University of Tennessee Office of Alumni Development
Analytical Study of Donor Habits

Executive Summary

With so many activities that occur on the University of Tennessee's campus being funded through the generous donations of our alumni and friends, the importance of being knowledgeable of these resources is becoming more and more critical. The goal of this analytical study is to find out what qualities our "ideal donors" share and to leverage this information to increase the Office of Alumni Affair's effectiveness. More specifically, we have studied what qualities lead to alumni making donations in the next three years and what qualities lead to increased sizes of these donations.

Over 20 predictors, including information about gender, location, contact information availability, degree/school, student involvement, wealth, and prior giving history, were used predict the amount that an individual has donated to UT. Using this information, we are able to predict the amount an alumnus will donate with 87.1% accuracy.

To help our donor relations teams, we have determined that the highest value targets (those who will donate the most) are those who are married to a UT graduate, have two points of contact on file, and are in the second, third, or forth wealth estimate bracket. On the other hand, our donors in the first wealth estimate bracket are not donating as much as individuals in each of the other brackets.

Next, an original data set of over 60 predictors, including the information used in the previous model as well as more information about spouse's UT degrees/involvement, current job location/title, and historical giving amounts and length of giving history, was leveraged to create a model with a goal of determining whether an individual will make a donation to UT in the next three years. While I was limited by the processing power available, we can still correctly determine that an alumnus will make a donation to the university in the next three years with 88.2% accuracy. For our team, this information can be used to assure that we are devoting our time to those individuals who are most likely to be at the top of our donor lists, giving them much more individualized attention.

As with the last model, there were qualities that individuals who are most likely to give a gift shared, including individuals having a UT graduate as a spouse, having an email address on file, and having made a small gift to the university in the past. By targeting individuals who share these traits, we can continue to increase the effectiveness of our donor relations officers.

With this information, I recommend we give our team access to the qualities/traits that our most significant donors share or create a donor "score" which contains the likelihood of receiving a gift. While donations of all sizes are important and the relationships with our donors/alumni is of prime importance, we want to be sure we are properly utilizing our resources and the information we have available to us.

There is data that was used during this project that we could develop in order to increase the accuracy of our work. For example, information about current job titles for alumni was very inconsistent. By updating this data, we can more effectively use it in our predictions, as a CEO or President is likely to have the financial resources to give more than a bus driver. Similarly, for these models to remain accurate, the information used to make these predictions must continue to remain accurate. There was a lack of information about student organization involvement which is a great predictor towards the likelihood of a student staying involved with the university. Finally, one of the limitations of these models is that the data used was from 2015 and prior. When attempting to predict anything from more current years, there is going to be a significant drop in accuracy.

Technical Analysis

When trying to determine the amount of money that an individual will donate to the university, there are some factors that are more significant than others, including:

- Gender, with males donating significantly more than females or unknown genders
- The number of points of contact the university has with an individual, with those who the university can contact in a wider variety of ways donating more
- Having a spouse who also graduated from the university

While trying to determine whether or not an individual will donate in the next three years, however, the most important factors are:

- Having an email address on file
- Being in one of the lower brackets for lifetime giving to the university, under $25,000
- Having a spouse who also graduated from the university

There were many other insights that were brought forward during this study relating to the habits of UT's donors and by carefully studying these relationships we can continue to understand what our "ideal" donors look like and what individuals are not worth investing as much time into.

Gender & Most Recent Graduation Year

Furthermore, as you can see in Figure 1 and Figure 2 below, there is a significant relationship between an individual's most recent graduation year and their average donation amount. Figure 1 shows the distribution of average donations amounts by most recent graduation year for females, while Figure 2 shows the same for males. While the average amounts between the two genders differ significantly, you can also see that individuals who graduated between the 1960's and the 1980's were giving more, on average, then their counterparts who graduated in prior or post years. This bump in donation amounts can be credited to the fact that individuals who graduated thirty to fifty years ago are now at a point in their life's where they have much more to give then younger or older counterparts.

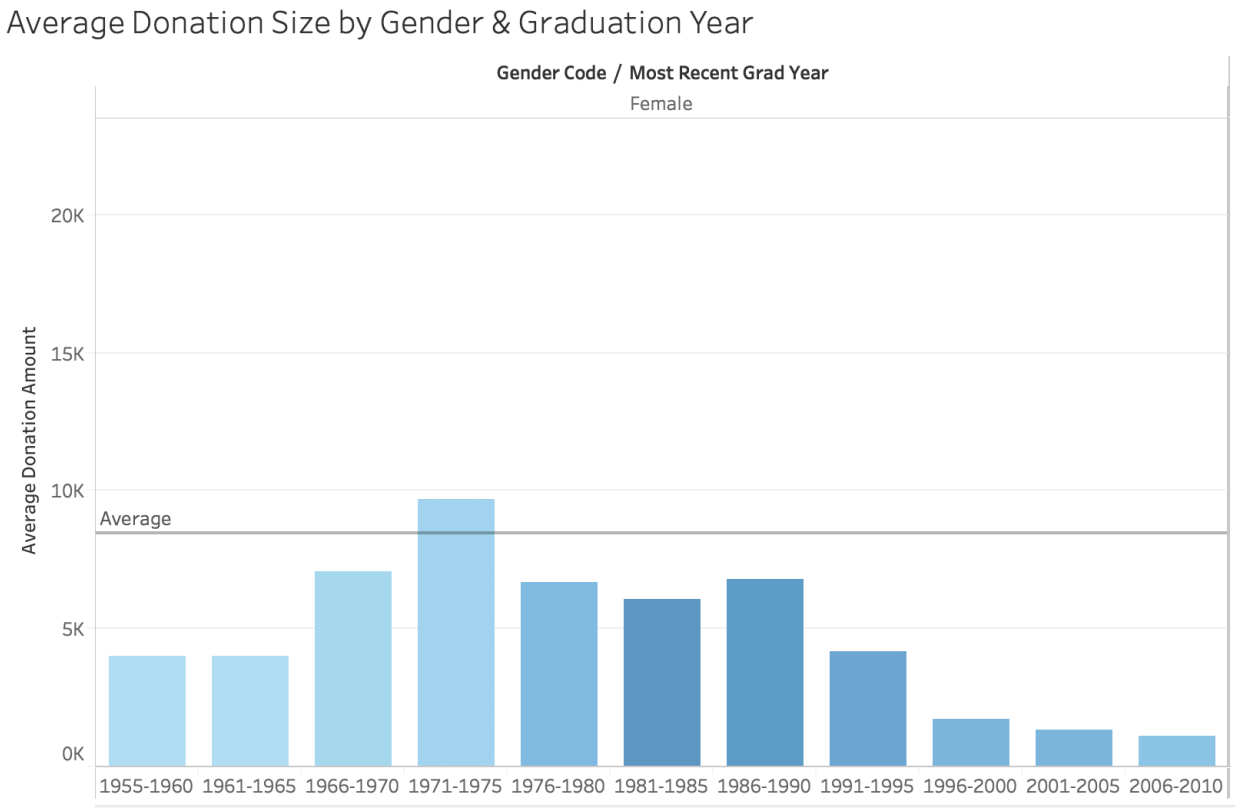

Figure 1

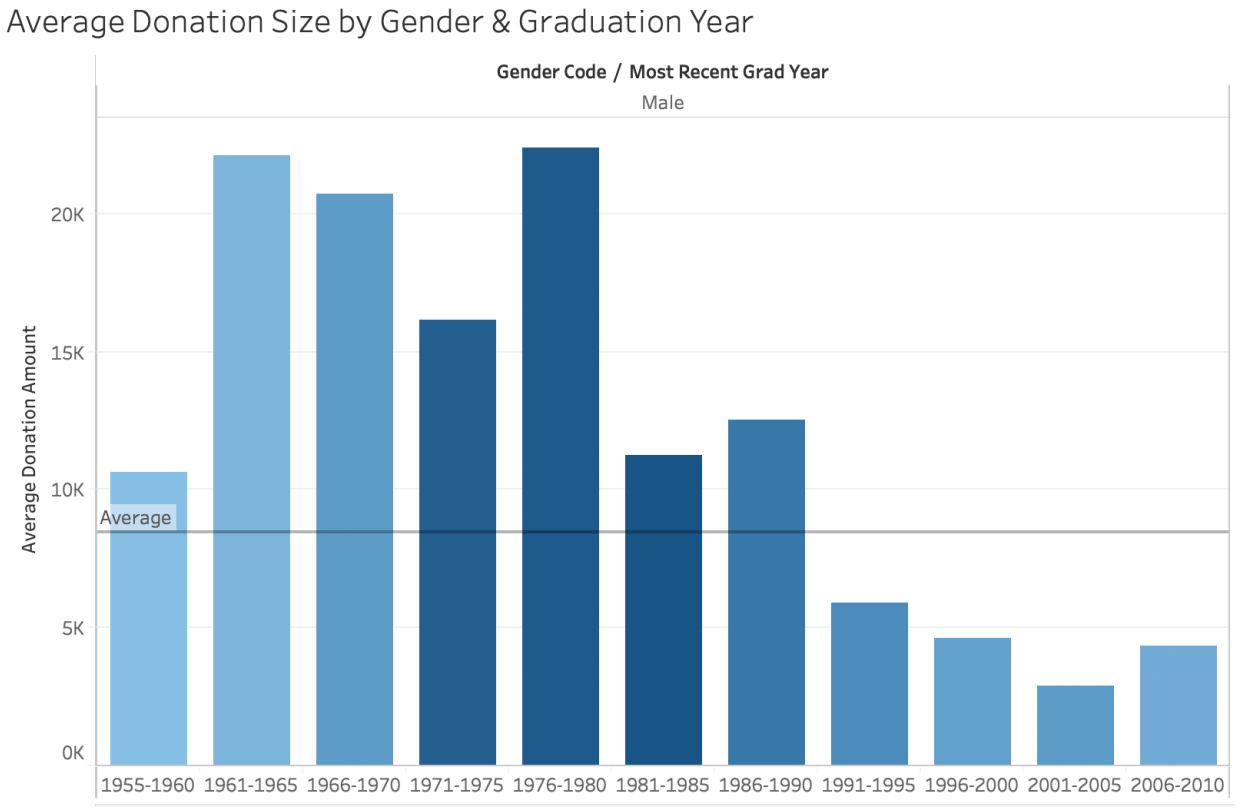
Average Donation Size by Gender & Graduation Year
Gender Code / Most Recent Grad Year
Male
Average Donation Amount
20K
15K
10K
5K
0K
Average
1955-1960
1961-1965
1966-1970
1971-1975
1976-1980
1981-1985
1986-1990
1991-1995
1996-2000
2001-2005
2006-2010

Figure 2

Available Points of Contact

When analyzing donation habits, there is an expected relationship between donation amount and the amount of time the university invests in soliciting these donations. As seen in Figure 3 below, the more points of contact that are available for an alumnus, the more they are likely to donate. Luckily for us, 63% of our alumni within the amount data set have at least two points of contact. With both an email address and a phone number it is easily to keep an alumnus up to date on the current life of the campus as well as to show the donor how much we appreciate their commitment to the university. However, this information is also skewed based on the fact that we are going to have more up to date and complete information on someone who is constantly investing in the university then someone who hasn't made contact since they graduated.

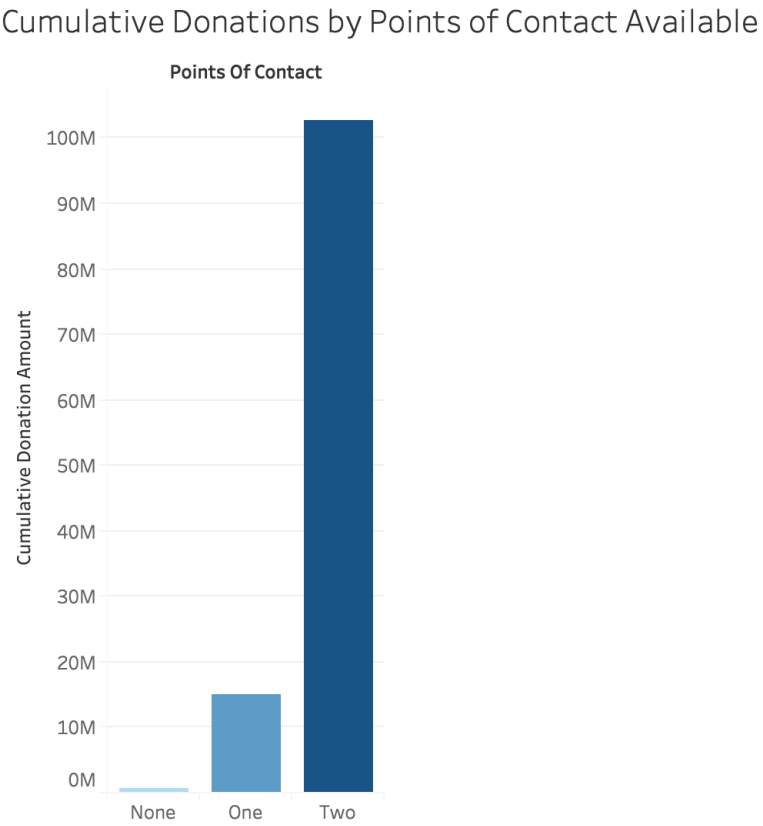

Figure 3

Location Information

One of the most interesting observations throughout our study was the analysis of donation habits based on the known current locations of our alumni base. When reviewing Figure 4 below, the states in the darker shades of blue have the higher average donation amounts, while the number displayed within the state is the number of donors from that state. As expected, over 60% of our donors are located in the state of Tennessee. However, there is an interesting observation that Colorado is the state with the highest donation average, an average of over $500 per resident. After further analysis, this anomaly from the fact that out of 88 donors one of them has donated over 2 million dollars. With this location information, our donor officers throughout the country can have much more insight into how their area compares to others.

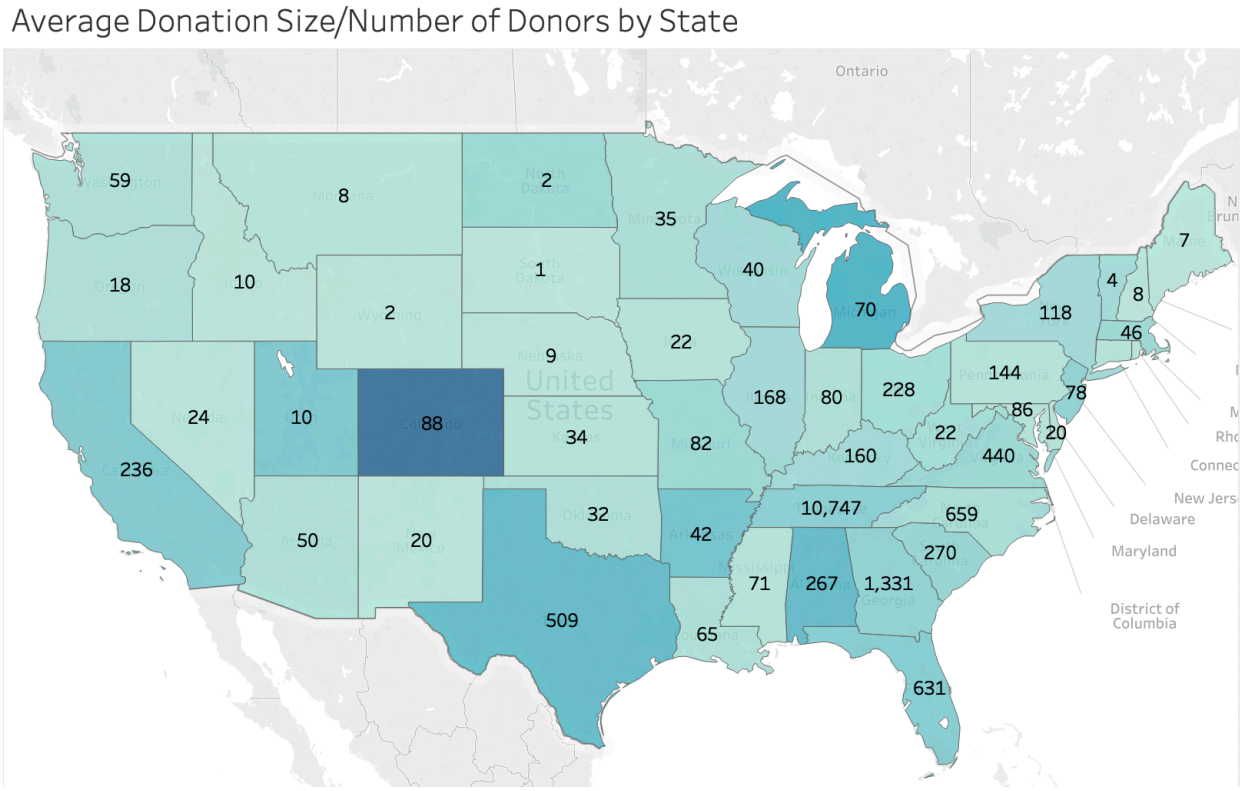

Figure 4

School Comparison (UTK Only)

For our flagship campus, the University of Tennessee in Knoxville, we conducted a study comparing the average donation amount of individuals based off of the college at which they earned their first degree in the UT system. In Figure 5, each bar represents the average donation amount, while the shades of the bars get darker for each individual that has donated from a college. Quickly, one notices that the College of Architecture and the College of Nursing have the highest donation averages, due to the positive relationship between these schools and their graduates higher pay rates. One can see that only a small portion of individuals from these schools are making donations to the university. My recommendation is that these colleges increase their outreach and begin tapping into more of these high donor bracket graduates. The Haslam College of Business, on the other hand, has the highest number of donors (16,407). There is a lot of potential for our other colleges to learn from Haslam as we look into what makes them so successful at receiving donations.

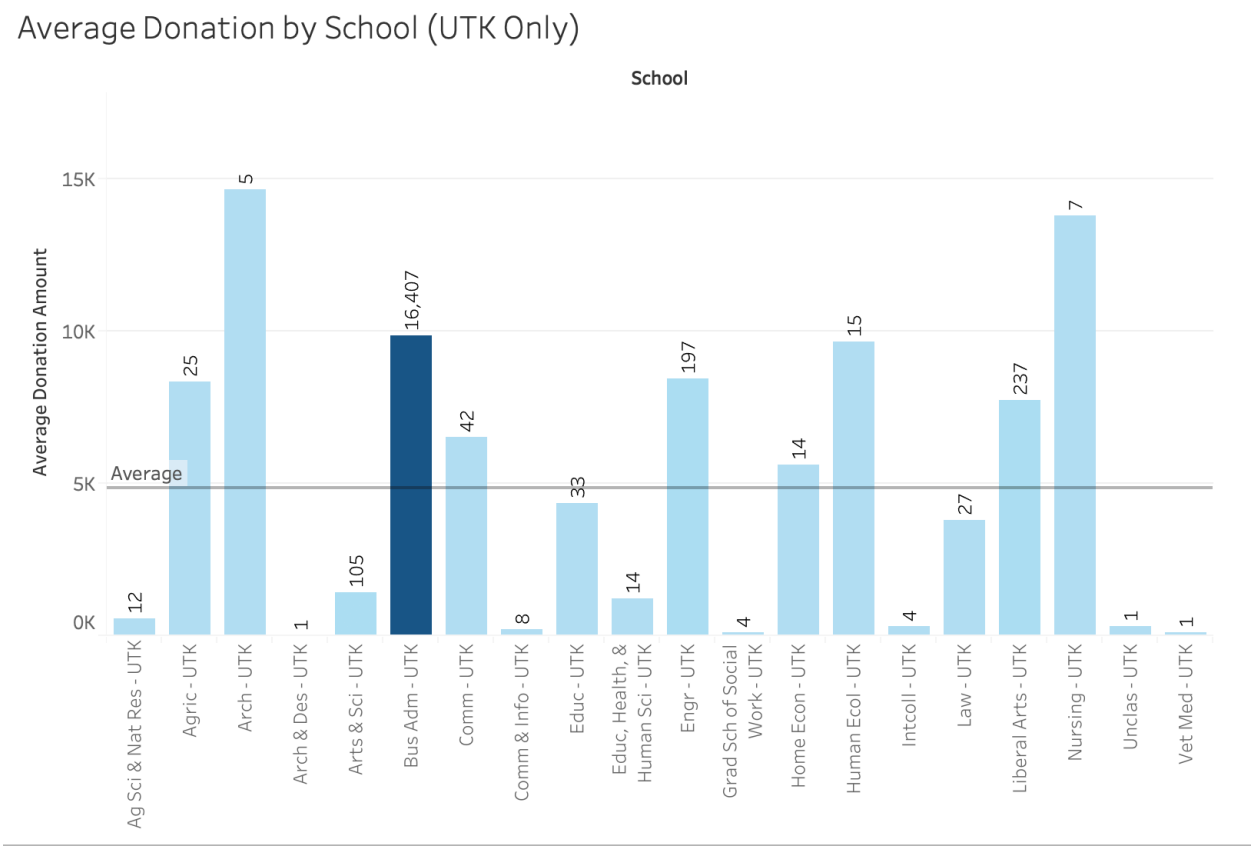

Figure 5

Wealth Estimate

One of the other areas of interest that I looked into was the effectiveness in the wealth estimate prediction showing us information that leads to donations. In Figure 6, you see the number of donors (non-donors excluded) for each of the wealth estimate categories. The darker the shade of a bar, the higher the average donation of individuals within this wealth category. As expected, 52% of our donors fall between the two highest wealth estimate ranges ($50,000 to $250,000). At the same time, the average size of these donations continues to increase as we climb higher into the wealth estimate range. An individual's ability to give more sizeable gifts come as their personal income grows.

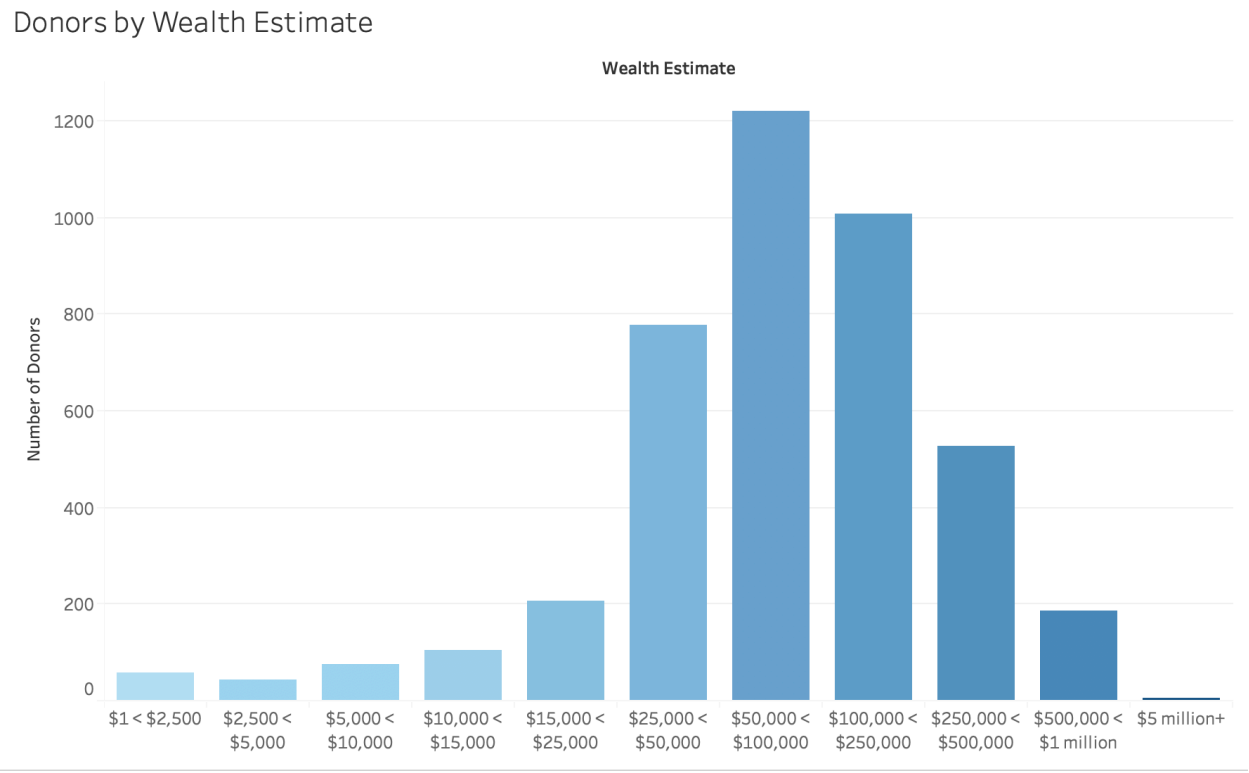

Figure 6

Data Cleaning & Manipulation Process

With a goal of determining what factors most significantly affect the donation habits of the University of Tennessee's alumni base, I first tasked myself with developing an in-depth understanding of the information that was available. Specifically, there are two data frames that were needed during analysis which had to be prepared, cleaned, and studied. The "Amount" dataset contains all of the information needed to predict how much money the alumnus has donated to the university and the "Donate" dataset contains the variables that were used to classify whether or not the alumnus will donate to UT in the next three years. While there were some identical variables between the two datasets, they were prepared independently of each other and varying predictor variables were used between the two models that were products.

Before beginning to dive into specific variables, I made the decision to convert all of the predictors into multi-level factors as I believe this is the most useful format for predictive model building. There were many predictor variables that were already provided in a fashion that made this conversion easy.

Directly Converted to Factors

Predictor variables such as gender, state, campus, degree, school, non-UT degrees, spouse UT grad, relations with UT, student life involvement, year of first gift, and wealth estimate had few enough levels where the conversion could be made directly to a multi-level factor. Many of these predictors such as the non-UT degrees, UT relations, and student life involvement were just yes or no levels. While others such as state code and the school at which a degree was earned on had many more levels. Nevertheless, all of these were able to be reduced down to 50 or less levels without much manipulations. For rows where a variable like state, gender, or student life involvement where there were missing values, "unknown" was added. While for variables like the 3rd and 4th degrees that were unknown, a "none" was added, since it was likely these individuals only earned their first degree at UT. I treated all information provided about spouses identically to the main degree data.

Combined Levels

Next, variables such as current city, job title, the city of employment, and major had too many levels to be converted to a factor that could be used for predictive modeling. There are relationships between some levels that need to be combined. To solve this problem, I used to

propose new levels function to combine the less common levels into groups of similar predictors. For current city and city of employment, I had a target of around 20 levels. While for the predictors such as major, I had a much higher target of around 50. There is potentially more information in the variation of majors, which led to the decision to aim for the larger amount of new levels.

Used to Create New Predictors

Afterwards, there were some variables that I decided to use to create new predictors that encapsulated information from multiple variables. For the email address and phone numbers, I created a factor of the number of points on contact we have with an individual, being either 0, 1, or 2. The thought behind this is that the more points of contact we have with an individual, the more likely we will be able to reach out and solicit donations. For the graduation years, I created a most recent graduation year variable by looping through an individual's varying graduation years for different degrees and returning the highest one. Then, I created bins in 5-year intervals to hold this information while reducing the number of factor levels. Lastly, I used a for loop to determine how many degrees an individual has received from UT, with the understanding that an individual who has received 5 degrees from UT is more likely to donate then an individual who earned one. These three new variables allowed me to collect information from multiple predictors and reduce the number of variables needed to use that information in a model.

Manually Cleaned or Binned

Similarly, the Greek affiliation and years that an individual has donated to the university variables had more levels then necessary. For the Greek affiliation there seemed to be a few data integrity issues causing some data to be duplicated. To solve this, I manually went through and combined the levels that seemed identically. For example, an individual who had affiliation with "UTK - Sigma Nu" and an individual who had affiliation with "UTK – Sigma Nu, UTK – Sigma Nu were likely in the same organization. To more easily tap into the information provided by the years an individual has donated to UT, I created 16 bins of varying donation levels, ranging from less than $100 to over $1 million dollars.

Removed from Datasets

Finally, the zip code and area code columns were removed from the data sets, as it was determined that the information provided by the variables duplicated the information that was provided by the state predictor.

Model Selection Process

Once the steps to manipulate the data into a useable form were complete, I began to audition models to find one appropriate for analysis.

With a goal of predicting the amount that an individual has donated to the University of Tennessee, I tested a random forest, boosted tree, and vanilla linear regression model. The boosted tree model preformed best, having a RMSE of 0.871. I believed that by combining phone and email address availability into one predictor I could increase my model's efficiency. However, when studying variable importance, email address availability still came out as a significant variable, above both phone number availability and my new combined predictor. As I continued to add trees, the model seemed to improve. However, I selected 1000 trees for the model.

Next, the goal of predicting whether or not an individual would donate in the next three years presented challenges. First, there were significantly more predictor variables available which had the potential to slow down processing and hide what was important for the predictions. I auditioned three models, a random forest, regularized logistic regression, and k-nearest neighbor model. While the random forest preformed best, with a ROC of 0.882 I was continually finding myself limited by time and processing power availability. I began with an mtry of 40, 50, and 60, hoping to maximize the use of all of my predictor variables, but after almost a day had no results. In the end, I was limited to using a max of 25 randomly selected predictors to create a model that would perform in a reasonable time.