I. Introduction

Trinity University's Office of Admissions is responsible for managing each applicant's admission process and materials. The goal of this project is to construct a classification model that will determine whether an accepted applicant will decide to attend the University or not. The information collected in the dataset pertains to applicants ranging in entry term from Fall 2017 to Fall 2021. The response variable in this dataset is "Decision," with 1 representing acceptance of the offer and 0 representing non-acceptance.

The importance of this project lies in the need for the University's admissions team to increase the yield of accepted applicants. By constructing a classification model, we can generate predictions regarding the likelihood of accepting the offer, and insights gained from this project can inform future recruitment and admissions strategies, enabling the Office of Admissions to better understand which factors are most important to students in their decision-making process. This information can be used to make more informed decisions about marketing and outreach efforts, as well as to tailor the admissions process to better meet the needs and preferences of prospective students. It is generally crucial for the university's financial health and overall success, because universities rely on a certain number of students to enroll each year to meet their revenue goals and maintain their academic programs and campus facilities.

II. Brief Description of the Original Dataset

The original dataset contains information on applicants who were accepted to Trinity University for the Fall entry term from 2017 to 2021. The dataset includes 15,144 observations and 69 variables. Each row corresponds to an applicant, and each column represents a different characteristic or attribute of the applicant. The response variable, which is the variable that we are interested in predicting, is the "Decision" column. The Decision column indicates whether the applicant accepted the offer of admission (a value of 1) or declined it (a value of 0).

The dataset includes various types of predictors, including demographic information (such as sex, race, and ethnicity), academic information (such as GPA and test scores), and information related to the admissions process (such as application submission date and source of application). In addition, there are many other important variables in the dataset that could potentially be useful in predicting an applicant's decision to attend Trinity University. For example, the "Count of Campus Visits" column indicates whether an applicant visited Trinity's campus for a daily visit, overnight stay, athletic camp, on-campus interview, or group visit. The "Merit Award" column provides information on the amount of financial aid that an applicant was awarded, and the "ACT" and "SAT" columns provide information on the applicant's standardized test scores.

One important note about the dataset is that in Fall 2021, Trinity University made standardized test scores optional for applicants. As a result, some applicants chose not to submit their test scores, and their scores

are not included in the dataset. This may have an impact on the accuracy of our predictive model, as test scores have historically been a strong predictor of whether an applicant will decide to attend Trinity University, and I will explain my strategy with this later in the next section.

Generally, the dataset contains a wealth of information that could potentially be used to construct a predictive model for an applicant's decision to attend Trinity University. By analyzing the data and building a model, the Office of Admissions at Trinity University could gain valuable insights into factors that influence an applicant's decision and use this information to increase the yield of accepted students who choose to attend the University.

III. Cleaning the Dataset

Before starting cleaning the dataset, I noticed that there are many blanks in the data, so I make blanks read as NA to make it more convenient to deal with missing values.

1) Handling missing values

- Categorical variables: For missing values in categorical variables, I replace the missing values by the most frequent level of the variable. For example, for **Permanent.Geomarket** variable, there was 1 NA value, so I use summary() function to determine which level is the most frequent level out of all levels within the variable, then I imputed this value to NA value. Additionally, there are some missing values of categorical variables, for example, **Academic.Interest.1** that are related to other variables, or are intentionally left blank and can be self-explanatory. To deal with this situation, I assign the value of the corresponding variable to the missing values, or I will determine the keyword based on my understanding and impute a new value for the missing values. For example, for **Academic.Interest.1**, there were 6 NA values, and most of the NA values actually have a value for **Academic.Interest.2**. Therefore, I assign the corresponding values in **Academic.Interest.2** to NA values of **Academic.Interest.1**. When a student does not even have **Academic.Interest.2**, it is likely that they have not decided their major interests yet, therefore, I find it reasonable to impute "Undecided" to the rest of the missing values. Examples are provided below.
- Numeric variables: One method I use to deal with missing values of numeric variables is to impute the mean value. One example is the standardize scores including ACT and SAT, either of which can be used to apply for Trinity. For convenience purposes, I decided to only use ACT scores for assessing and evaluating. Most of the missing values of **ACT.Composite** is because the student submitted their SAT scores instead of ACT scores, hence I converted all SAT scores into ACT scores based on the ACT-SAT-Concordance Table file provided on Tlearn. After this

step, most of the missing values of **ACT.Composite** were already resolved; however, given that starting Fall 2021, Trinity University made standardized test scores optional for applicants, and a lot of students applied for Trinity without ACT or SAT, there are still a lot of missing values in **ACT.Composite** to deal with. I calculated the mean value of exisiting values of **ACT.Composite**, then impute it to missing values. Finally, I dropped all columns related to SAT scores, and ACT partial scores (only keeping **ACT.Composite)** for standardize scores evaluation. Code example is provided below.

2) Correcting typos or measurement errors

For each categorical variable, I always use **summary()** function and check the data description to see if there are any typos of the categories. If typos are identified, I would replace them with the correct values. For example, for **Merit.Award,** there are a lot of categories that do not match the variable description because the description overlooked some levels. For example, for TT category (Trinity Tiger Award Scholarship), there were no "TT10", "TT12", or "TT125" in the description, and the description obviously overlooks those levels, because the website states that this scholarship is up to $12,000. For this specific situation, I did not base on the description file to resolve the "typos" of the categories, and I originally group levels based on the letter (which indicates what type of scholarship that they are classified as). However, given that even if combined, there are still not a lot of observations in each level, and after careful consideration for aliased issues later in the modeling stage, I decided to recategorize them into three main levels: Full Ride, Non Full Ride and No Merit.

3) Recoding and/or augmenting existing variables

In the cleaning stage, I also recode or augment some existing variables. As mentioned above, when dealing with "confusing" levels of **Merit.Award** and when observations in each original level only account for a small number of the whole dataset, I decided to recategorize them into three levels. I did the same thing for **Permanent.Geomarket** since there are a lot of levels of **Permanent.Geomarket** contaning very small number of observations out of 15143 observations, which will not provide much insightful information. I re-grouped them based on Census Regions information file of the U.S. provided on Tlearn, and classified foreign countries as "International"; however, after this step, there are still a lot of undefined categories and their information was not provided in the file. By researching, I found out that they are classified as unincorporated territory (US-GU refers to Guam for example), meaning the island is controlled by the U.S. but is separate from the mainland. Additionally, they do not account for the large number of values in the dataset, so I think it is reasonable to group them into "International" level as well. Example is provided below.

4) Dealing with outliers and Transformations

Since most of the variables of the dataset are categorical variables, I did not encounter outlier issues, and I did not have to do a lot in terms of transformations either. One example is when I clean **School.1.GPA.Recalculated** and checked for skewness, it turned out to be moderately skewed. However, the left skew is understandable because a lot of students got into Trinity with a high GPA, almost 4.0, so I think that it is unnecessary to do any transformations. Example is below.

5) Creating new variables

In the cleaning stage, I also create new variables. For example, when viewing column 11-13, which includes **First_Source.Origin.First.Source.Date** (the date Trinity was given by a third party (such as the College Board or others) that the person was seeking admission to college), **Inquiry.Date** (the date the person inquired about admissions to TU), and **Submitted** (the date the application is submitted), I think it would be more interesting to see how the time differences between First_Source date and submission date, or maybe the time differences between submission date and inquiry date affect Decision the response variable. Therefore, I created new variables to calculate the time differences between each pair. One issue I encountered was that there are NAs values in **Inquiry.Date**, which leads to NAs in the new variables it was related to, so imputed NA those new variables with median values. Example is provided below.

6) Removing variables

After creating new variables, I remove the existing variables because they have been combined into new variables. For example, as mentioned aboved, I removed column **11-13** after they are being used to calculate the time differences (Submit_Inquiry and Submit_FirstSource). Another example is **Permanent.Country**, which was eventually removed because I already have the information needed from **Permanent.Geomarket**.

I also removed all variables related to SAT scores, since I was using ACT scores as the main academic factor for evaluation. I also removed variables that are ACT partial scores, such as **ACT.English**, **ACT.Math**, etc since **ACT.Composite** is enough to tell about the student's performance on this standardized test.

Additionally, I encountered aliased coefficients issue after running the logistic model. I use vif() and alias() function to identify the culprits, and eventually removed **Sport.1.Rating** and **Sport.1.Sport** to avoid aliased issues.

IV. Report Kappa Scores from Implemented Models

Model	Kappa
Logistic Regression	0.5350381
KNN	0.1077172

Simple Classification Tree	0.3849844
Pruned Tree	0.4192191
Bagging	0.4740288
Random Forest	0.4991938
Boosting	**0.5384955**
SVM with Linear Kernel	0.449827
SVM with Polynomial Kernel	0.4671609
SVM with Radial Kernal	0.3288783

Among all the techniques, Boosting yields the highest Kappa score and seems to give us the best model.

V. Insights

After determining Boosting gives out the best model, I look at summary() of the method to identifying important predictors in the dataset. The summary() function produces a relative influence plot and also outputs the relative influence statistics of each variable in the model. For "rel.inf", the higher the rel.inf of a variable is, the more important that variable is in the boosted classification tree. Based on this, top factors that are the most important and significantly affect a student's decision to attend Trinity or not are:

- Total Event Participation: This variable indicates whether an applicant has attended any of the University's open house events or other non-campus visits events. It is important because it reflects the level of interest that an applicant has in the University, and their willingness to engage with the University's community. Applicants who have attended such events may have a better understanding of the University's culture and may be more likely to enroll.
- Decision Plan: This variable indicates whether an applicant applied through early action, regular decision, or another type of decision plan. It is important because it may reflect an applicant's level of interest in the University. Early action applicants, for example, may be more likely to enroll because they have made a commitment to the University earlier in the application process.
- Count of Campus Visits: This variable indicates whether an applicant has visited the University's campus for various activities such as a daily visit, overnight stay, athletic camp, on-campus interview, or group visit. It is important because it reflects an applicant's level of engagement with the University, and their willingness to take the time and effort to visit the campus.

Applicants who have visited the campus may have a better understanding of the University's resources and facilities, which may increase their likelihood of enrollment.

- Athlete: This variable indicates whether an applicant is an athlete or not. It is important because it may reflect an applicant's level of interest in the University's sports programs. Athletes may be more likely to enroll because of their desire to participate in the University's athletic programs, or because of the scholarships or other benefits offered to athletes.
- Second Academic Interest: This variable indicates an applicant's second choice of academic interest, after their primary interest. It is important because it may reflect an applicant's level of interest in the University's academic programs. Applicants who have indicated an interest in more than one academic program may be more likely to These factors provide valuable insights into an applicant's level of engagement and interest in the University, and can be used to develop targeted strategies to increase the yield of accepted applicants.

These findings are not out of my expectations. However, before I run the models, there are some factors that I personally found them important at first and thought they would play a role in the student's decision to attend Trinity, but eventually they turned out to be not as much important.

- Ethnicity, Race, and Religion: One thing I notice while studying at Trinity is that the school seems to be mainly White with a sizeable Hispanic population, probably due to its location in Texas. Therefore, I intially assumed that these demographic variables could affect a student's decision to attend a university, and thought that these variables would be important predictors. However, it turned out that these variables do not play a significant role in the dataset.
- School 1 Top Percent in Class and School 1 GPA Recalculated: These variables are commonly used in college admissions and are often seen as important predictors. However, in this dataset, these variables did not have a significant impact on the decision of students to attend Trinity University.
- ACT Composite: Like GPA and top percent in class, standardized test scores are often considered important predictors in college admissions. However, in this dataset, the ACT composite score were not significant predictors of a student's decision to attend Trinity University.

Generally, the Office of Admissions can use this information to develop strategies to increase the likelihood of enrollment. For example, the Office of Admissions could prioritize outreach to applicants who have attended open house events or visited the campus, or who have expressed interest in more than one academic program. This information can also be used to tailor marketing materials or to develop new initiatives that will appeal to potential applicants. In addition, the findings of the Boosting model may have broader

implications for the University's admissions policies and practices. For example, the University may consider offering more campus visit opportunities, or developing new outreach programs that target applicants who are more likely to enroll. By leveraging these insights, the University can increase the yield of accepted applicants and improve its overall enrollment rates.

VI. Reflections

Constructing a classification model to determine whether an accepted applicant will decide to attend Trinity University or not was both an interesting and challenging experience for me. I actually learned a lot of new things when working on the project, as well as gained a better sense of dealing with data when cleaning in general. I also learned more about critical steps when building an effective predictive model. It is very essential to have a thorough understanding of the data and the context in which it was collected to develop meaningful and useful insights.

During the project, I encountered several obstacles, including dealing with categorical variables with many levels and handling missing values. Dealing with categorical variables with many levels was a challenge because some variables had a large number of categories, which made it difficult to interpret the data. For example, the Merit.Award variable had various different categories, and each category had a different degree of number of observations. Grouping these categories was challenging, as it was not immediately clear how they should be grouped together, and different groupings could lead to different interpretations of the data. Dealing with missing values was another challenge that I encountered, because some variables had a large number of missing values, which could potentially bias the results or reduce the accuracy of the predictive model (for example, Athelete variable had 13120 NAs, and also had many categories with only a few cases). Another obstacle I faced was related to the Boosting model. Initially, when I tried to run the model, I encountered a strange error, and I was not sure how to resolve this issue and spent considerable time trying to figure it out. To address all of these issues, I consulted relevant sources (StackOverflow for example) to understand how other people have handled similar problems, and I also sought guidance from my professor, who provided valuable insights and suggested different techniques to preprocess and transform the data. Eventually, I successfully dealt with categorical variables with various levels, successfully handled missing values, and also discovered that I needed to increase the memory size in R to run the Boosting model successfully. This experience taught me the importance of seeking help and resources when facing technical challenges that I am not familiar with. By being resourceful and open to learning from others, I was able to overcome this obstacle and complete the project successfully.

This experience has taught me valuable lessons in data analysis and machine learning. I have gained a better understanding of the challenges and complexities involved in developing predictive models. I hope that

the insights generated from this project will be useful for the Office of Admissions at Trinity University, and I am grateful for the opportunity to contribute to this important work.