.0 is significantly different. This group of students has retention rates that vary from 23.6% to 11.3%, with Hispanic students being the most successful, continuing enrollment at under 24% and disabled students being the least successful. African American students with less than a 2.0 GPA continued enrollment at a rate of 23.2%, and first generation students at a rate of 19.9%. Students also transferred at higher rates within this subset; Hispanic students transferred at 13.7%, first generation at 16.5%, African American at 19.5%, and disabled at 35%. The dropout rate for students with lower than a 2.0 GPA was staggeringly high with rates as high as 63.6% for first generation students. Attrition rates were high for other minority groups as well: 62.7%, 57.3%, and 53.6% for Hispanic, African American,

and disabled students, respectively. Students not fitting any minority groups (categorized as other) were actually the least successful at this GPA break.

**Table 3: Count and percentage of all freshman students having first year of college GPA < 2.0 who Stayed, Transferred, and Dropped at MTSU.**

	Total	Stayed	Transferred	Dropped
African American	1912	443	373	1096
Disabled	274	31	96	147
First Generation	357	71	59	227
Hispanic	233	55	32	146
Other	11664	1388	3339	6937

**(a) Count**

	Total	Stayed	Transferred	Dropped
African American	1	23.2%	19.5%	57.3%
Disabled	1	11.3%	35.0%	53.6%
First Generation	1	19.9%	16.5%	63.6%
Hispanic	1	23.6%	13.7%	62.7%
Other	1	11.9%	28.6%	59.5%

**(b) Percentage**

The next GPA break that I analyzed was for students with a 2.4 GPA or higher. Students in this GPA range stay enrolled at a rate of up to 83.7% for first generation students, an improvement from the previously analyzed subset of students above a 2.0 GPA. Hispanic students rank second highest, staying at MTSU at 76.1%, followed by African American students at 74.7%, and disabled students at 63.9%. Other students with a 2.4 GPA or above are retained at 65.1%. The transfer rate for these students is only slightly lower than students in the above 2.0 GPA range. Hispanic students transfer at the lowest rate (5.5%) with first generation students at a close 5.7% transfer rate. African American students and other students both transfer at a rate of 8.1%. Disabled students have the highest transfer rate at 11.5%. First generation students in this subset have the smallest dropout rate (10.6%) and disabled students have the highest dropout rate of the

minority groups (24.5%). However, other students drop out at a rate of 26.9%, which is higher than any of the defined minority groups. Hispanic and African American students drop at 18.4% and 17.2%, respectively.

**Table 4: Count and percentage of all freshman students having first year of college GPA  $\geq$  2.4 who Stayed, Transferred, and Dropped at MTSU.**

	Total	Stayed	Transferred	Dropped
African American	4273	3193	345	735
Disabled	607	388	70	149
First Generation	1227	1027	70	130
Hispanic	1033	786	57	190
Other	63297	41187	5096	17014

**(a) Count**

	Total	Stayed	Transferred	Dropped
African American	1	74.7%	8.1%	17.2%
Disabled	1	63.9%	11.5%	24.5%
First Generation	1	83.7%	5.7%	10.6%
Hispanic	1	76.1%	5.5%	18.4%
Other	1	65.1%	8.1%	26.9%

**(b) Percentage**

In the subset of students with a GPA less than 2.4, the retention rate ranged from 21.0% (disabled) to as high as 41.5% (Hispanic). African American students had a 38.0% rate and first generation students had a 36.8% success rate. Students transferred out of the university at rates of 30.6 (disabled), 20.4% (other), 15.0% (African American), 12.3% (first generation), and 9.5% (Hispanic). Dropout rates for this subset of students are all close to 50% (47.0% to 52.2%).

**Table 5: Count and percentage of all freshman students having first year of college GPA < 2.4 who Stayed, Transferred, and Dropped at MTSU.**

	Total	Stayed	Transferred	Dropped
African American	2727	1037	409	1281
Disabled	353	74	108	171
First Generation	487	179	60	248
Hispanic	347	144	33	170
Other	18065	4949	3692	9424

**(a) Count**

	Total	Stayed	Transferred	Dropped
African American	1	38.0%	15.0%	47.0%
Disabled	1	21.0%	30.6%	48.4%
First Generation	1	36.8%	12.3%	50.9%
Hispanic	1	41.5%	9.5%	49.0%
Other	1	27.4%	20.4%	52.2%

**(b) Percentage**

The final GPA subsets include students with GPAs over/under 2.75. This GPA is specifically relevant because of the Tennessee HOPE Lottery Scholarship. After a student's first year of study, they must retain at least a 2.75 GPA to continue to receive the lottery scholarship. For the years analyzed in this study (2007-2013), the scholarship amounted to \$2000 per semester, and \$4000 for a full year.

Students with GPA higher than 2.75 have a much better success rate. First generation students rank highest, continuing enrollment at a rate of over 86%. African American and Hispanic students are both retained at near 78%, and disabled students are retained at 66.3%. Transfer rates remain near 5-8%, with the exception of disabled students, who transfer from MTSU at almost 11%. First generation students dropped from the university at a rate of 8.3%, the lowest that we have seen thus far. African American students drop at a rate of 14.4%, Hispanic students at a rate of 16.3%, and disabled students at 22.7%.

**Table 6: All freshman students having first year of college GPA  $\geq 2.75$  who Stayed, Transferred, and Dropped at MTSU.**

	Total	Stayed	Transferred	Dropped
African American	3120	2444	227	449
Disabled	466	309	51	106
First Generation	979	848	50	81
Hispanic	822	641	47	134
Other	51457	34942	3804	12711

**(a) Count**

	Total	Stayed	Transferred	Dropped
African American	1	78.3%	7.3%	14.4%
Disabled	1	66.3%	10.9%	22.7%
First Generation	1	86.6%	5.1%	8.3%
Hispanic	1	78.0%	5.7%	16.3%
Other	1	67.9%	7.4%	24.7%

**(b) Percentage**

The next subset is students who have below a 2.75 GPA. If these students had the HOPE lottery scholarship when they began enrollment at MTSU, they would have lost it after their first year due to failing to meet the GPA requirement. Hispanic students in this set stay at MTSU at a rate of 51.8%, first generation students at 48.7%, African American students at 46.0%, and disabled students at 31.0%. Students transfer from MTSU to another university at up to 25.7% for disabled students. 13.6% of African American students, 10.9% of first generation students, and 7.7% of Hispanic students transfer. Drop out rates for African American, Hispanic, and first generation students are around 40% and disabled students drop at 43.3%.

**Table 7: Count and percentage of all freshman students having first year of college GPA < 2.75 who Stayed, Transferred, and Dropped at MTSU.**

	Total	Stayed	Transferred	Dropped
African American	3880	1786	527	1567
Disabled	494	153	127	214
First Generation	735	358	80	297
Hispanic	558	289	43	226
Other	29905	11194	4984	13727

**(a) Count**

	Total	Stayed	Transferred	Dropped
African American	1	46.0%	13.6%	40.4%
Disabled	1	31.0%	25.7%	43.3%
First Generation	1	48.7%	10.9%	40.4%
Hispanic	1	51.8%	7.7%	40.5%
Other	1	37.4%	16.7%	45.9%

**(b) Percentage**

Disabled students consistently have one of the highest drop out rates among the four minority groups, excluding the subset of students with GPA below 2.0, where disabled students are actually the most resilient with the lowest drop out rate. Disabled students also tend to transfer at a higher rate than any other minority group. African American students and Hispanic students have similar retention rates. First generation students with higher GPAs tend to continue enrollment at the highest rate.

### 3.2 Attribute Selection

The first step of attribute selection for each minority group was completing a chi-square test on each attribute to determine which attributes have the highest predictive correlation on a students' status after their first year. The Weka software aided in performing these tests and ranked the attributes from most predictive to least predictive. As expected, there was some variation between the minorities groups' ranking of

attributes. The number of attributes selected for each group correlated to the sample size of the group. The African American student group was the largest (7337 students) and had the largest number of attributes selected as predictive for student retention (51 attributes). The number of attributes selected for first generation, Hispanic, and disabled students also differed depending on the group's sample size.

For African American students, there were 51 attributes that showed predictive correlation to the status of students' retention. The top two ranking attributes, INST\_LGPA\_GPA and INST\_TGPA\_GPA, are both relating to the grade point average of the student. LGPA is defined as the cumulative institutional GPA for the student, and TGPA is another calculation of the student's GPA based on the term and GPA hours. FIN\_SUS, Financial Suspension Status, was the next highest predictive attribute. A student can be put on Financial Aid Suspension due to unsatisfactory progress. Satisfactory academic progress is monitored by the student's GPA, percentage of credit hours passed, and total number of credit hours. The next highest ranked attributes include percentage of courses withdrawn, high-school GPA, student type, financial probation, total credit hours, family total income, scholarship received, age, housing, compass English scores, independent status, loans received, and high school math classes. Other interesting attributes that are predictive for African American students include the students' nontraditional status or if the student has children. All 51 of the attributes along with their ranked score can be found in Table 8 below.

The analysis for disabled students showed 16 attributes with predictive correlation to retention status. Similar to African American students, the top two attributes for disabled students were based on the students' GPA. Percentage of courses withdrawn and

financial probation and suspension were the next highest ranked attributes. The remaining selected attributes include student's status as an honors student, 3 different attributes associated with enrollment in math courses at a college or high school level, institutional aid and grants, and the education level of a student's father and mother.

First generation students have 21 attributes identified to be important in predicting student retention. These attributes include the GPA and financial status attributes mentioned above. Other attributes of note include many from the student's high school, such as: high school GPA, math classes and ACT math score, and AP courses. Home environment, family size, and work hours also show as selected attributes.

Hispanic students have 19 attributes selected as identifiers for retention status. The top two attributes are once again related to a student's institutional GPA. Percentage of courses with reported difficulty during academic progress reports (PC\_PROG\_COURSES) is the third highest predictive attribute. The next highest attributes all relate to student's home life, including: parent total income, dependent status, family size, and gross income. Age, home environment, father's education level, single marital status, and nontraditional student status are also included in the selected attributes.

Many attributes appear as selected features for multiple target groups. The students' institutional GPA is ranked as the highest predictive attribute for student retention for all four target minority groups. Financial status seems to be the next best predictor, with FIN\_PROB and FIN\_SUS appearing near the top of each group. Other financial predictors, such as total income, parent total income, scholarships, loans, and grants appear for one or more of the target groups. There are 8 attributes that appear in

three of the four target groups: independent status, family size, father’s education level, help education plans, advanced high school math, calculus high school math, dollar amount of scholarships, and percentage of courses with grade “W.” The attribute indicating father’s education level appears for three groups: African American, disabled, and Hispanic, while mother’s education level only appears for the African American and disabled target groups. Student age appears to impact retention for African American and Hispanic students, but does not show up for the disabled or first generation groups.

**Table 8: Selected Attributes in descending order** – Attributes identified by chi-square analysis with the most predictive correlation to a student’s retention status for the four minority target groups in descending order of importance.

African American	Disabled	First Generation	Hispanic
INST_LGPA_GPA	INST_LGPA_GPA	INST_LGPA_GPA	INST_LGPA_GPA
INST_TGPA_GPA	INST_TGPA_GPA	INST_TGPA_GPA	INST_TGPA_GPA
FIN_SUS	PC_ATTEND_W_COURSES	PC_ATTEND_W_COURSES	PC_PROG_COURSES
PC_ATTEND_W_COURSES	FIN_PROB	FIN_SUS	PARENT_1_TOT_INCOME
HS_GPA2	FIN_SUS	HS_GPA2	Dstatus
Student_Type2	HONORS	PC_WITHDRAW_COURSES	FAMILY_SIZE2
FIN_PROB	PCOL_ASSOCIATES	PC_PROG_COURSES	GROSS_INCOME_FAFA
TotalCreditHours	HSCourseBegCalc	FIN_PROB	Student_Type2
FM_TOTAL_INCOME	HSCourseOtherAdvMath	PC_DFWN_COURSES	FIN_SUS
SCHLD2	FatherEdLevel	HSMathClasses	SCHLD2
Age	INSTAIDD2	SZBSTER_ACT_MATH	Age
Housing	HelpEducationPlans	Dstatus	HOME_ENVIR
COMPASS_ENGLISH	GRNTD2	HOME_ENVIR	Single
PARENT_1_TOT_INCOME	Gender2	COMPASS_MATH	FIN_PROB
Dstatus	MotherEdLevel	FAMILY_SIZE2	Nontraditional
PC_DFWN_COURSES	MATH1710	ACT_Work_Hours_21	HSCourseOtherAdvMath
LOAND2		SCHLD2	HSCourseBegCalc
HSMathClasses		Medium_High_School	HelpEducationPlans
Nontraditional		ANYPRESC	FatherEdLevel

**Table 8 Continued: Selected Attributes in descending order**

<b>African American</b>	<b>Disabled</b>	<b>First Generation</b>	<b>Hispanic</b>
HSCourseBegCalc		AP	
HSCourseOtherAdvMath			
Dependents			
Learning_Community			
S_Parent			
INSTAIDD2			
PC_ONLINE_COURSES WRODP			
PC_WITHDRAW_COUR SES			
HAVE_CHILDREN			
HS Curr Collprep			
Gender2			
EVENING STUDENT			
PC_ONLINE_COURS ES			
ECHSHIGH			
GRNTD2			
Accomp3			
ECCOLHIGH			
HelpMathSkills			
HelpEducationPlans			
Medium_High_School			
MotherEdLevel			
FatherEdLevel			
FAMILY_SIZE2			
HelpReading			
MATH1530			
MATH1710			
UNMET_NEED2			
Single			
WAIVER			
HelpEdPlans TWO areas			
ATHLD2			

**Table 9: Selected Attributes in alphabetical order** – Most predictive attributes to a student’s retention status in alphabetical order, showing the number of groups in which each attribute appears.

**Key:** Attribute appears in  4 groups  3 groups  2 groups  1 group

African American	Disabled	First Generation	Hispanic
Accomp3			
		ACT_Work_Hours_21	
Age			Age
		ANYPRESC	
		AP	
ATHLD2			
COMPASS_ENGLISH			
		COMPASS_MATH	
Dependents			
Dstatus		Dstatus	Dstatus
ECCOLHIGH			
ECHSHIGH			
EVENING_STUDENT			
FAMILY_SIZE2		FAMILY_SIZE2	FAMILY_SIZE2
FatherEdLevel	FatherEdLevel		FatherEdLevel
FIN_PROB	FIN_PROB	FIN_PROB	FIN_PROB
FIN_SUS	FIN_SUS	FIN_SUS	FIN_SUS
FM_TOTAL_INCOME			
Gender2	Gender2		
GRNTD2	GRNTD2		
			GROSS_INCOME_FAFSA
HAVE_CHILDREN			
HelpEdPlans_TWO_areas			
HelpEducationPlans	HelpEducationPlans		HelpEducationPlans
HelpMathSkills			
HelpReading			
		HOME_ENVIR	HOME_ENVIR
	HONORS	HONORS	
Housing			
HS_Curr_Collprep			
HS_GPA2		HS_GPA2	
HSCourseBegCalc	HSCourseBegCalc		HSCourseBegCalc
HSCourseOtherAdvMath	HSCourseOtherAdvMath		HSCourseOtherAdvMath

**Table 9 Continued: Selected Attributes in alphabetical order**

African American	Disabled	First Generation	Hispanic
INST_LGPA_GPA	INST_LGPA_GPA	INST_LGPA_GPA	INST_LGPA_GPA
INST_TGPA_GPA	INST_TGPA_GPA	INST_TGPA_GPA	INST_TGPA_GPA
INSTAIDD2	INSTAIDD2		
Learning_Community			
LOAND2			
MATH1530			
MATH1710	MATH1710		
Medium_High_School		Medium_High_School	
MotherEdLevel	MotherEdLevel		
Nontraditional			Nontraditional
PARENT_1_TOT_INCOME			PARENT_1_TOT_INCOME
PC_ATTEND_W_COURSES	PC_ATTEND_W_COURSES	PC_ATTEND_W_COURSES	
PC_DFWN_COURSES		PC_DFWN_COURSES	
PC_ONLINE_COURSES			
PC_ONLINE_COURSES_WRODP			
		PC_PROG_COURSES	PC_PROG_COURSES
PC_WITHDRAW_COURSES		PC_WITHDRAW_COURSES	
	PCOL_ASSOCIATES		
S_Parent			
SCHLD2		SCHLD2	SCHLD2
Single			Single
Student_Type2			Student_Type2
		SZBSTER_ACT_MATH	
TotalCreditHours			
UNMET_NEED			
UNMET_NEED2			
WAIVER			

### 3.3 Decision Tree Model

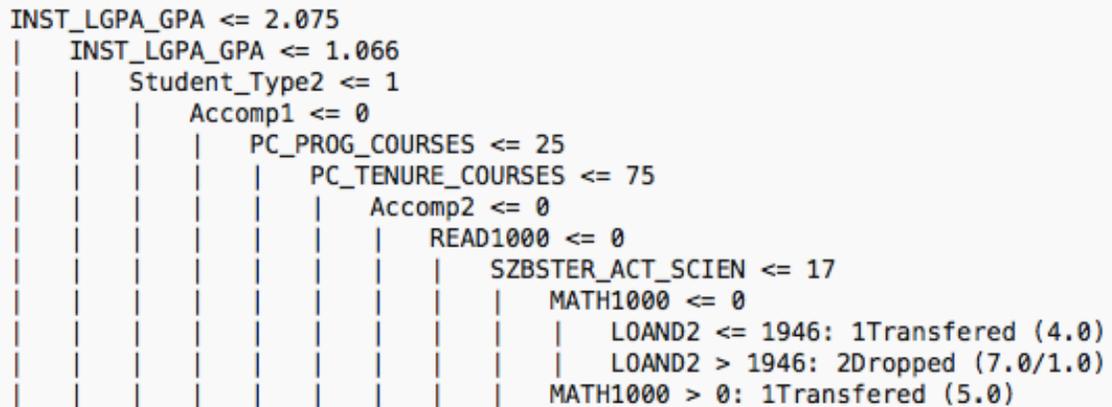
The next step of this study was to create the predictive models for student retention. A separate model was built for each target minority group. Because different attributes were selected as predictive for different minority groups in the last step, having

different models should allow us to predict more accurately retention for minority students that fit in one of these target groups. The nodes that appear closer to the root of the tree tend to be the most predictive, as these are the attributes that are able to initially break down the data set into like classifications. The leaves of the tree are the terminal nodes of the tree and represent the resulting classification.

The decision tree model for the African American group is by far the largest and most complex model. This complexity is due to the large sample size and number of attributes that were selected as predictive for retention. This model has 515 leaves and 1029 total nodes. I will walk through the first few if statements until we reach the first set of leaves. The decision tree for these steps is represented in Figure 4. The complete tree can be found in the appendix on page 53.

If a student:

- has GPA less than or equal to 2.075, AND
- has GPA less than or equal to 1.066, AND
- has student type less than or equal to 1, AND
- does not have 1 recorded accomplishment, AND
- has less than or equal to 25% of their courses reported academic difficulties, AND
- has less than or equal to 75% of their courses are taught by tenured professors, AND
- does not have 2 recorded accomplishments, AND
- is not enrolled in READ 1000, AND
- scored 17 or below on the Science portion of their ACT, AND
  - is not* enrolled in MATH 1000, AND
    - has less than \$1,946 in student loans,
      - Then the student is predicted to **transfer**.
    - has loans totaling more than \$1,946,
      - Then the student is predicted to **drop** from the university.
  - is* enrolled in MATH 1000,
    - Then the student is predicted to **transfer**.



**Figure 4: Portion of Decision Tree Model for African American Students from the root to the first set of leaves.**

The decision tree model developed for first generation students is expectedly much smaller than the tree for African American students because of the total sample size of first generation students. This tree has 81 leaves and 161 total nodes. An interesting attribute of this tree is that some students can be classified based on the values of only 2 attributes: Institutional GPA and Parent Total Income. The complete tree can be found in the Appendix on page 63, and a portion of the tree is described below and represented in Figure 5.

- If a student:
  - has a GPA less than or equal to 1.692, AND
  - parents' total income is *less than or equal to* \$82,996,
    - Then the student is predicted to **drop** from the university.
  - parents' total income is *greater than* \$82,996, AND
  - scored *less than or equal to* 20 on the math portion of their ACT,
    - Then the student is predicted to **transfer** from the university.
  - scored *greater than* 20 on the math portion of their ACT, AND
  - takes a W in less than or equal to 0% of courses,
    - Then the student is predicted to **stay** at the university.
  - takes a W in greater than 0% of courses,
    - Then the student is predicted to **drop** from the university.

```

INST_TGPA_GPA <= 1.692
| PARENT_1_TOT_INCOME <= 82996: 2Dropped (77.0/13.0)
| PARENT_1_TOT_INCOME > 82996
| | SZBSTER_ACT_MATH <= 20: 1Transferred (6.0)
| | SZBSTER_ACT_MATH > 20
| | | PC_ATTEND_W_COURSES <= 0: 0Stayed (4.0/1.0)
| | | PC_ATTEND_W_COURSES > 0: 2Dropped (2.0)

```

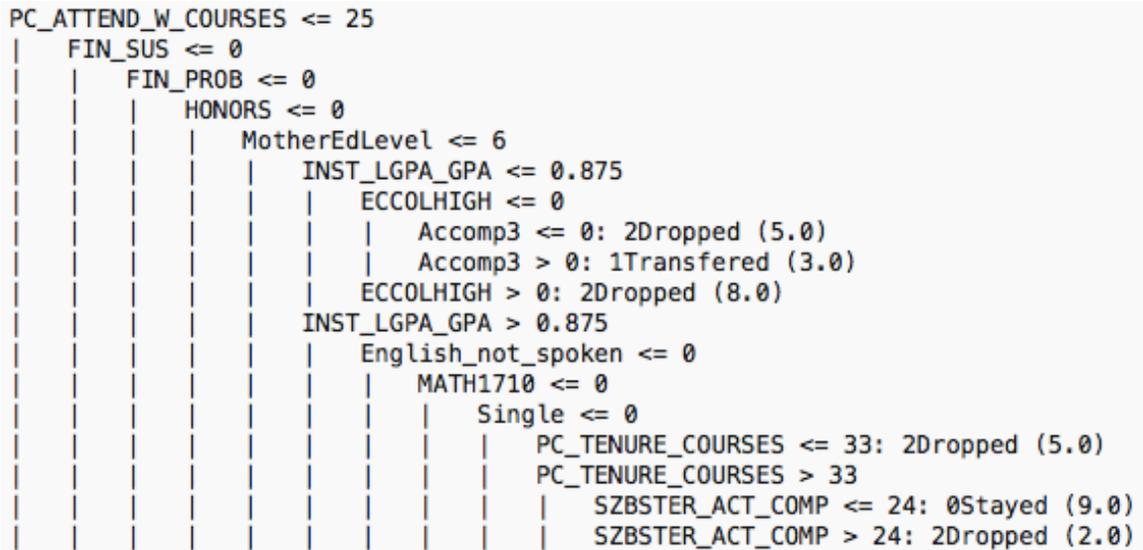
**Figure 5: Portion of Decision Tree Model for first generation Students from the root to the first set of leaves.**

The decision tree model developed for disabled students is also much smaller than the African American students' model, and about the same size as the model for first generation students. This tree has 86 leaves and 171 total nodes. The nodes closest to the root selected to initially break down the data set into classified status groups use PC\_ATTEND\_W\_COURSES, FIN\_SUS, FIN\_PROB, HONORS, Mother Education Level, and GPA as the predictive attributes. For this tree, I would like to explain the nodes that lead to the first leaf that is classified as *stayed*. The complete tree can be found in the Appendix on page 64.

If a student:

- takes a W in less than or equal to 25% of courses, AND
- is not on financial suspension, AND
- is not on financial probation, AND
- is not in Honors, AND
- mother's education level is less than or equal to 6, AND
- has a GPA greater than 0.875, AND
- speaks English, AND
- is not enrolled in MATH 1710, AND
- has a single marital status, AND
- has a tenured professor for more than 33% of their courses, AND
- has a composite score of 24 or below on their ACT,

Then the student is predicted to **stay**.



**Figure 6: Portion of Decision Tree Model for disabled Students from the root to the first set of leaves.**

The decision tree model for Hispanic students has 67 leaves and 133 total nodes.

A portion of these students can also be classified by the use of only three attributes:

- 1) Institutional GPA, if MTSU was selected as their second choice, and their score on the Math portion of the ACT being below 15

**OR**

- 2) Institutional GPA, if MTSU was not selected as their second choice, and if they are receiving any financial aid.

The steps of the decision tree for these groups of students are explained below and represented by Figure 7. The complete decision tree model for Hispanic students can be found in the Appendix on page 66.

If a student:

has a GPA less than or equal to 2.231, AND  
does not have MTSU listed as their second choice, AND  
has a score of 15 or below on the Math section of their ACT,  
Then the student is expected to **stay** at the university.  
has a score higher than 15 on the Math section of their ACT,

*... see the full decision tree model to see this result*

has MTSU listed as their second choice, AND  
is receiving no financial aid,  
Then the student is expected to **stay** at the university.  
is receiving some amount of financial aid,  
Then the student is expected to **drop** from the university.

```
INST_TGPA_GPA <= 2.231
|   Second_Choice <= 0
|   |   SZBSTER_ACT_MATH <= 15: 0Stayed (10.0/1.0)
|   |   SZBSTER_ACT_MATH > 15
|
|   ...
|
|   Second_Choice > 0
|   |   AnyAid <= 0: 0Stayed (2.0)
|   |   AnyAid > 0: 2Dropped (13.0)
```

**Figure 7: Portion of Decision Tree Model for Hispanic Students that can be classified using only 3 attributes.**

## 3.4 Model Validation

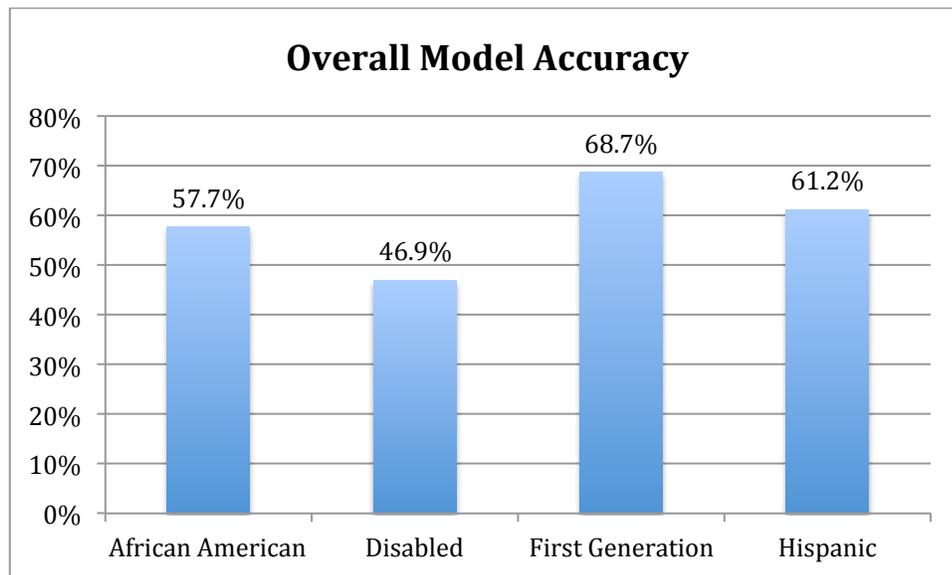
### 3.4.1 Cross-Validation of Existing Models

After the models were created, we wanted to know how well they performed when classifying new, never before seen data. To determine the accuracy of our classification models, we performed ten fold cross validation on each of the decision tree models using a 90% training set and 10% testing set. The accuracy results were averaged across 10 iterations for each model. We used the following criteria to evaluate the

performance of the model: True Positive (TP Rate), False Positive (FP Rate), Precision, and Recall.

The True Positive (TP) rate is the proportion of examples which were classified as class  $x$ , among all examples which truly have class  $x$ , i.e. how much part of the class was captured. It is equivalent to Recall. The False Positive (FP) rate is the proportion of examples which were classified as class  $x$ , but belong to a different class, among all examples which are not of class  $x$ . The Precision is the proportion of the examples which truly have class  $x$  among all those which were classified as class  $x$ . ( $x$  may refer to class stayed, class transferred, or class dropped). Recall measures the percentage of all the data belong to a class that are labeled correctly (Li, Wu, Wallin, & Hains, 2015).

An overview of the success rates of the models can be found in Figure 8 below, and more descriptive explanations can be found at the end of each section.



**Figure 8: Overview of accuracy for each minority group model.**

The decision tree model derived from the African American student data was able to classify correctly 57.7% of the tested instances. However, this means that 42.3% of students were classified incorrectly. The model does best at classifying students who will

be staying at the university. 72.6% of students who actually stayed at the university were classified correctly. This also means that 27.4% of students who actually stayed at the university were incorrectly classified as either transferred or dropped by the model. In practice, this would mean that 27.4% of students who will stay at the university may falsely end up on a university watch-list of students who are at higher risk of dropping or transferring from the school. However, this situation would not be detrimental, as it would be seen as being overly precautionary. An more undesirable result would be to classify someone who will drop from the university as stayed. This student would likely benefit from an intervention but would not be put on a watch-list. This means that we would like to minimize False Positive results for the stayed class. This model has a 47.7% FP rate for the stayed class. The model is not very accurate in predicting students who will transfer from the university; only 14% of students who actually transfer from the university were classified correctly.

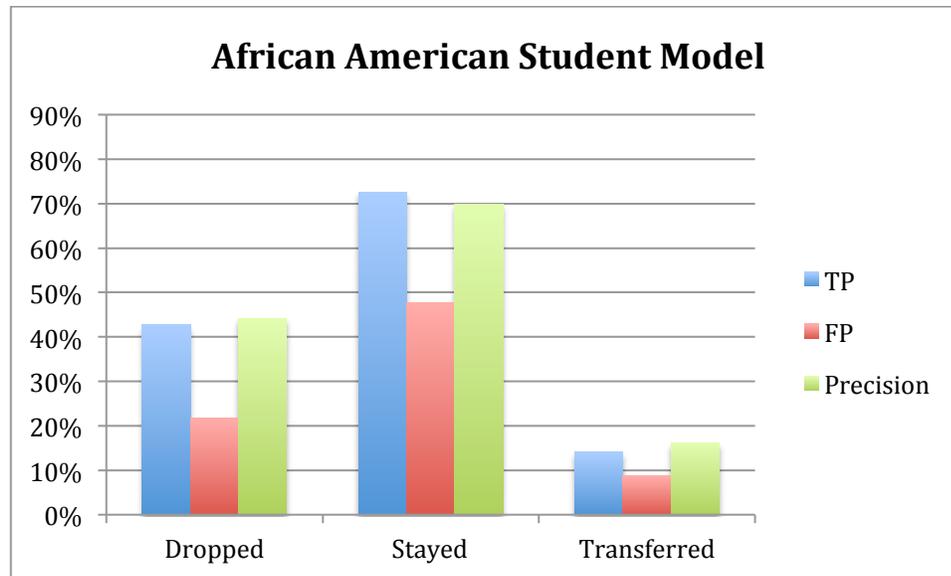
**Table 10: Performance of the decision tree model for the African American student group.**

Class	TP	FP	Precision	Recall
Dropped	0.427	0.218	0.441	0.427
Stayed	0.726	0.477	0.699	0.726
Transferred	0.141	0.088	0.161	0.141

**Table 11: Classification results of the decision tree model for the African American student group.**

		Classified as		
		Dropped	Stayed	Transferred
Actual	Class			
	Dropped	430	457	120
	Stayed	423	1536	156
	Transferred	121	203	53

 Correctly Classified



**Figure 9: Detailed accuracy for the African American student model.**

The first generation students’ decision tree model was able to classify correctly 68.7% of the tested instances, slightly better than the model for African American students. This model also does a good job of classifying students who will be staying at the university. About 86.1% of students who will stay at the university were classified correctly, the highest percentage of any of the derived models. 78.2% of students who were classified as stayed actually continued enrollment at the university. Unfortunately, this model results in a 57.1% false positive rate for the stayed class. About 57% of students who drop or transfer from the university are incorrectly classified as stayed with this model. However, 36% of students who do drop from the university are classified correctly. This means that over 30% of students who will drop out could be placed on a watch-list and have a chance for intervention from the university. While we would love this percentage to be higher, having the chance to help 30% of students who will drop out is definitely a positive.

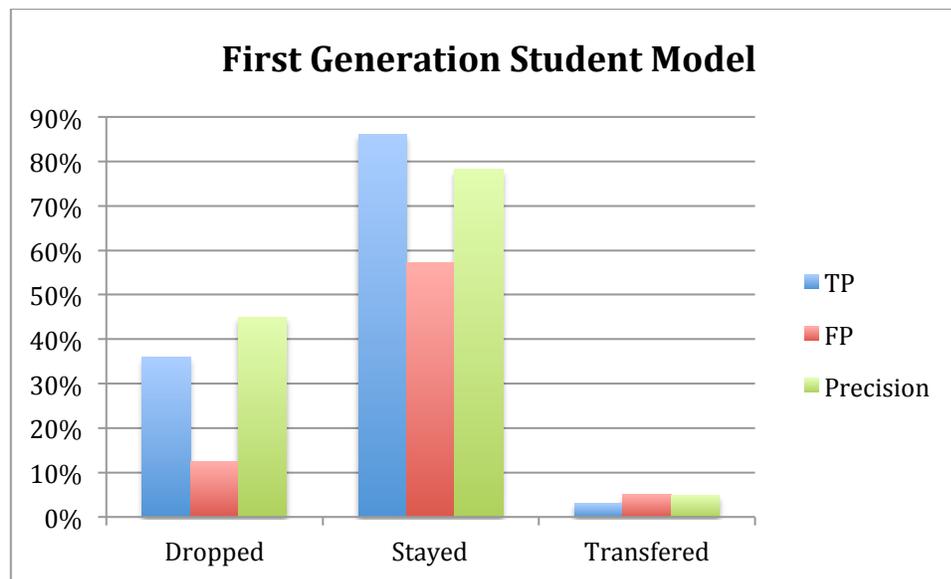
**Table 12: Performance of the decision tree model for the first generation student group.**

Class	TP	FP	Precision	Recall
Dropped	0.36	0.124	0.45	0.36
Stayed	0.861	0.571	0.782	0.861
Transferred	0.031	0.051	0.048	0.031

**Table 13: Classification results of the decision tree model for the first generation student group.**

		Classified as		
		Dropped	Stayed	Transferred
Actual	Class			
	Dropped	68	102	19
	Stayed	63	519	21
	Transferred	20	43	2

 Correctly Classified



**Figure 10: Detailed accuracy for the first generation student model.**

Using the decision tree model for disabled students, 46.9% of students in this minority group were classified correctly. This is the lowest overall accuracy rate. However, this model has the lowest FP rate for the stayed class at 39%. Students who

dropped were classified correctly 36.9% of the time, and transfer students were classified correctly 25.8% of the time. All of these students would end up on a watch-list with the chance of intervention. This is the most successful model for predicting students who will transfer from the university.

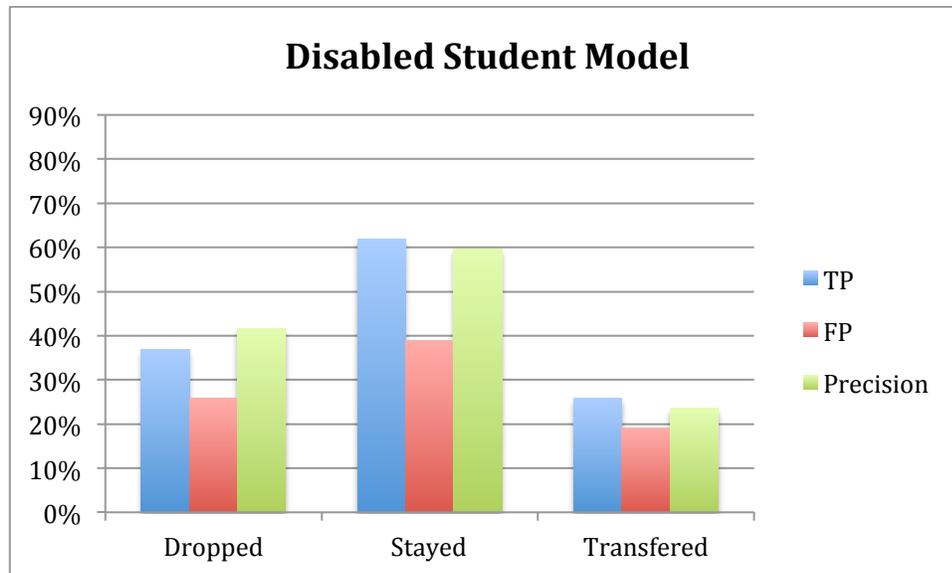
**Table 14: Performance of the decision tree model for the disabled student group.**

Class	TP	FP	Precision	Recall
Dropped	0.369	0.259	0.415	0.369
Stayed	0.619	0.39	0.596	0.619
Transferred	0.258	0.192	0.235	0.258

**Table 15: Classification results of the decision tree model for the disabled student group.**

		Classified as		
		Dropped	Stayed	Transferred
Actual	Class			
	Dropped	59	62	39
	Stayed	52	143	36
	Transferred	31	35	23

 Correctly Classified



**Figure 11: Detailed accuracy for the disabled student model.**

The decision tree model derived from the Hispanic student data was able to classify correctly 61.2% of the tested instances. Students who drop from the university were classified correctly 32.8% of the time. This is the lowest rate of any of the models for successfully classifying dropped students. Students who are retained were classified correctly over 78% of the time. However, there was also a 61% false positive rate for the stayed class, making it the least successful model for the stayed FP rate. This model was not able to classify correctly any students who transfer from the university. Although the overall classification success rate is relatively high, this model is the least successful for detecting students who will drop or transfer from the university.

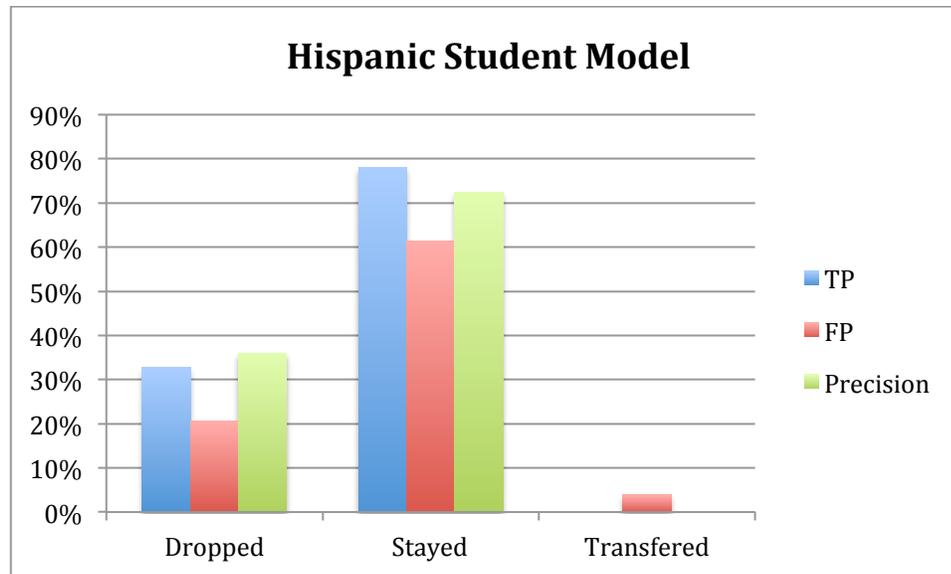
**Table 16: Performance of the decision tree model for the Hispanic student group.**

Class	TP	FP	Precision	Recall
Dropped	0.328	0.206	0.36	0.328
Stayed	0.781	0.613	0.725	0.781
Transferred	0	0.039	0	0

**Table 17: Classification results of the decision tree model for the Hispanic student group.**

		Classified as		
		Class	Dropped	Stayed
Actual	Dropped	59	107	14
	Stayed	91	363	11
	Transferred	14	31	0

 Correctly Classified



**Figure 12: Detailed accuracy for the Hispanic student model.**

### 3.4.2 Creating a New Model

Because the model derived from the Hispanic student data was the least successful, we decided to create a new model. The inconsistency of the model could be due to over-fitting of the model on the training data. To control for this, I reduced the data set to include only the top predictive attributes. I removed all of the attributes that did not appear as predictive during the attribute selection step and formed a new, smaller data set. I then created a new decision tree model with this data set and performed another round of ten fold cross validation.

The decision tree model for the reduced Hispanic student data set is much simpler than the original model. This model contains 50 leaves and 90 total nodes compared to the previous model's 67 leaves and 133 nodes. The deepest leaf on the new tree is 11 nodes, while the original tree is up to 18 nodes deep. The full model can be found in the appendices on page 67.

Overall, the new model is more successful than the original. Instances were classified correctly at a rate of 67.4%, while the old model was only 61.2% successful. Although the model is more successful overall, it is less successful for predicting students who will drop from the university. The model has a 67.6% false positive rate for the stayed class and only a 31.1% true positive for the dropped class. The model derived from the full set of Hispanic student data was able to achieve a higher 32.8% true positive rate for dropped students and a lower 61.3% false positive for stayed students. The over prediction of stayed students is something that we would like to avoid, so we will continue to use the model derived from the full data set.

**Table 18: Performance of the decision tree model for the reduced Hispanic student group.**

Class	TP	FP	Precision	Recall
Dropped	0.311	0.122	0.475	0.311
Stayed	0.88	0.676	0.729	0.88
Transferred	0	0.017	0	0

**Table 19: Classification results of the decision tree model for the reduced Hispanic student group.**

		Classified as		
		Dropped	Stayed	Transferred
Actual	Class			
	Dropped	10	117	7
	Stayed	52	409	4
	Transferred	10	35	0

 Correctly Classified

# Chapter 4

## Discussion

### 4.1 MTSU's Current Initiatives

This research lends well to one of the major initiatives that MTSU has taken on over the past few years - the Quest for Student Success. MTSU has been focused on student success throughout its history as an institution, but has defined a comprehensive initiative to improve retention rates that is set to be carried out between 2013-2016. This strategic focus on student success specifically addresses the new and changing challenging economic environment, dwindling annual budgets of the university, and the Complete College Tennessee Act of 2010, which is focused on educational outcomes (Middle Tennessee State University, 2013). The initiative can be defined more clearly by its three goals:

- I. Middle Tennessee State University will recruit students who value student success and have the potential to achieve in a student-centered culture.
- II. Middle Tennessee State University will enhance the academic experience of students to better ensure their success.
- III. Middle Tennessee State University will facilitate student success through innovation and the use of data-informed best practices.

MTSU's implementation of the Quest for Student Success initiative shows the university's

forward-thinking and strong desire to improve the environment of student retention at our university. This data-mining study lends itself specifically to Goal III of the Quest for Student Success plan which acknowledges the importance of "innovation and the use of data-informed best practices" (Middle Tennessee State University, 2013). MTSU has committed to "collect, analyze, and distribute data analyses to all student success stakeholders for use in making decisions in processes and practices related to student success" (Middle Tennessee State University, 2013).

MTSU also specifically acknowledges that the most at-risk students of failing or dropping out of school are Freshmen. The university has set to establish a Freshmen Involvement initiative that connects freshmen and sets expectations for their first year of study. By also discovering a set of freshman who are specifically at-risk, the university can continue to pinpoint students who are more likely to fail to continue enrollment at MTSU and possibly help them before it is too late.

The data-mining research completed in this study of first year students adds to the collection of analyzed student data being used to improve retention rates at MTSU in hopes of assisting the university in its efforts of maintaining enrollment of current students, specifically minority students after their first year.

## **4.2 Conclusions**

Overall, first generation students are retained at the highest rate when the student has a high GPA. However, first generation students with lower GPAs drop at the highest rate. Both African American and Hispanic students with lower GPAs are retained at a

higher rate than any of the other minority groups at the same GPA range. Disabled students consistently have higher drop out and transfer rates than the other target groups.

During this study, we have learned that the best predictors of a student's success are related to the student's institutional GPA and financial situation. Other high ranking predictive features that appear for most minority groups include: the percentage of courses from which the student withdraws (takes a grade of "W"), the math courses the student takes both in high school and in college, and the parent's education level.

The decision tree models are relatively different for each target minority group. The model developed for the African American student group is by far the most complex and largest model, due to the large sample size of students in this group. GPA attributes are the most highly predictive attributes and are the root node for every model, except for the model derived from the disabled student data. This model uses the percentage of classes withdrawn as its root node. Other nodes that appear towards the root of each tree include financial suspension/probation, ACT math scores, total parental income, and MTSU as a second choice. Overall, the decision tree generated for disabled students was the least successful, only classifying 46.9% of students correctly. The most successful model was for first generation students, classifying correctly 68.7% of the tested instances. However, the least successful model is actually the model created for the Hispanic student group. This model is able to classify correctly 61.2% of the tested instances but strongly overestimates the number of students who will stay at the university. Incorrectly classifying students into the stayed class will prevent us from detecting that they are at risk of attrition.

When compared to a chance prediction of one of the 3 classes (1 out of 3 chance of success; a 33.3% success rate), the model for the disabled student class may not be substantially more effective. However, the model for first generation students is over twice as successful as a chance model would be, leading us to believe that this model is relatively useful.

Predictive models are not always 100% accurate. There will always be outliers who do not fit within a specific formula. However, if these models were implemented to identify students at risk of leaving the university, as transfers or as drop outs, we could identify hundreds of minority students who are at potential risk.

# Appendices

## Appendix A

### Attribute Dictionary

This dictionary was compiled and created by Dr. Cen Li. It includes variable names and descriptions for TBR variables used in this study.

ACAD_PROGRESS,	Academic progress reported by the instructor early in the semester -> PC_PROG
Accomp1, <i>One Accomplishment</i>	0 No, 1 Yes
Accomp2, <i>Two Accomplishments</i>	0 No, 1 Yes
Accomp3, <i>Three or More Accomplishments</i>	0 No, 1 Yes
ACT_Disability, <i>Students with Disabilities</i>	0 No, 1 Yes
ACT_Work_Hours_21, <i>Worked 21 Hours or more</i>	0 No, 1 Yes
ACTIVE_DUTY, <i>Active Military Duty</i>	1 Yes, 2 No
ADMISSION_REQ	whether the program has candidacy requirements
Age	age
AnyAid, <i>Any Type of Aid</i>	0 No, 1 Yes
ANYPRESC, <i>Any Prescribed</i>	0 No, 1 Yes
AP, <i>AP Course Credits</i>	0 No, 1 Yes
ATHL, <i>Athletic Aid</i>	0 - no aid, more - aid
ATHL2, <i>Athletic Aid</i>	0 No Aid, 1 Received Aid
ATHLD2, <i>Dollar amount of Athletic Aid</i>	0 None
ATHLD2R, <i>Athletic Aid</i>	Should use PC_ATTENDANCE
ATTENDANCE,	Should use PC_ATTENDANCE
ATTENDANCE2,	Should use PC_ATTENDANCE
AttendFullPartTime, <i>ACT Status</i>	1 Full-Time, 2 Part-Time
Black	0 No, 1 Yes
Citizenship_Status	1 - US resident, 2 - temp resident, 3 - foreign citizen
College	IF (COLL_CODE = "BA") College = 1 . IF (COLL_CODE = "BH") College = 2 . IF (COLL_CODE = "BU") College = 3 . IF (COLL_CODE = "E") College = 4 . IF (COLL_CODE = "LA") College = 5 . IF (COLL_CODE = "MC") College = 6 . IF (COLL_CODE = "RO") College = 7 . IF (COLL_CODE = "00") College = 8 . IF (COLL_CODE = "UC") College = 9
Combined_Disability, <i>Disability</i>	0 No, 1 Yes
COMPASS_ENGLISH	The Compass level a student test into. There is a separate data explanation file to understand the different levels.
COMPASS_MATH	The Compass level a student test into. There is a separate data explanation file to understand the different levels.
COMPASS_READ	The Compass level a student test into. There is a separate data explanation file to understand the different levels.
COURSE_NUMBER	like 1170 in CSCI 1170
CRN_KEY	do not use it, unless you want to extract the students who took one section of CSCI 1170 together.
DDY1	definitely dropped(or transferred to college that does not share information), 1 -yes, 0 - no
DDY2	definitely dropped(or transferred to college that does not share information), 1 -yes, 0 - no
DDY3	definitely dropped(or transferred to college that does not share information), 1 -yes, 0 - no
DDY4	definitely dropped(or transferred to college that does not share information), 1 -yes, 0 - no
DDY5	definitely dropped(or transferred to college that does not share information), 1 -yes, 0 - no
DDY6	definitely dropped(or transferred to college that does not share information), 1 -yes, 0 - no
DEGC_CODE	IF (DEGC_CODE = "000000" ) DEGC_CODE2 = 2. IF (DEGC_CODE = "BA" ) DEGC_CODE2 = 3. IF (DEGC_CODE = "BBA" ) DEGC_CODE2 = 4. IF (DEGC_CODE = "BFA" ) DEGC_CODE2 = 5. IF (DEGC_CODE = "BM" ) DEGC_CODE2 = 6. IF (DEGC_CODE = "BS" ) DEGC_CODE2 = 7. IF (DEGC_CODE = "BSN" ) DEGC_CODE2 = 8. IF (DEGC_CODE = "BSW" ) DEGC_CODE2 = 9. IF (DEGC_CODE = "BUS" ) DEGC_CODE2 = 10. IF (DEGC_CODE = "GCR" ) DEGC_CODE2 = 13. IF (DEGC_CODE = "MA" ) DEGC_CODE2 = 14. IF (DEGC_CODE = "MACC" ) DEGC_CODE2 = 15. IF (DEGC_CODE = "MAT" ) DEGC_CODE2 = 16. IF (DEGC_CODE = "MBA" ) DEGC_CODE2 = 17. IF (DEGC_CODE = "MBE" ) DEGC_CODE2 = 18. IF (DEGC_CODE = "MCJ" ) DEGC_CODE2 = 19. IF (DEGC_CODE = "MED" ) DEGC_CODE2 = 20. IF (DEGC_CODE = "MFA" ) DEGC_CODE2 = 21. IF (DEGC_CODE = "MPS" ) DEGC_CODE2 = 22. IF (DEGC_CODE = "MS" ) DEGC_CODE2 = 23. IF (DEGC_CODE = "MSN" ) DEGC_CODE2 = 24. IF (DEGC_CODE = "MST" ) DEGC_CODE2 = 25. IF (DEGC_CODE = "MSW" ) DEGC_CODE2 = 26. IF (DEGC_CODE = "NDGD" ) DEGC_CODE2 = 27. IF (DEGC_CODE = "NDUG" ) DEGC_CODE2 = 28. IF (DEGC_CODE = "DA" ) DEGC_CODE2 = 29. IF (DEGC_CODE = "EDS" ) DEGC_CODE2 = 30. IF (DEGC_CODE = "PHD" ) DEGC_CODE2 = 31. IF (DEGC_CODE = "XXX" ) DEGC_CODE2 = 32.
DEPEND_INDEPEND_IND	do not use this, use D_STATUS instead, 1 - yes, 0 - no
Dependents	0 No, 1 Yes

	IF (DEPT_CODE1 = "AERO" ) DEPT = 2. IF (DEPT_CODE1 = "ABAS" ) DEPT = 3. IF (DEPT_CODE1 = "ART" ) DEPT = 4. IF (DEPT_CODE1 = "BA" ) DEPT = 5. IF (DEPT_CODE1 = "BIOL" ) DEPT = 6. IF (DEPT_CODE1 = "BCEN" ) DEPT = 7. IF (DEPT_CODE1 = "CHEM" ) DEPT = 8. IF (DEPT_CODE1 = "INFS" ) DEPT = 9. IF (DEPT_CODE1 = "CSCI" ) DEPT = 10. IF (DEPT_CODE1 = "CIM" ) DEPT = 11. IF (DEPT_CODE1 = "CIA" ) DEPT = 12. IF (DEPT_CODE1 = "DYST" ) DEPT = 13. IF (DEPT_CODE1 = "ECON" ) DEPT = 14. IF (DEPT_CODE1 = "E" ) DEPT = 15. IF (DEPT_CODE1 = "EDLR" ) DEPT = 16. IF (DEPT_CODE1 = "ELED" ) DEPT = 17. IF (DEPT_CODE1 = "ETIS" ) DEPT = 18. IF (DEPT_CODE1 = "ENGL" ) DEPT = 19. IF (DEPT_CODE1 = "FLL" ) DEPT = 20. IF (DEPT_CODE1 = "GEOS" ) DEPT = 21. IF (DEPT_CODE1 = "HHP" ) DEPT = 22. IF (DEPT_CODE1 = "HIST" ) DEPT = 23. IF (DEPT_CODE1 = "HSC" ) DEPT = 24. IF (DEPT_CODE1 = "LA" ) DEPT = 25. IF (DEPT_CODE1 = "MGMT" ) DEPT = 26. IF (DEPT_CODE1 = "MC" ) DEPT = 27. IF (DEPT_CODE1 = "MATH" ) DEPT = 28. IF (DEPT_CODE1 = "MUSI" ) DEPT = 29. IF (DEPT_CODE1 = "NURS" ) DEPT = 30.
DEPT	
DESIRED_DEGREE_TYPE, <i>Desired Degree Type</i>	1 1st Bachelors
Dstatus, <i>Independent Status</i>	0 No, 1 Yes
DY1, <i>Dropped in a Year Period</i>	Flags available going out six years including graduation
DY2	did not appear n years later (DDY + transfers)
DY3	did not appear n years later (DDY + transfers)
DY4	did not appear n years later (DDY + transfers)
DY5	did not appear n years later (DDY + transfers)
DY6	did not appear n years later (DDY + transfers)
ECCOLHIGH	FROM ACT readiness, 0(does not participate) , 1(participate), expected to be involved in extracurriculum activities in college
ECHSHIGH	FROM ACT readiness, 0(does not participate) , 1(participate), expected to be involved in extracurriculum activities in high school
ECY1	enrolled in the same college for n years
ECY2	enrolled in the same college for n years
ECY3	enrolled in the same college for n years
ECY4	enrolled in the same college for n years
ECY5	enrolled in the same college for n years
ECY6	enrolled in the same college for n years
EMY1	enrolled in the same major for n years
EMY2	enrolled in the same major for n years
EMY3	enrolled in the same major for n years
EMY4	enrolled in the same major for n years
EMY5	enrolled in the same major for n years
EMY6	enrolled in the same major for n years
ENGL1008	1 - taken, 0 - not taken
ENGL1009	0 No, 1 Yes
ENGL1009	1 - taken, 0 - not taken
ENGL1010	0 No, 1 Yes
ENGL1010	1 - taken, 0 - not taken
English_not_spoken, <i>English Second Language</i>	0 No, 1 Yes
EnglishPrimaryLang, <i>English Spoken</i>	1 Yes
ENGLP, <i>Prescribed English</i>	0 No, 1 Yes
ENGLP2, <i>Prescribed English</i>	0 No, 1 Yes
	with Internationals 9 Not Specified 1 Alaskan Native 2 American Indian 3 Asian 4 Black or African American 5 Hispanic 7 White 6 Native Hawaiian or Other Pacific Islander 8 Two Or More Races.
ETHN_DESC2	
ETHN_DESC3, <i>Race or Ethnicity</i>	
EUY0	We do not need this
EUY1	We do not need this
EUY2	We do not need this
EUY3	We do not need this
EUY4	We do not need this
EUY5	We do not need this
EUY6	We do not need this
EVENING_STUDENT, <i>Evening Student</i>	0 No, 1 Yes
F_HOLD, <i>Financial Hold</i>	0 No, 1 Yes
F_HOLD_ALL	Last hold number
FAMILY_IN_COLLEGE, <i>Family in College</i>	
FAMILY_SIZE, <i>Family Size</i>	
	From FASFA IF (FAMILY_SIZE = 1) FAMILY_SIZE2 = 1. IF (FAMILY_SIZE = 2) FAMILY_SIZE2 = 2. IF (FAMILY_SIZE = 3) FAMILY_SIZE2 = 3. IF (FAMILY_SIZE = 4) FAMILY_SIZE2 = 4. IF (FAMILY_SIZE >= 5) FAMILY_SIZE2 = 5.
FAMILY_SIZE2	
FATHER_HIGHEST_ED, <i>Fathers Highest Education Level</i>	1 Middle School/Jr. High School
FatherEdLevel, <i>Education Father</i>	1 Less than High School
FIN_PROB, <i>Financial Probation</i>	0 No, 1 Yes
FIN_SUS, <i>Financial Suspension</i>	0 No, 1 Yes
FINANCIAL_HOLD	do not use it
FINANCIAL_HOLD_ALL	do not use it
First_Choice	from ACT readiness data, MTSU was first choice
First_Gen, <i>First Generation</i>	0 No, 1 Yes

FM_Income_Missing, <b>Family Missing Income</b>	0 No, 1 Yes
FM_TOTAL_INCOME, <b>Family Total Income</b>	
FM_TOTAL_INCOMER, <b>Family Total Income</b>	
Foreign_Student, <b>Foreign Student</b>	0 No, 1 Yes
Foreign2, <b>Foreign Student</b>	0 No, 1 Yes
FRAT_SOR	whether involved in fraternity or sorority
FTPT_Status	0 Full-Time, 1 Part-Time
FTPTStatus	IF (FTPTStatus = "Part-Time") Status=1. IF (FTPTStatus = "Full-Time") Status=0.
G, <b>G</b>	subsidized/not subsidized from GPSU -> Loans, do not use, use G2 instead
G2, <b>G</b>	0 No Aid, 1 Received Aid
GCY1	graduated from the same college in n years
GCY2	graduated from the same college in n years
GCY3	graduated from the same college in n years
GCY4	graduated from the same college in n years
GCY5	graduated from the same college in n years
GCY6	graduated from the same college in n years
GD2R, <b>G</b>	0 None
GED, <b>GED</b>	0 No, 1 Yes.
Gender2, <b>Gender</b>	0 Male, 1 Female
GMV1	graduated from the same college in n years
GMV2	graduated from the same college in n years
GMV3	graduated from the same college in n years
GMV4	graduated from the same college in n years
GMV5	graduated from the same college in n years
GMV6	graduated from the same college in n years
GRADE_CODE	A, B, C
GRNT, <b>Grant</b>	do not use it, use GRant2 instead
GRNT2, <b>Grant</b>	0 No Aid, 1 Received Aid
GRNTD2, <b>Dollar amount of Grants</b>	
GRNTD2R, <b>Grant</b>	0 None
GROSS_INCOME_FAFSA, <b>Gross Income</b>	
GUY1	graduated from MTSU in n years
GUY2	graduated from MTSU in n years
GUY3	graduated from MTSU in n years
GUY4	graduated from MTSU in n years
GUY5	graduated from MTSU in n years
GUY6	graduated from MTSU in n years
HAS_LEGAL_DEPEND, <b>Dependent</b>	1 Yes, 2 No
HAVE_CHILDREN, <b>Have Children</b>	1 Yes, 2 No
HaveDisability, <b>Disability</b>	1 No disability reported
HelpEdPlans_ALL_areas, <b>3 Areas Help</b>	0 No, 1 Yes
HelpEdPlans_ONE_area, <b>1 Area Help</b>	0 No, 1 Yes
HelpEdPlans_TWO_areas, <b>2 Areas Help</b>	0 No, 1 Yes
HelpEducationPlans, <b>Help Education</b>	0 No, 1 Yes
HelpMathSkills	0 No, 1 Yes
HelpReading	0 No, 1 Yes
HelpStudySkills	0 No, 1 Yes
HelpWriting	0 No, 1 Yes
Highschool or lower	
HighSchoolAverage, <b>ACT High School GPA</b>	GPA from ACT readiness
HighSchoolCurriculum, <b>School Curriculum</b>	1 Business
HighSchoolGPA, <b>High School GPA</b>	1 0.5-0.9
HighSchoolType, <b>HS Type</b>	1 Public
HOME_ENVIR,	
Home_School, <b>Home School</b>	0 No, 1 Yes
HONORS, <b>Honors Student</b>	0 No, 1 Yes
HoursToWorkCollege, <b>Work Hours</b>	1 None
Housing	(HOUSING_IND = "N") Housing=0. IF (HOUSING_IND = "Y") Housing=1. on campus - 1, off campus - 2
HOUSING_CODE	1 live on campus, 0 live off campus
HOUSING_PARENT	1 lives with parents, 0 - without
HS_COMPLETION_STATUS, <b>High School Completion Status</b>	1 High School Diploma
HS_Curr_Collprep, <b>Coll Prep Curriculum</b>	0 No, 1 Yes
HS_ENG_Below_B, <b>English GPA &lt; 3.0</b>	0 No, 1 Yes
HS_GPA, <b>HS GPA</b>	HS GPA in Banner, i.e., high school transcript GPA
HS_GPA_A, <b>GPA 3.5-4.0</b>	0 No, 1 Yes
HS_GPA_B, <b>GPA 3.0-3.4</b>	0 No, 1 Yes
HS_GPA2, <b>HS GPA- Combined</b>	combined from HS_GPA and ACT high school GPA, if Banner does not have high school GPA, pull in the ACT high school GPA
HS_MATH_Below_B, <b>Math GPA &lt; 3.0</b>	0 No, 1 Yes
HS_MathLevel_High	
HS_MathLevel_Low	
HS_NATSCI_BELOW_B, <b>Natural Science Grade &lt; 3.0</b>	0 No, 1 Yes
HS_SOCSTUD_BELOW_B	0 No, 1 Yes
HSCourseAlgebra1	0 No, 1 Yes
HSCourseAlgebra2	0 No, 1 Yes
HSCourseBegCalc	0 No, 1 Yes
HSCourseGeometry	0 No, 1 Yes
HSCourseOtherAdvMath	0 No, 1 Yes
HSCourseTrig	0 No, 1 Yes
HSEngGrade, <b>HS English GPA</b>	
HSGradeEnglish	obtained from ACT readiness test data. Likely redundant to other features, do not use
HSGradeMath	obtained from ACT readiness test data. Likely redundant to other features, do not use
HSGradeSocStud, <b>HS Social Studies GPA</b>	obtained from ACT readiness test data. Likely redundant to other features, do not use
HSMathClasses, <b>HS Math Courses</b>	1 HSCourseAlgebra1
HSMathGrade, <b>HS Math GPA</b>	
HSocStudies	obtained from ACT readiness test data. Likely redundant to other features, do not use
Income_Level, <b>Income Level</b>	PARENT_1_TOT_INCOME; Missing data from midpoint of ACT ranges for the variable LevelParentsIncome
INST_LGPA_GPA	cummulative for MTSU

INST_TGPA_GPA, <i>Institutional Term GPA</i>	<p>IF (INST_TGPA_GPA &lt; .67 And INST_TGPA_HOURS&gt;0) GPA_Range = 1 .  IF (INST_TGPA_GPA &gt;= .67 And INST_TGPA_GPA &lt; 1) GPA_Range = 2 .  IF (INST_TGPA_GPA &gt;= 1 And INST_TGPA_GPA &lt; 1.33) GPA_Range = 3 .  IF (INST_TGPA_GPA &gt;= 1.33 And INST_TGPA_GPA &lt; 1.67) GPA_Range = 4 .  IF (INST_TGPA_GPA &gt;= 1.67 And INST_TGPA_GPA &lt; 2) GPA_Range = 5 .  IF (INST_TGPA_GPA &gt;= 2 And INST_TGPA_GPA &lt; 2.33) GPA_Range = 6 .  IF (INST_TGPA_GPA &gt;= 2.33 And INST_TGPA_GPA &lt; 2.67) GPA_Range = 7 .  IF (INST_TGPA_GPA &gt;= 2.67 And INST_TGPA_GPA &lt; 3) GPA_Range = 8 .  IF (INST_TGPA_GPA &gt;= 3 And INST_TGPA_GPA &lt; 3.33) GPA_Range = 9 .  IF (INST_TGPA_GPA &gt;= 3.33 And INST_TGPA_GPA &lt; 3.67) GPA_Range = 10 .  IF (INST_TGPA_GPA &gt;= 3.67 And INST_TGPA_GPA &lt; 4) GPA_Range = 11 .  IF (INST_TGPA_GPA = 4) GPA_Range = 12 .</p> <p>IF (INST_TGPA_GPA &gt;= 3.67) TGPA_LETTER = 1.  IF (INST_TGPA_GPA &gt;= 2.67 AND INST_TGPA_GPA &lt; 3.67) TGPA_LETTER = 2.  IF (INST_TGPA_GPA &gt;= 1.67 AND INST_TGPA_GPA &lt; 2.67) TGPA_LETTER = 3.  IF (INST_TGPA_GPA &gt;= .67 AND INST_TGPA_GPA &lt; 1.67) TGPA_LETTER = 4.  IF (INST_TGPA_GPA &lt; .67) TGPA_LETTER = 5.  Execute.</p> <p>IF (INST_TGPA_GPA &lt; 2.0 And INST_TGPA_HOURS &gt; 0) TGPA_UNDER_2.0 = 1.  IF (INST_TGPA_GPA &gt;= 2.0 And INST_TGPA_HOURS &gt; 0) TGPA_UNDER_2.0 = 0.  Execute.</p>
INSTAID, <i>Institutional Aid</i>	<p>Variable Labels INSTAID Institutional Aid.  Variable Labels INSTAID2 Institutional Aid.  Value Labels INSTAID2  0 No Aid  1 Received Aid .  Variable Labels INSTAID2R Institutional Aid .  Value Labels INSTAID2R  0 None  1 &lt; \$1,000  2 \$1,000 - \$1,999  3 \$2,000 - \$2,999  4 \$3,000 - \$3,999  5 \$4,000 - \$4,999  6 \$5,000 or more</p>
INSTAID2, <i>Institutional Aid</i>	0 No Aid, 1 Received Aid
INSTAID2R, <i>Institutional Aid</i>	0 None
INSTAID2, <i>Institutional Aid</i>	0 No, 1 Yes
International	0 No, 1 Yes
Large_High_School	high school classified as large school according to ACT standards
LEARNING	belong to the learning community (not living)
Learning_Community	belong to Raider Learning community (both learning and living community)
Legacy, <i>First Generation</i>	0 No, 1 Yes
Legacy2, <i>First Generation</i>	0 No, 1 Yes
LevelParentsIncome, <i>Parents Income</i>	1 Less than \$24,000
LEVL_CODE	UG, GR
LOAN, <i>Loan</i>	0 No Aid 1 Received Aid
LOAN_DEFAULT	0 No, 1 Yes
LOAN_DEFAULT_COHORT	0 No, 1 Yes
LOAN2, <i>Loan</i>	0 No Aid, 1 Received Aid
LOAND2, <i>Dollar amount of Loans</i>	
LOAND2R, <i>Loan</i>	0 None
Major	<p>IF (MAJR_CODE1 = "ACIN" ) Major = 10.  IF (MAJR_CODE1 = "LIBS" ) Major = 20.  IF (MAJR_CODE1 = "AESE" ) Major = 30.  IF (MAJR_CODE1 = "AESL" ) Major = 40.  IF (MAJR_CODE1 = "ADSU" ) Major = 50.  IF (MAJR_CODE1 = "ASTL" ) Major = 60.  IF (MAJR_CODE1 = "AERO" ) Major = 70.  IF (MAJR_CODE1 = "AEED" ) Major = 75.  IF (MAJR_CODE1 = "AGBS" ) Major = 80.  IF (MAJR_CODE1 = "ALLP" ) Major = 90.  IF (MAJR_CODE1 = "ANSC" ) Major = 100.  IF (MAJR_CODE1 = "ANTH" ) Major = 110.  IF (MAJR_CODE1 = "ARMA" ) Major = 120.  IF (MAJR_CODE1 = "ART" ) Major = 130.  IF (MAJR_CODE1 = "ARED" ) Major = 140.  IF (MAJR_CODE1 = "ARH" ) Major = 150.  IF (MAJR_CODE1 = "ATTR" ) Major = 160.  IF (MAJR_CODE1 = "AVAD" ) Major = 170.  IF (MAJR_CODE1 = "BIOC" ) Major = 180.  IF (MAJR_CODE1 = "BIOL" ) Major = 190.  IF (MAJR_CODE1 = "BUAD" ) Major = 200.  IF (MAJR_CODE1 = "BUED" ) Major = 210.  IF (MAJR_CODE1 = "CHEM" ) Major = 220.  IF (MAJR_CODE1 = "COMS" ) Major = 230.  IF (MAJR_CODE1 = "COSC" ) Major = 240.  IF (MAJR_CODE1 = "CIM" ) Major = 250.  IF (MAJR_CODE1 = "CM" ) Major = 260.  IF (MAJR_CODE1 = "CRJU" ) Major = 270.  IF (MAJR_CODE1 = "CUID" ) Major = 280.</p>
MARITAL_STATUS_FAFSA, <i>Marital Status</i>	1 Single
MATH1000	0 No, 1 Yes
MATH1010	0 No, 1 Yes
MATH1530	0 No, 1 Yes
MATH1710	0 No, 1 Yes
MATHK, <i>Lower Prescribed Math</i>	0 No, 1 Yes
MATHK2, <i>Lower Prescribed Math</i>	0 No, 1 Yes
MATHP, <i>Prescribed Math</i>	0 No, 1 Yes
MATHP2, <i>Prescribed Math</i>	0 No, 1 Yes
MaxOfCompositeScore, <i>Composite</i>	maximum composite ACT score
MaxOfEnglishScore, <i>English</i>	maximum ACT English score
MaxOfMathScore, <i>Math</i>	maximum ACT math score

MaxOfReadingScore, <i>Reading</i>	maximum ACT reading score
MaxOfScienceReasonScore, <i>Science</i>	maximum ACT science score
Medium_High_School	high school classified as medium school according to ACT standards
MOTHER_HIGHEST_ED, <i>Mothers Highest Education Level</i>	1 Middle School/Jr. High School
MotherEdLevel, <i>Education Mother</i>	1 Less than High School
Nontraditional, <i>Nontraditional Age</i>	0 Less than 24, 1 25 or older
NRTUITION, <i>Non-Resident Tuition</i>	0 No, 1 Yes
NTB, <i>Intl Bacc Course Credits</i>	0 No, 1 Yes
Off_Campus_Loc, <i>Off Campus</i>	0 No, 1 Yes
ONEPRESC, <i>One Prescribed</i>	0 No, 1 Yes
ONLINE_ONLY, <i>Online Only</i>	0 No, 1 Yes
Other_Minority	0 No, 1 Yes
P_Income_Missing, <i>Parent Missing Income</i>	0 No, 1 Yes
P2, <i>P</i>	0 No Aid, 1 Received Aid
PARENT_1_TOT_INCOME, <i>Parent Total Income</i>	
PARENT_1_TOT_INCOMER, <i>Parent Total Income</i>	1 <= 10000
Parents_Income_High, <i>High Income</i>	0 No, 1 Yes
Parents_Income_Middle, <i>Middle Income</i>	0 No, 1 Yes
PC_ATTEND_W_COURSES	% (percentage) of courses one withdraw 'W'
PC_ATTEND1, <i>Courses Attendance</i>	
PC_ATTEND2, <i>Courses Attendance</i>	
PC_DFWN_COURSES	% of courses end with grade DFWN (N are the special type of courses, considered difficult, students are allowed to take again wit
PC_Large_COURSES	% of courses have large student size
PC_Medium_COURSES	% of courses have medium student size
PC_ONLINE, <i>Courses Online</i>	Divided by total credit hours
PC_ONLINE_COURSES	% of courses are online courses (do not include RODP)
PC_ONLINE_COURSES_WRODP	% of courses are online courses including RODP
PC_PROG, <i>Courses Academic</i>	
PC_PROG_COURSES	% of courses have reported academic difficulties (per instruction early semester report)
PC_Small_COURSES	% of courses are considered small course
PC_TENURE_COURSES	% of courses are taught by tenured professors
PC_WITHDRAW, <i>Courses Withdrawn</i>	
PC_WITHDRAW_COURSES	% percentage of courses one withdraw
PCATHL, <i>Percent Athletic Aid</i>	
PCGRANT, <i>Percent Grant</i>	
PCLOAN, <i>Percent Loan</i>	
PCOL_ASSOCIATES, <i>Prior Associates</i>	0 No, 1 Yes
PCOL_BACHELORS, <i>Prior Bachelors</i>	0 No, 1 Yes
PCOL_GRADUATE, <i>Prior Graduate</i>	0 No, 1 Yes
PCSCHL, <i>Percent Scholarship</i>	
PCWORK, <i>Percent Work</i>	
PCWSCH, <i>Percent Work Scholarship</i>	
PD2R, <i>P</i>	0 None
	PELLD2R
	0 None
	1 < \$1,000
	2 \$1,000 - \$1,999
	3 \$2,000 - \$2,999
	4 \$3,000 - \$3,999
	5 \$4,000 - \$4,999
	6 \$5,000 or more
	So PELL is just an amount of Grant
PELL, <i>Pell Grant</i>	
PELL2, <i>Pell Grant</i>	0 No Aid, 1 Received Aid
PELLD2	0 No Aid
PELLD2R, <i>Pell Grant</i>	1 Received Aid
PELLD2R, <i>Pell Grant</i>	0 None
PGY1	Pending gratuation after 1st year
PGY2	Pending gratuation
PGY3	Pending gratuation
PGY4	Pending gratuation
PGY5	Pending gratuation
PGY6	Pending gratuation
PINSTAID, <i>Percent Institutional Aid</i>	
PlanPT, <i>Plan Attend Part-time</i>	0 Full-Time, 1 Part-Time
PRE_PROFESSIONAL, <i>Pre-professional Program</i>	0 No, 1 Yes
Private_High_School, <i>Private High School</i>	0 No, 1 Yes
Probation	0 No, 1 Yes
READ1000	0 No, 1 Yes
Readmitted	0 No, 1 Yes
READP2, <i>Prescribed Reading</i>	0 No, 1 Yes
Resident_Status, <i>Resident Status</i>	0 Out of State, 1 In State
RODP	0 No, 1 Yes
S, <i>Subsidized Loan</i>	do not use it, use S2 instead
S_Income_High, <i>Student High Income</i>	0 No, 1 Yes
S_Income_Middle, <i>Student Middle Income</i>	0 No, 1 Yes
S_Parent, <i>Single Parent</i>	0 No, 1 Yes
S2, <i>Subsidized Loan</i>	0 No Aid, 1 Received Aid
SCHL, <i>Scholarship</i>	do not use it, use Scholarship2 instead
SCHL2, <i>Scholarship</i>	0 No Aid, 1 Received Aid
SCHLD2, <i>Dollar amount of Scholarships</i>	
SCHLD2R, <i>Scholarship</i>	0 None
SD2R, <i>Subsidized Loan</i>	0 None
Second_BA, <i>Second Bachelors</i>	0 No, 1 Yes
Second_Choice	MTSU is his second choice college
Single	0 No, 1 Yes
Small_High_School	high school classified as small school according to ACT standards
Status, <i>Status</i>	0 Full-Time, 1 Part-Time
Student_Type2	IF (SZBSTER STUD_LEVEL = "01" And STYP_CODE = "N") Student_Type2=1.
STUDENT_WORKER, <i>On-campus Job</i>	IF (SZBSTER STUD_LEVEL = "01" And STYP_CODE = "A") Student_Type2=1.
SUBJ_CODE	major code, CSCI for computer science, etc.
Suspension	0 No, 1 Yes
SY1	first does not work at all, do not use the SY* features
SY2	suspention year
SY3	suspention year
SY4	suspention year

SYS	suspension year
SYS6	suspension year
SYS_DISABILITY	0 No, 1 Yes
SZBSTER_ACT_COMP, <i>ACT Composite</i>	SZBSTER are data sent to TBR, max ACT composite score
SZBSTER_ACT_ENGL, <i>ACT English</i>	max ACT English score
SZBSTER_ACT_MATH, <i>ACT Math</i>	max ACT math score
SZBSTER_ACT_READ, <i>ACT Reading</i>	max ACT reading score
SZBSTER_ACT_SCIEN, <i>ACT Science</i>	max ACT science score
SZBSTER_AGE, <i>Age</i>	do not use it
TERM_CODE_KEY	10-spring, 50-summer, 80-fall term
TERMGPA, <i>Term GPA</i>	
THREEPRES, <i>Three Prescribed</i>	0 No, 1 Yes
TotACMP, <i>Total Accomplishments</i>	Total number of awards received in high school, data obtained from the ACT data
TotAid, <i>Total Aid</i>	
TotAidR, <i>Total Aid</i>	0 None
TotalCreditHours, <i>Credit Hours</i>	
TRANSFER_GPA_ALL,	Do not use it
Transfer2yr, <i>Community College Transfer</i>	0 No, 1 Yes
TRY1, <i>Transferred in a 1 Year Period</i>	transferred within a year, transfer info found in the transfer house clearing house
TRY2	Transfer year, comes from LOAN file
TRY3	Transfer year, comes from LOAN file
TRY4	Transfer year, comes from LOAN file
TRY5	Transfer year, comes from LOAN file
TRY6	Transfer year, comes from LOAN file
TWOPRES, <i>Two Prescribed</i>	0 No, 1 Yes
U, <i>Unsubsidized Loan</i>	use U2 instead
U2, <i>Unsubsidized Loan</i>	0 No Aid, 1 Received Aid
UD2R, <i>Unsubsidized Loan</i>	0 None
Undeclared	0 No, 1 Yes
UNIV1010	Whether one takes UNIV 1010
UNMET_NEED, <i>Unmet Need</i>	
UNMET_NEED2, <i>Unmet Need</i>	0 No, 1 Yes
UNMET_NEEDR, <i>Unmet Need</i>	0 None
VET_FAFSA, <i>FAFSA U.S. Veteran</i>	1 Yes, 2 No
VETC_CODE	A, B, F, G, I, N, T
Veteran	0 No, 1 Yes
WAIVER, <i>Tuition Discount</i>	0 No, 1 Yes
WantHonorsCourses, <i>Honor Courses</i>	0 No, 1 Yes
WORK, <i>Work</i>	work study
WORK2, <i>Work</i>	0 No Aid, 1 Received Aid
WORKD2, <i>Dollar amount of Work</i>	
WORKD2R, <i>Work</i>	0 None
WSCH, <i>Work Scholarship</i>	Samount received from work study
WSCH2, <i>Work Scholarship</i>	0 No Aid, 1 Received Aid
WSCHD2, <i>Dollar amount of Work Scholarship</i>	
WSCHD2R, <i>Work Scholarship</i>	0 None
YEAR	year of the term

## Appendix B

### Decision Tree Models

These models were created using the Weka Data Mining software.

#### B.1 Decision Tree Model for African American students

```

1 | INST_LGPA_GPA <= 2.075
2 | | INST_LGPA_GPA <= 1.066
3 | | | Student_Type2 <= 1
4 | | | | Accompl1 <= 0
5 | | | | | PC_PROG_COURSES <= 25
6 | | | | | | PC_TENURE_COURSES <= 75
7 | | | | | | | Accompl2 <= 0
8 | | | | | | | | READ1000 <= 0
9 | | | | | | | | | SZBSTER_ACT_SCIEN <= 17
10 | | | | | | | | | | MATH1000 <= 0
11 | | | | | | | | | | | LOAND2 <= 1946: 1Transferred (4.0)
12 | | | | | | | | | | | | LOAND2 > 1946: 2Dropped (7.0/1.0)
13 | | | | | | | | | | | | MATH1000 > 0: 1Transferred (5.0)
14 | | | | | | | | | | | | | SZBSTER_ACT_SCIEN > 17
15 | | | | | | | | | | | | | | Second_Choice <= 0
16 | | | | | | | | | | | | | | | Dstatus <= 0
17 | | | | | | | | | | | | | | | | FIN_PROB <= 0
18 | | | | | | | | | | | | | | | | | MATH1000 <= 0
19 | | | | | | | | | | | | | | | | | | PC_DFWN_COURSES <= 22: 1Transferred (2.0)
20 | | | | | | | | | | | | | | | | | | | PC_DFWN_COURSES > 22
21 | | | | | | | | | | | | | | | | | | | | SZBSTER_ACT_ENGL <= 18: 1Transferred (5.0/1.0)
22 | | | | | | | | | | | | | | | | | | | | | SZBSTER_ACT_ENGL > 18
23 | | | | | | | | | | | | | | | | | | | | | | SZBSTER_ACT_MATH <= 25
24 | | | | | | | | | | | | | | | | | | | | | | | HelpEducationPlans <= 0
25 | | | | | | | | | | | | | | | | | | | | | | | | PC_Small_COURSES <= 75: 2Dropped (26.0)
26 | | | | | | | | | | | | | | | | | | | | | | | | | PC_Small_COURSES > 75
27 | | | | | | | | | | | | | | | | | | | | | | | | | | TotAid <= 8863: 1Transferred (2.0)
28 | | | | | | | | | | | | | | | | | | | | | | | | | | | TotAid > 8863: 2Dropped (4.0)
29 | | | | | | | | | | | | | | | | | | | | | | | | | | | HelpEducationPlans > 0
30 | | | | | | | | | | | | | | | | | | | | | | | | | | | | LEARNING <= 0

```





```

247 | | | | | MATH1710 > 0
248 | | | | | SCHLD2 <= 3742
249 | | | | | | Accompl <= 0
250 | | | | | | | UNMET_NEED2 <= 0
251 | | | | | | | UNMET_NEED <= -2757: 2Dropped (2.0)
252 | | | | | | | UNMET_NEED > -2757
253 | | | | | | | SZBSTER_ACT_COMP <= 17: 0Stayed (5.0/1.0)
254 | | | | | | | SZBSTER_ACT_COMP > 17: 1Transferred (8.0/1.0)
255 | | | | | | UNMET_NEED2 > 0
256 | | | | | | | PC_DFWN_COURSES <= 10: 1Transferred (2.0)
257 | | | | | | | PC_DFWN_COURSES > 10
258 | | | | | | | INST_LGPA_GPA <= 1.435: 1Transferred (4.0/1.0)
259 | | | | | | | INST_LGPA_GPA > 1.435
260 | | | | | | | SZBSTER_ACT_COMP <= 16: 0Stayed (4.0/1.0)
261 | | | | | | | SZBSTER_ACT_COMP > 16
262 | | | | | | | | FatherEdLevel <= 4: 2Dropped (24.0/1.0)
263 | | | | | | | | FatherEdLevel > 4
264 | | | | | | | | FAMILYSIZE <= 3: 0Stayed (3.0)
265 | | | | | | | | FAMILYSIZE > 3: 2Dropped (5.0)
266 | | | | | | Accompl > 0: 1Transferred (3.0/1.0)
267 | | | | | | SCHLD2 > 3742: 0Stayed (4.0)
268 | | | | | Nontraditional > 0
269 | | | | | | DEPT <= 21: 0Stayed (5.0)
270 | | | | | | DEPT > 21
271 | | | | | | | S_Parent <= 0: 2Dropped (11.0)
272 | | | | | | | S_Parent > 0
273 | | | | | | | UNMET_NEED <= 3917: 2Dropped (3.0/1.0)
274 | | | | | | | UNMET_NEED > 3917: 0Stayed (3.0)
275 | | | | | FIN_PROB > 0
276 | | | | | | INSTAIDD2 <= 749
277 | | | | | | | Large_High_School <= 0
278 | | | | | | | TotAid <= 9538
279 | | | | | | | | PC_PROG_COURSES <= 0
280 | | | | | | | | Accompl2 <= 0: 2Dropped (47.0/8.0)
281 | | | | | | | | Accompl2 > 0
282 | | | | | | | | Gender2 <= 0: 0Stayed (2.0)
283 | | | | | | | | Gender2 > 0: 2Dropped (2.0/1.0)
284 | | | | | | | | PC_PROG_COURSES > 0
285 | | | | | | | | MATH1710 <= 0
286 | | | | | | | | | HelpEducationPlans <= 0: 2Dropped (3.0)
287 | | | | | | | | | HelpEducationPlans > 0
288 | | | | | | | | | MATH1000 <= 0: 0Stayed (4.0)
289 | | | | | | | | | MATH1000 > 0: 2Dropped (3.0/1.0)
290 | | | | | | | | | MATH1710 > 0: 0Stayed (2.0)
291 | | | | | | | | TotAid > 9538: 1Transferred (3.0)
292 | | | | | | | Large_High_School > 0
293 | | | | | | | | PC_DFWN_COURSES <= 29: 1Transferred (4.0)
294 | | | | | | | | PC_DFWN_COURSES > 29: 2Dropped (3.0)
295 | | | | | | | INSTAIDD2 > 749: 0Stayed (3.0)
296 | | | | | FIN_SUS > 0
297 | | | | | | MATH1010 <= 0
298 | | | | | | | HSCourseOtherAdvMath <= 2
299 | | | | | | | | LEARNING <= 0
300 | | | | | | | | | INST_LGPA_GPA <= 1.756: 2Dropped (22.0/1.0)
301 | | | | | | | | | INST_LGPA_GPA > 1.756
302 | | | | | | | | | HSMathClasses <= 5
303 | | | | | | | | | | Nontraditional <= 0
304 | | | | | | | | | | MATH1710 <= 0: 2Dropped (8.0)
305 | | | | | | | | | | MATH1710 > 0: 0Stayed (3.0/1.0)
306 | | | | | | | | | | Nontraditional > 0: 0Stayed (2.0)
307 | | | | | | | | | | HSMathClasses > 5: 0Stayed (4.0)
308 | | | | | | | | | LEARNING > 0: 0Stayed (3.0)
309 | | | | | | | | HSCourseOtherAdvMath > 2
310 | | | | | | | | | PC_Large_COURSES <= 29: 2Dropped (7.0)
311 | | | | | | | | | PC_Large_COURSES > 29: 1Transferred (2.0)
312 | | | | | | | | | MATH1010 > 0: 0Stayed (2.0/1.0)
313 | | | | | | Student_Type2 > 11
314 | | | | | | | Combined_Disability <= 0: 2Dropped (26.0/2.0)
315 | | | | | | | Combined_Disability > 0: 0Stayed (3.0/1.0)
316 | | | | | INST_LGPA_GPA > 2.075
317 | | | | | | FIN_SUS <= 0
318 | | | | | | | Second_BA <= 0
319 | | | | | | | | Student_Type2 <= 9
320 | | | | | | | | FIN_PROB <= 0
321 | | | | | | | | | COMPASS_ENGLISH <= 2
322 | | | | | | | | | | HelpEdPlans_ALL_areas <= 0
323 | | | | | | | | | | PC_PROG_COURSES <= 0
324 | | | | | | | | | | LEARNING <= 0
325 | | | | | | | | | | | PC_Medium_COURSES <= 29
326 | | | | | | | | | | | COMPASS_MATH <= 3
327 | | | | | | | | | | | NRTUITIION <= 0
328 | | | | | | | | | | | Accompl <= 0
329 | | | | | | | | | | | | Citizenship_Status <= 2
330 | | | | | | | | | | | | INST_LGPA_GPA <= 3.031
331 | | | | | | | | | | | | | Single <= 0: 2Dropped (2.0)
332 | | | | | | | | | | | | | Single > 0
333 | | | | | | | | | | | | | | HAVE_CHILDREN <= 1: 1Transferred (2.0)
334 | | | | | | | | | | | | | | HAVE_CHILDREN > 1
335 | | | | | | | | | | | | | | COMPASS_ENGLISH <= 0: 1Transferred (3.0)
336 | | | | | | | | | | | | | | COMPASS_ENGLISH > 0
337 | | | | | | | | | | | | | | Accompl2 <= 0
338 | | | | | | | | | | | | | | | COMPASS_ENGLISH <= 1: 0Stayed (3.0/1.0)
339 | | | | | | | | | | | | | | | COMPASS_ENGLISH > 1
340 | | | | | | | | | | | | | | | | HelpWriting <= 0
341 | | | | | | | | | | | | | | | | | PC_ATTEND_W_COURSES <= 0
342 | | | | | | | | | | | | | | | | | | PC_Small_COURSES <= 86
343 | | | | | | | | | | | | | | | | | | HelpMathSkills <= 0
344 | | | | | | | | | | | | | | | | | | College <= 4
345 | | | | | | | | | | | | | | | | | | | FAMILY_IN_COLLEGE <= 1
346 | | | | | | | | | | | | | | | | | | | UNMET_NEED2 <= 0: 2Dropped (7.0)
347 | | | | | | | | | | | | | | | | | | | UNMET_NEED2 > 0
348 | | | | | | | | | | | | | | | | | | | | PC_DFWN_COURSES <= 50: 0Stayed (4.0/1.0)
349 | | | | | | | | | | | | | | | | | | | | PC_DFWN_COURSES > 50: 2Dropped (3.0)
350 | | | | | | | | | | | | | | | | | | | | FAMILY_IN_COLLEGE > 1: 0Stayed (5.0/1.0)
351 | | | | | | | | | | | | | | | | | | | | College > 4: 1Transferred (4.0/1.0)
352 | | | | | | | | | | | | | | | | | | | | HelpMathSkills > 0
353 | | | | | | | | | | | | | | | | | | | | MotherEdLevel <= 4: 1Transferred (4.0)
354 | | | | | | | | | | | | | | | | | | | | MotherEdLevel > 4: 0Stayed (2.0)
355 | | | | | | | | | | | | | | | | | | | | | PC_Small_COURSES > 86: 1Transferred (6.0/1.0)
356 | | | | | | | | | | | | | | | | | | | | | PC_ATTEND_W_COURSES > 0: 2Dropped (2.0)
357 | | | | | | | | | | | | | | | | | | | | | HelpWriting > 0: 2Dropped (2.0)
358 | | | | | | | | | | | | | | | | | | | | | Accompl2 > 0: 0Stayed (5.0)
359 | | | | | | | | | | | | | | | | | | | | | INST_LGPA_GPA > 3.031
360 | | | | | | | | | | | | | | | | | | | | | ANYPRES <= 0: 2Dropped (2.0/1.0)
361 | | | | | | | | | | | | | | | | | | | | | ANYPRES > 0
362 | | | | | | | | | | | | | | | | | | | | | | PRE_PROFESSIONAL <= 0
363 | | | | | | | | | | | | | | | | | | | | | | PC_Medium_COURSES <= 22
364 | | | | | | | | | | | | | | | | | | | | | | TotAid <= 8374: 0Stayed (30.0)
365 | | | | | | | | | | | | | | | | | | | | | | TotAid > 8374
366 | | | | | | | | | | | | | | | | | | | | | | | COMPASS_MATH <= 2: 1Transferred (5.0/1.0)
367 | | | | | | | | | | | | | | | | | | | | | | | COMPASS_MATH > 2: 0Stayed (2.0)
368 | | | | | | | | | | | | | | | | | | | | | | | PC_Medium_COURSES > 22: 1Transferred (2.0)
369 | | | | | | | | | | | | | | | | | | | | | | | PRE_PROFESSIONAL > 0

```











```

985 | | Second_BA > 0
986 | | COMPASS_READ <= 2: 0Stayed (4.0)
987 | | COMPASS_READ > 2
988 | | PCOL_ASSOCIATES <= 0
989 | | Undeclared <= 0
990 | | PC_WITHDRAW_COURSES <= 0
991 | | PARENT_1_TOT_INCOME <= 22607
992 | | PCOL_BACHELORS <= 0: 2Dropped (2.0)
993 | | PCOL_BACHELORS > 0
994 | | Resident_Status <= 0: 2Dropped (2.0)
995 | | Resident_Status > 0
996 | | DEPT <= 18: 0Stayed (7.0)
997 | | DEPT > 18
998 | | TotAid <= 4405: 0Stayed (3.0)
999 | | TotAid > 4405
1000 | | PC_TENURE_COURSES <= 86: 2Dropped (7.0)
1001 | | PC_TENURE_COURSES > 86: 0Stayed (3.0/1.0)
1002 | | PARENT_1_TOT_INCOME > 22607: 2Dropped (2.0)
1003 | | PC_WITHDRAW_COURSES > 0: 2Dropped (4.0/1.0)
1004 | | Undeclared > 0: 2Dropped (3.0/1.0)
1005 | | PCOL_ASSOCIATES > 0: 0Stayed (2.0)
1006 | FIN_SUS > 0
1007 | | Accom3 <= 0
1008 | | TWOPRESC <= 0
1009 | | PRE_PROFESSIONAL <= 0
1010 | | MATH1010 <= 0
1011 | | PCOL_ASSOCIATES <= 0
1012 | | GRNTD2 <= 1450
1013 | | FAMILY_IN_COLLEGE <= 1
1014 | | FAMILY_IN_COLLEGE <= 0: 2Dropped (3.0/1.0)
1015 | | FAMILY_IN_COLLEGE > 0: 0Stayed (12.0/2.0)
1016 | | FAMILY_IN_COLLEGE > 1: 2Dropped (5.0)
1017 | | GRNTD2 > 1450: 2Dropped (27.0/3.0)
1018 | | PCOL_ASSOCIATES > 0: 0Stayed (3.0/1.0)
1019 | | MATH1010 > 0: 2Dropped (5.0)
1020 | | PRE_PROFESSIONAL > 0
1021 | | TotalCreditHours <= 15: 2Dropped (2.0)
1022 | | TotalCreditHours > 15: 0Stayed (2.0/1.0)
1023 | | TWOPRESC > 0: 2Dropped (7.0)
1024 | | Accom3 > 0
1025 | | ECHSHIGH <= 0: 1Transferred (3.0/1.0)
1026 | | ECHSHIGH > 0
1027 | | SZBSTER_ACT_READ <= 21: 0Stayed (8.0)
1028 | | SZBSTER_ACT_READ > 21: 2Dropped (3.0/1.0)

```

## B.2 Decision Tree Model for first generation students

```

1 | INST_TGPA_GPA <= 1.692
2 | | PARENT_1_TOT_INCOME <= 82996: 2Dropped (77.0/13.0)
3 | | PARENT_1_TOT_INCOME > 82996
4 | | | SZBSTER_ACT_MATH <= 20: 1Transferred (6.0)
5 | | | SZBSTER_ACT_MATH > 20
6 | | | | PC_ATTEND_W_COURSES <= 0: 0Stayed (4.0/1.0)
7 | | | | PC_ATTEND_W_COURSES > 0: 2Dropped (2.0)
8 | INST_TGPA_GPA > 1.692
9 | | HOME_ENVIR <= 0
10 | | | INST_LGPA_GPA <= 2.667
11 | | | | FIN_SUS <= 0
12 | | | | | Private_High_School <= 0
13 | | | | | Student_Type2 <= 1
14 | | | | | | HS_GPA2 <= 2.617
15 | | | | | | | ADMISSION_REQ <= 0: 1Transferred (4.0)
16 | | | | | | | ADMISSION_REQ > 0: 2Dropped (3.0/1.0)
17 | | | | | | | HS_GPA2 > 2.617
18 | | | | | | | MATH1010 <= 0
19 | | | | | | | | English_not_spoken <= 0
20 | | | | | | | | PC_Medium_COURSES <= 20
21 | | | | | | | | | SZBSTER_ACT_ENGL <= 29
22 | | | | | | | | | HONORS <= 0
23 | | | | | | | | | HOUSING_PARENT <= 0
24 | | | | | | | | | Dependents <= 0
25 | | | | | | | | | | SZBSTER_ACT_COMP <= 20
26 | | | | | | | | | | | Other_Minority <= 0
27 | | | | | | | | | | | Large_High_School <= 0
28 | | | | | | | | | | | TWOPRESC <= 0
29 | | | | | | | | | | | | HelpEdPlans_ALL_areas <= 0
30 | | | | | | | | | | | | | LOAND2 <= 1594: 2Dropped (6.0)
31 | | | | | | | | | | | | | LOAND2 > 1594
32 | | | | | | | | | | | | | | HelpReading <= 0
33 | | | | | | | | | | | | | | | PC_TENURE_COURSES <= 29: 2Dropped (9.0/2.0)
34 | | | | | | | | | | | | | | | PC_TENURE_COURSES > 29: 0Stayed (8.0/1.0)
35 | | | | | | | | | | | | | | | HelpReading > 0: 0Stayed (5.0)
36 | | | | | | | | | | | | | | | HelpEdPlans_ALL_areas > 0: 2Dropped (5.0)
37 | | | | | | | | | | | | | | | TWOPRESC > 0: 0Stayed (9.0/1.0)
38 | | | | | | | | | | | | | | | Large_High_School > 0: 2Dropped (2.0)
39 | | | | | | | | | | | | | | | Other_Minority > 0: 0Stayed (2.0)
40 | | | | | | | | | | | | | | | SZBSTER_ACT_COMP > 20
41 | | | | | | | | | | | | | | | | HSCourseBegCalc <= 0: 2Dropped (3.0/1.0)
42 | | | | | | | | | | | | | | | | HSCourseBegCalc > 0
43 | | | | | | | | | | | | | | | | | LEGACY2 <= 0
44 | | | | | | | | | | | | | | | | | | MotherEdLevel <= 2
45 | | | | | | | | | | | | | | | | | | | HelpStudySkills <= 0: 2Dropped (3.0)
46 | | | | | | | | | | | | | | | | | | | HelpStudySkills > 0: 0Stayed (3.0/1.0)
47 | | | | | | | | | | | | | | | | | | | MotherEdLevel > 2: 0Stayed (4.0)
48 | | | | | | | | | | | | | | | | | | | LEGACY2 > 0: 0Stayed (25.0)
49 | | | | | | | | | | | | | | | | | | | Dependents > 0: 2Dropped (3.0)
50 | | | | | | | | | | | | | | | | | | | HOUSING_PARENT > 0
51 | | | | | | | | | | | | | | | | | | | Accompl <= 0

```







## B.4 Decision Tree Model for Hispanic students

```

1 INST_TGPA_GPA <= 2.231
2   Second_Choice <= 0
3     SZBSTER_ACT_MATH <= 15: 0Stayed (10.0/1.0)
4     SZBSTER_ACT_MATH > 15
5       ACT_Work_Hours_21 <= 0
6         PC_TENURE_COURSES <= 80
7           Undeclared <= 0
8             ENGL1009 <= 0
9               PC_TENURE_COURSES <= 17: 0Stayed (5.0/1.0)
10              PC_TENURE_COURSES > 17
11                HS_GPA2 <= 2.26: 2Dropped (10.0/1.0)
12                HS_GPA2 > 2.26
13                  MATH1000 <= 0
14                    FIN_SUS <= 0
15                      College <= 5
16                        Learning_Community <= 0
17                          MATH1010 <= 0
18                            PC_PROG_COURSES <= 20
19                              HelpReading <= 0
20                                SZBSTER_ACT_ENGL <= 19: 2Dropped (7.0/1.0)
21                                SZBSTER_ACT_ENGL > 19
22                                  FAMILY_SIZE2 <= 2: 2Dropped (7.0/1.0)
23                                  FAMILY_SIZE2 > 2: 0Stayed (23.0/4.0)
24                                  HelpReading > 0: 0Stayed (3.0/1.0)
25                                  PC_PROG_COURSES > 20: 2Dropped (3.0)
26                                  MATH1010 > 0: 0Stayed (3.0)
27                                  Learning_Community > 0: 0Stayed (5.0/1.0)
28                                  College > 5: 2Dropped (8.0/1.0)
29                                  FIN_SUS > 0: 2Dropped (8.0/1.0)
30                                  MATH1000 > 0: 2Dropped (2.0)
31                                  ENGL1009 > 0: 0Stayed (5.0/1.0)
32                                  Undeclared > 0
33                                    Housing <= 0: 2Dropped (5.0/1.0)
34                                    Housing > 0: 1Transferred (2.0)
35                                  PC_TENURE_COURSES > 80
36                                    PARENT_1_TOT_INCOME <= 19189: 0Stayed (2.0)
37                                    PARENT_1_TOT_INCOME > 19189: 1Transferred (2.0)
38                                  ACT_Work_Hours_21 > 0
39                                    TotalCreditHours <= 14: 1Transferred (4.0/1.0)
40                                    TotalCreditHours > 14: 2Dropped (9.0/1.0)
41          Second_Choice > 0
42            AnyAid <= 0: 0Stayed (2.0)
43            AnyAid > 0: 2Dropped (13.0)
44 INST_TGPA_GPA > 2.231
45   Student_Type2 <= 10
46     Single <= 0
47       Second_Choice <= 0
48         ONEPRESC <= 0
49           Veteran <= 0
50             PC_Large_COURSES <= 44
51               HelpEducationPlans <= 0
52                 FatherEdLevel <= 2: 2Dropped (2.0)
53                 FatherEdLevel > 2
54                   ADMISSION_REQ <= 0
55                     PCOL_ASSOCIATES <= 0
56                       LOAND2 <= 3034
57                         Gender2 <= 0
58                           Age <= 21: 2Dropped (3.0)
59                           Age > 21: 0Stayed (2.0)
60                         Gender2 > 0: 2Dropped (5.0)
61                       LOAND2 > 3034: 0Stayed (2.0)
62                     PCOL_ASSOCIATES > 0: 0Stayed (3.0)
63                   ADMISSION_REQ > 0: 0Stayed (15.0/1.0)
64                   HelpEducationPlans > 0: 0Stayed (4.0)
65                 PC_Large_COURSES > 44: 2Dropped (4.0)
66             Veteran > 0: 0Stayed (5.0)
67           ONEPRESC > 0: 0Stayed (7.0)
68         Second_Choice > 0: 2Dropped (2.0)
69     Single > 0
70       FIN_PROB <= 0
71         FAMILY_SIZE2 <= 1
72           HS_Curr_Collprep <= 0
73             ECHSHIGH <= 0
74               SZBSTER_ACT_ENGL <= 23
75                 Citizenship_Status <= 2: 0Stayed (32.0/6.0)
76                 Citizenship_Status > 2
77                   EVENING_STUDENT <= 0: 0Stayed (2.0)
78                   EVENING_STUDENT > 0: 2Dropped (2.0)
79                 SZBSTER_ACT_ENGL > 23: 2Dropped (2.0)
80             ECHSHIGH > 0: 2Dropped (2.0)
81           HS_Curr_Collprep > 0: 2Dropped (3.0)
82         FAMILY_SIZE2 > 1
83           FAMILY_IN_COLLEGE <= 2
84             COMPASS_ENGLISH <= 2
85             ECCOLHIGH <= 0

```





## Bibliography

- Aldenderfer, M.S., Blachfield, R.K., (1984). *Cluster Analysis*. Thousand Oaks, CA: SAGE Publications, Inc. <http://dx.doi.org/10.4135/9781412983648.n1>
- Arnold, K.E., Zeynep, T., & King, A.S., (2010). Administrative perceptions of data-mining software *signals*: promoting student success and retention. *The Journal of Academic Administration in Higher Education*, 6(2), 29-40.
- Bresciani, M. J., & Carson, L., (2002). A study of undergraduate persistence by unmet need and percentage of gift aid. *NASPA Journal*, 40(1), 104–123.
- Chandola, V., Banerjee, A., Kumar, V., (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3). New York, NY: ACM.
- Chen, R., & DesJardins, S. L. (2010). Investigating the impact of financial aid on student dropout risks: Racial and ethnic differences. *The Journal of Higher Education*, 81(2), 179–208.
- De Bruyne, S., & Plastria, F. (2010). Process, data and classifier models for accessible supervised classification problem solving. Brussels, Belgium : VUBPress, 2010.
- Federal Student Aid (2010). What are graduation, retention, and transfer rates? *U.S. Department of Education*. Retrieved from <https://fafsa.ed.gov/help/fotw91n.htm>
- Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques (2nd ed.)*. San Francisco, CA: Elsevier, Inc.
- Hastie, T., Tibshirani, R., Friedman, J. (2008). *Elements of statistical learning: Data mining, inference, and prediction (2nd ed.)*. Stanford, CA: Springer.

- Herzog, S., (2005). Measuring determinants of students returns vs dropout/stopout vs transfer: a first to second year analysis of new freshmen. *Research in higher education*, 46(8).
- Li, C., Wu, Q., Wallin, J., & Hein, M. (2015). Project report: Improving minority student success through data driven analysis.
- Middle Tennessee State University (2013). Retention rates 2012-2013. *Integrated Postsecondary Educational Data System*, pp. 6.
- Middle Tennessee State University (2013). Quest for student success 2013-2016. Retrieved from <http://www.mtsu.edu/docs/QuestforStudentSuccess.pdf>.
- Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning*. Cambridge, MA: MIT Press.
- National Center for Education Statistics (2014). Retention of first-time degree-seeking undergraduates at degree-granting postsecondary institutions: 2006 to 2012. Retrieved from [http://nces.ed.gov/programs/digest/d13/tables/dt13\\_326.30.asp](http://nces.ed.gov/programs/digest/d13/tables/dt13_326.30.asp)
- Ronco, S.L., Cahill, J., (2006). Does it matter who's in the classroom? Effect of instructor type on student retention, achievement and satisfaction. *AIR Professional File*, 100.
- Slotnik, W.J., Orland, M., (2010). Data rich but information poor. *Education Week*, 29. Retrieved from <http://www.edweek.org/ew/articles/2010/05/06/31slotnik.h29.html>
- Tinto, V., (1975). Dropout from higher education: a theoretical synthesis of recent research. *Review of educational research*, 45(1), 89-125.