Data 610 – Fall 2017

Assignment 3 – Fundraising Model Development

Kenneth Lulie

Dr. Moretto

University of Maryland University College

October 29, 2017

**Initial Data Review & Data Refinement and Correction**

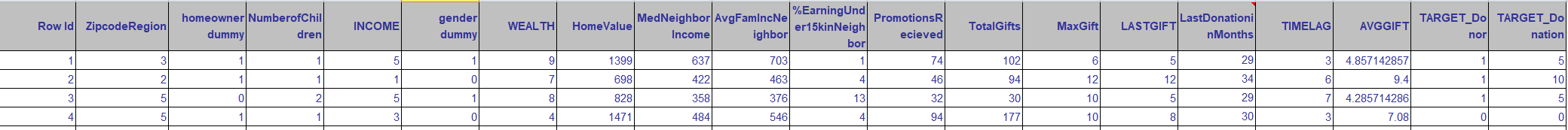
The data set selected for this exploratory data analysis was a database of fundraising donors provided as a class resource. This data set has 3,120 observations with a total of 24 variables per observation. Based on the “Data Codes” tab included with the data set, this data set was presumably created in 1998 and was used to decide which donors would be selected for outreach, along with how much to request from each donor.

The data set includes variables showing each donor’s zip code grouping, which was converted into one column with five categories to make understanding the data easier. The data set also contains information on each donor’s neighborhood, including average home value, average family income, median family income, and percent of people earning less than 15,000. The observations noting dollar values are shown in units of hundreds, but no information is given on the time period used to measure income.

Additionally, the data set contains information on the donors themselves, including gender, number of children, if they are homeowners, the number of promotions they have received, as well as the average, max, total, and last amount gifted in dollars. There is also an income observation which measures household income, but shows no indication of time period or units. Lastly, there is a wealth index field relative to each state which is calculated based on the median family income and population statistics, with zero being the lowest and nine being the highest. Variables were also renamed as needed for clarity.

When reviewing the data for mistakes, bad records, data entry errors and outliers, it was noted that the average home value for 23 observations was zero, and the last gift amount for six observations was zero. These observations were removed from the data set. It was also noted that only four observations were in a zip code group and thus were removed from the dataset due to the imbalance in the number compared to the other groups (Osborne & Overbay, 2004). Additionally, outliers in the TotalGifts, MaxGift, and LastDonationInMonths fields were identified and removed, as well as an extraneous column titled ‘RowId.’. A total of 41 observations were removed from the data set.

Before refinement, there were 3,120 observations with 24 variables and a data quality score of 67%. After refinement, there are 3,079 observations with 20 columns and a data quality score of 68%. A sample screenshot showing the data after refinement is below.



It appears that this data set was created to assist in fundraising efforts of an unspecified organization. The information related to the donors was likely used to create the TargetDonor column and the target donation amount column. For the purpose of this assignment, it will be assumed that the unknown fundraising organization assigned target values to this data set by hand during its routine fundraising efforts, and found the results to be satisfactory. The organization presumably attempted to reverse engineer this dataset in order to determine the logic used to select donors and the amounts requested from these donors as part of an effort to maximize revenue from fundraising.

**Initial Predictive Model for Target Donors**

To help understand the categorical target variable of TargetDonors, Watson was asked to provide a breakdown of TargetDonors and provided a bar graph showing an even breakdown of 1,538 targeted versus 1,541 not targeted (see Figure 1). Next, Watson was asked to make a predictive model for target donors using all available variables in the data set, except for target donations (see Figure 2). Watson created a CHAID classification tree to produce a decision tree predictive model for this categorical target variable (IBM, 2017).

The confusion matrix indicates that the decision tree produced by Watson was not powerful, with only 58% of the observations being correctly predicted (see Figure 3). As the target variable is almost evenly split into two groups, this model is only 8% better than a baseline of random guessing (Ray, 2016). The matrix also shows that the model had specificity of 60% and sensitivity of 56%, indicating that this decision tree prediction model would be slightly better at correctly guessing which donors were not selected versus donors who were selected (Markham, 2014).

**Decision Rules Analysis for Target Donors**

Watson’s predictive model used the LastGift, HomeValue, LastDonationinMonths, PromotionsReceived and TotalGifts variables in the decisions rules, of which LastGift was the most predictive. The decision rules varied in accuracy, ranging from 50% to 67% (see Figures 4 & 5). The rule with the highest accuracy of 67% predicted that if a donor had a LastGift of over 14, a LastDonation of less than or equal to 30, a LastGift of 14 to 20, and a TotalGifts of less than or equal to 66, the donor should be selected as a target donor. The rule with the second highest accuracy of 66% predicted that, if a donors LastGift was less than or equal to 6, their LastDonationinMonths was less than or equal to 31, and their PromotionsReceived was greater than 26, the donor should be selected as a target donor.

The rule with the third highest accuracy of 61% predicted that, if a donors LastGift was less than or equal to 6, their LastDonationinMonths was less than or equal to 31, and their PromotionsReceived was less than or equal to 26, the donor should **not** be selected as a target donor. The rule with the fourth highest accuracy of 60% predicted that, if a donors LastGift was greater than 14, LastDonationInMonths was less than or equal to 30, and LastGift was greater than 20, the donor should **not** be selected as a target donor. The rule with the fifth highest accuracy of 60% predicted simply that if the LastGift was greater than 14 and LastDonationInMonths was greater than 30, the donor should **not** be selected as a target donor.

All five of these most predictive rules were weighted highly on LastGift and LastDonationinMonths, with some rules also using PromotionsReceived, and TotalGifts. An organization could use these rules to educate its employees that the most useful information when targeting donors is data relating to history of donations, as opposed to information regarding neighborhood demographics.

**Initial Predictive Model for Target Donations**

Next, to help understand how to predict the continuous target variable of TargetDonations, the data set was filtered to ignore non targeted donors. Watson was asked to use this data set of 1,563 observations to show a distribution of TargetDonations, and showed that the majority of TargetDonations are in the 1-21 grouping (see Figure 6). Next, Watson was asked to make a predictive model for TargetDonations using all remaining variables (see Figure 7). Watson created a CHAID classification tree to produce a decision tree predictive model for this continuous target variable (Berson, 1999). The decision tree model Watson produced had 43% accuracy.

The decision tree model used the LastGift, AvgGift, MaxGift, LastDonationinMonths, and %EarningUnder15k variables to predict TargetDonations (see Figure 8). The LastGift variable was found to have a predictor importance of .77 and AvgGift was found to have a value of .22, showing that LastGift has the most predictive power of all variables, and that LastGift and AvgGift combined have the most predictive power over other variable combinations (IBM, 2017). A fundraising organization could use this information when deciding how much to request from a donor based on how much that donor gave last time and what their average gift is.

**New Predictive Model for Target Donors**

To improve the model for predicting TargetDonors, a new variable of TotalGiftTimes was added representing the TotalGifts divided by the AverageGift in order to calculate the total number of times a donor has gifted. This variable was added under the theory that the total number of times a person has donated could influence if that person would be considered a good target. Additionally, a new variable was created called MaxGiftDividedByLastGift to calculate the size of the last gift given by the donor compared to the total gifts given by that donor. This variable was added under the theory that a donor would not be selected as a target donor if they had leveled off in their donations. Lastly, all other variables except for LastGift, HomeValue, LastDonationinMonths, PromotionsReceived, and Total gifts, which were used in the original decision tree, were removed from the data set to see if the data set could be simplified.

Using this refined data set, another decision tree was created using Watson (see Figure 9). While the result of these changes lowered the predictive strength from 58% to 57% (see Figure 10), the variables available for prediction in the data set went from 17 to 7, which is an improvement in simplicity of the data set.

The rules determined by the new decision tree were similar to the original decision tree (see Figures 11 & 12). The rule with the most accurate grouping of 71% for 270 records predicted that donors with a LastGift greater than 14, MaxGiftDividedbyLastGift greater than or equal to 4.93, HomeValue of under or equal to 1,577.5, LastDonationinMonths greater than 29 and a TotalGifts greater than 40 should **not** be selected. Despite new variables being added, the new and original models did not significantly differ from each other, as the new model did use the added MaxGiftDividedbyLastGift variable, but did not use the added TotalGiftTimes variable. Both the original and new models had 58% predictive strength. This low level could indicate that the target donors were not selected in a standardized manner, or were chosen using data that was not available in this data set. It could be speculated that the donors were hand selected by local fundraising volunteers using human intuition and personal interactions with the donors by adding a proxy of generosity compared to the income level.

**New Predictive Model for Target Donations**

To attempt to improve the model for target donations, unused variables were deleted from the data set, and new variables were derived from available fields. For this model, the variable TotalGiftTimes was added again to the data set. A new variable of IncomeDividedbyAvgGift was also added to the dataset as a way of exploring if a combination of these variables could improve the model.

Watson was asked to produce another decision tree model (see Figure 13). This decision tree had a predictive strength of 45%, a 2% increase over the original decision tree model. The new model differed slightly from the original, as it used the IncomeDividedbyAvgGift rule to help guide the predicted value for those donors who had a LastGift of over 17 and a max Gift of over 20; however, the decision rules are otherwise very similar to the previous model (see Figure 14). Those donors with an IncomeDividedByAverageGift ratio of 4.14 or higher were predicted to have a target donation value of 36.31, as opposed to a target donation value of 24.53 for those with a ratio of lower than 4.14. These findings indicate that the relationship between the AverageGift and Income level could be useful when determining how much could be requested from the biggest donors, as these categories had the two highest predicted values.

Overall, with the exception of the use of the IncomeDividedByAverageGift variable, the new model is very similar to the original model. To build a stronger model, new information would likely need to be added to the data set.

**Potential Use in a Fundraising Organization**

The decision tree models developed in this analysis could be used in a fundraising organization such as the Red Cross. Decision tree rules are easy to understand, and could be accepted as guidelines to use in donor outreach. For example, showing a spiral chart of the predictive variables (see Figure 15) alongside the previously referenced decision trees could help new volunteers quickly understand what to focus on. However, before attempting to replace human intuition with decision rules more research would need to be done on the effectiveness of each approach.

An advantage to changing from a human intuition model to a well-designed data driven model would almost certainly be higher response rates from the donors selected, and increased revenue from the average donor, ultimately resulting in more revenue per dollar spent on fundraising. The data driven model would also eliminate the need to have human volunteers go over the records by hand freeing up that human capital for other uses

**References**

Multimedia: September 16, 2016). Overview of Watson analytics [Web]. Retrieved from https://www.ibm.com/communities/analytics/watson-analytics-blog/video-overview-of-watsonanalytics/

Berson, A., Smith, S., & Thearling, K. (1999.).An overview of data mining techniques[section 2.2]. Retrieved October 28, 2017 from http://www.thearling.com/text/dmtechniques/dmtechniques.html

Markham, Kevin (March 25, 2014). Simple guide to confusion matrix terminology Retrieved October 27, 2017 from http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

Ray, S. (2016, January). A comprehensive guide to data exploration. Retrieved from http://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

Osborne, J., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research & Evaluation, 9(6). Retrieved December 2, 2015 from http://pareonline.net/getvn.asp?v=9&n=6

Appendix A

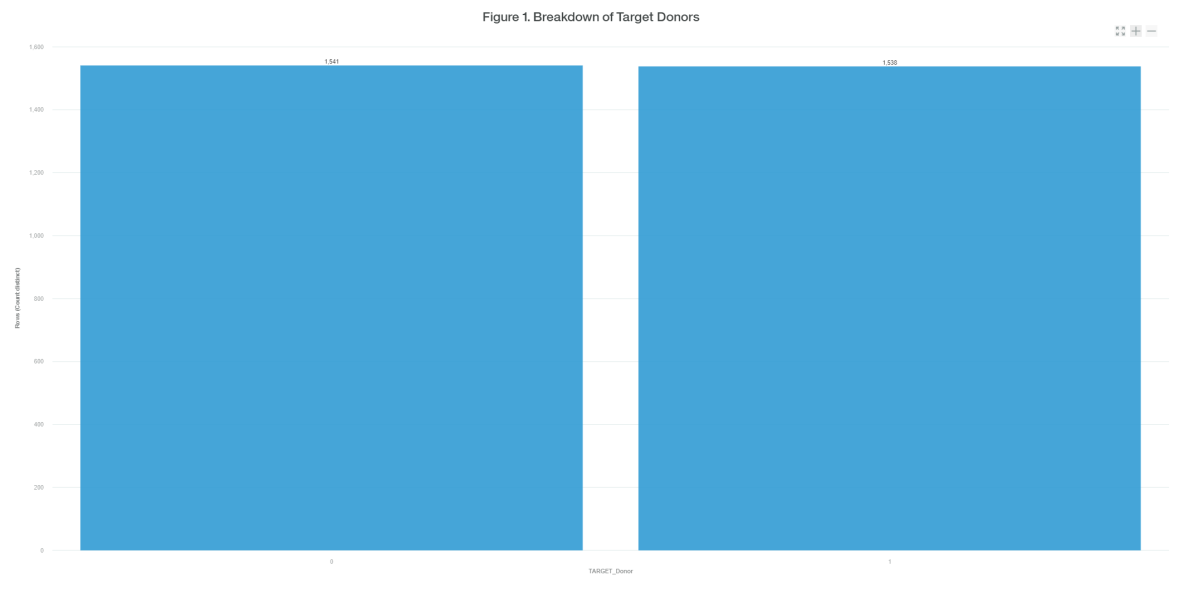
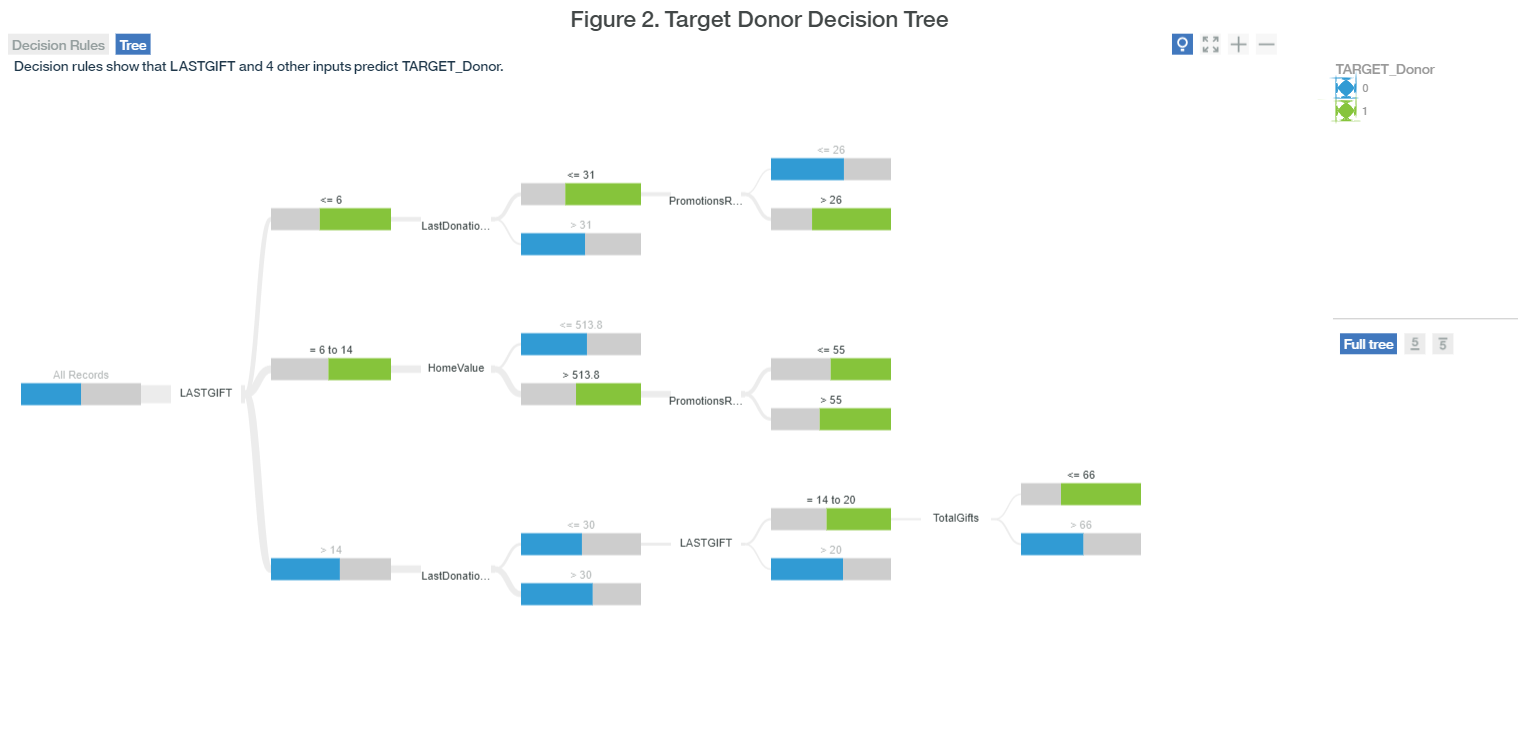
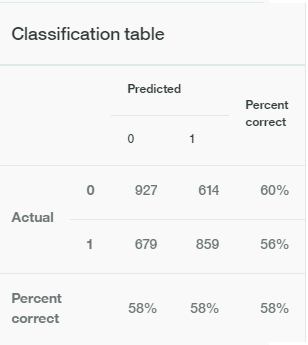


Figure 3. Target Donor Confusion Matrix









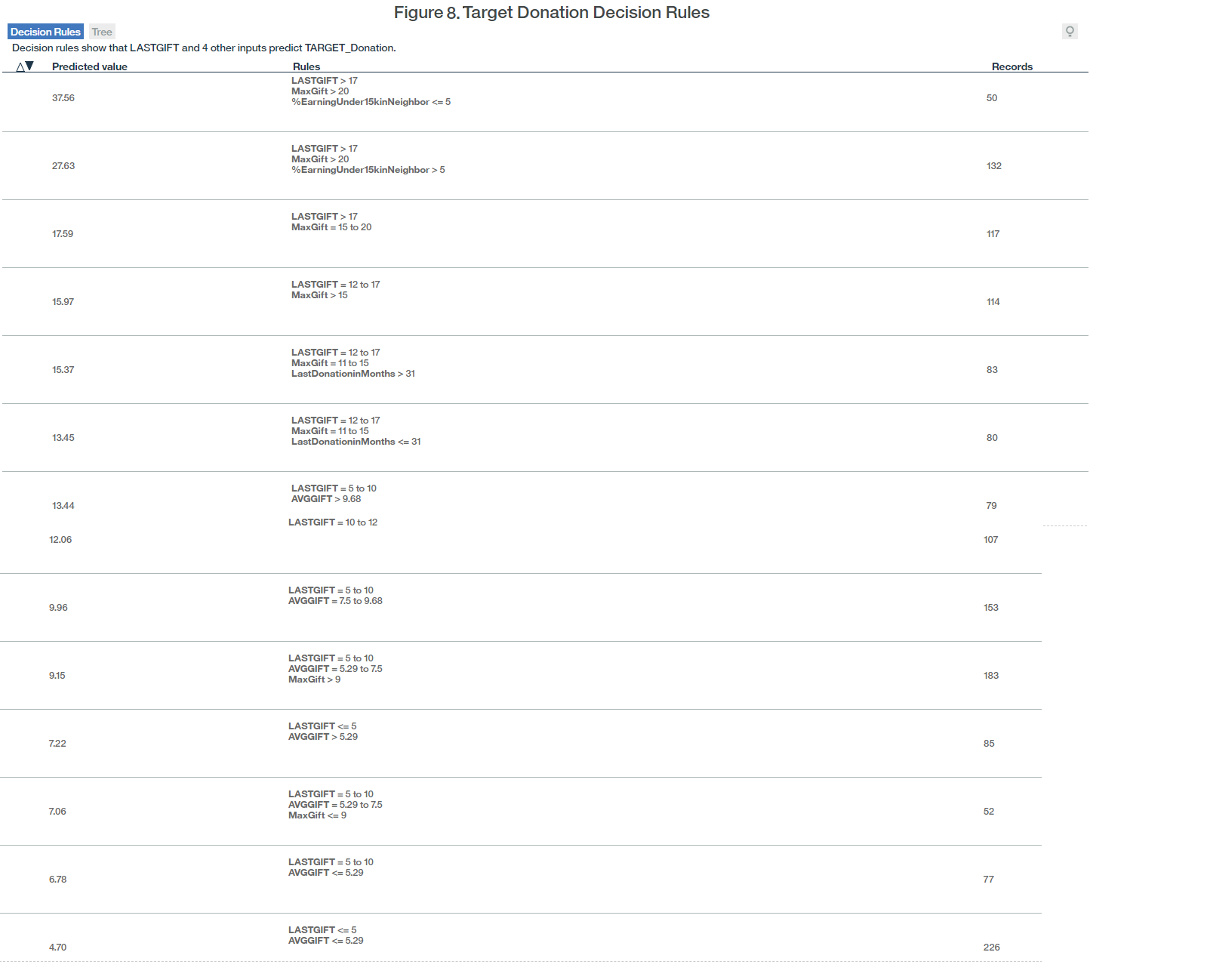
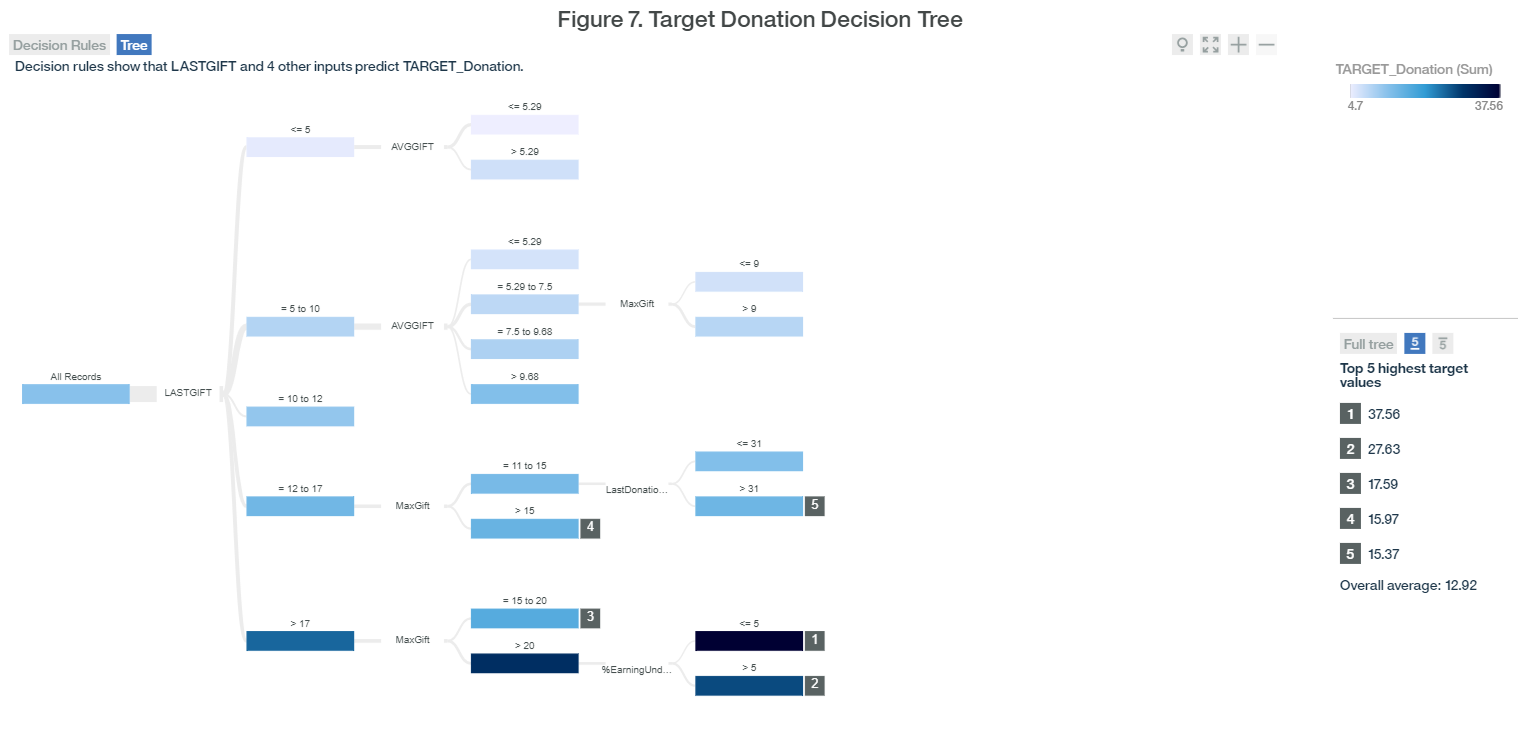
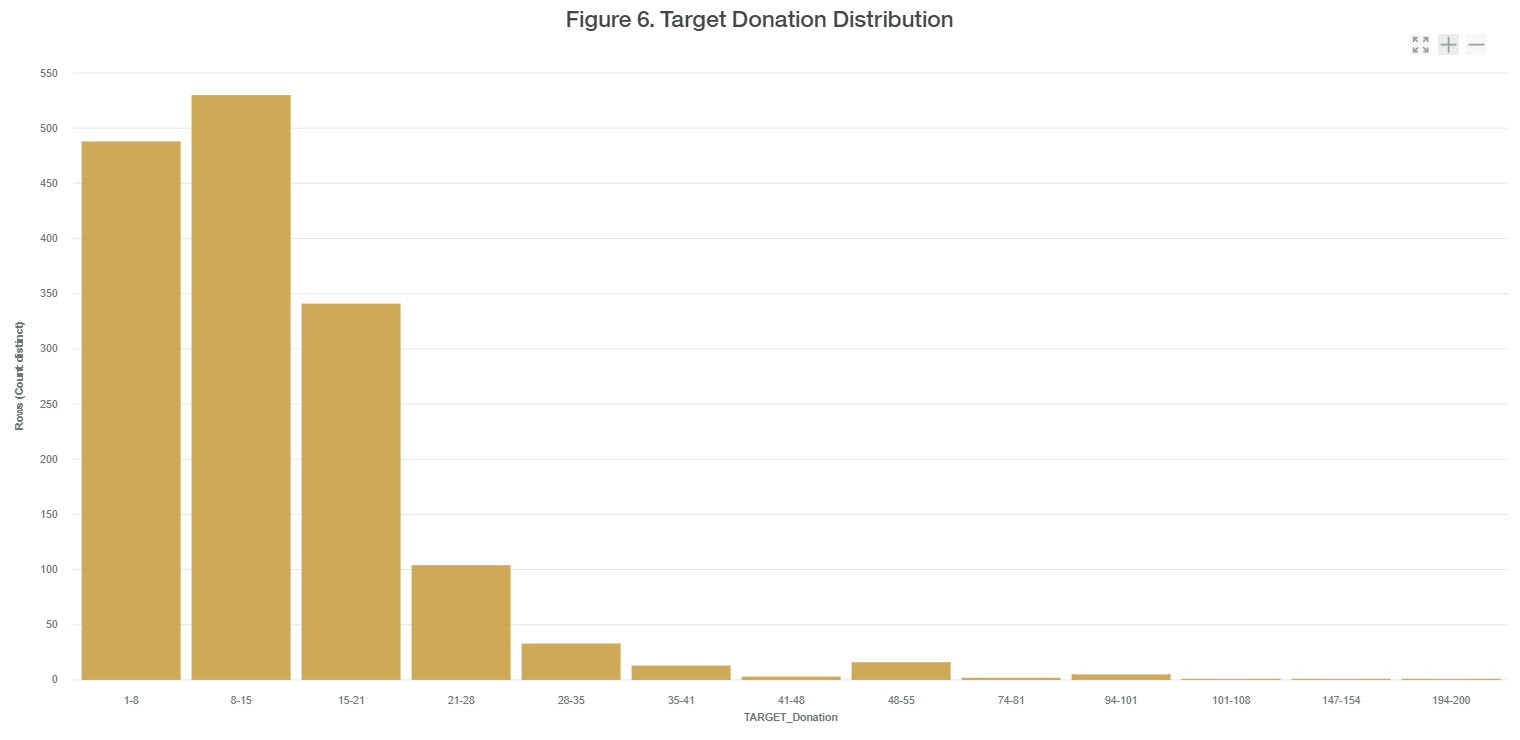




Figure 10. Improved Target Donor Confusion Matrix

