Predicting Miami University Alumni Donations

Client: Tim Jones Office of the Advancement, Miami University

Analysts: Palma Daawin, Craig Heard, Chelsea Hillenburg

Introduction

The Office of University Advancement is responsible for alumni engagement and fundraising. This department is integral to collecting donations from a variety of constituents. These constituents consist of about 212,000 alumni, 4,000 faculty and staff, 32,000 friends, and 100,000 parents. Donations can range anywhere from less than a dollar to millions of dollars. Advancement considers small gifts to be donations totalling less than ten thousand dollars. Large gifts are donations greater than ten thousand dollars. Between four and six million dollars are raised annually in donations.

Advancement employees may travel to host alumni events or build relationships with known-wealthy constituents. There are many funds to which people can donate. For instance, the parents' fund provides funding for student events and student-worker awards. The economy can influence the amount of money donated, and institutional changes, such as donor based seating (2013) and the changing of the mascot from Redskins to RedHawks (1990s), can make a difference as well. A person's involvement in athletic activities can also influence the funds they donate to. We seek to explore the relationships between donations and involvement in various activities such as athletics as well as other demographic characteristics

With a dataset having more than 1.5 million transactional records dating back further than 20 years, and a large amount of information on donors, there are many possible questions to answer and analysis approaches. The data contains a multitude of variables such as demographics, type of relationship to the University, giving history, special collections, Miami academic division (e.g. Farmer Business School), athletics involvement and information regarding involvement in other organizations. Particular groups of interest are young alumni (<10 Years), athletes, faculty/staff, consistent and new donors.

Mr. Jones, our client, suggested several questions that we might explore:

- What are the measures of giving patterns for athletes? Are they the same after 10 years?
- How long do constituents' relationship last? Why do they leave?
- How costly is losing a donor compared to keeping them?
- What is the participation in activities like amongst constituents?
- Do constituents upgrade or downgrade their donations? When and why?
- Is there a difference between class years? Between the levels of donors?
- Are constituents that are more likely to participate more likely to donate?

Because of the large and complex nature of the data, and time constraints imposed by the semester, we have not been able to address all of these questions. Instead, we performed an extensive exploratory analysis of the data and built models that provided further insights that may be of interest to Advancement. At the end of this report, we will provide some suggestions for taking this project further.

Data Description

The data set consisted of 11 different sets of .csv files. The data files used are described in Table A.2 of the appendix. The main data set relating to alumni giving was the “giving.csv” file. This consisted of a history of the various donations made by alumni to Miami University over the years. In our analysis this remains the main data of interest, but we chose to focus on entities graduating after 1986 due to a change in record keeping that made older data of less integrity. Originally, there were 186,839 unique entities with transactions from 1982 to 2016. Once the data had been cleaned and filtered for the analysis, the number of entities was reduced to 93,831. We give some details regarding the cleaning/filtering process below.

From the “giving” data file, we created a master dataset with an observation for each entity for each year from the graduation date to the years after graduation or the most current year in the dataset (2016). From here, we merged the predictors found in Table A.1 of the appendix to this master dataset. In order to do this merging, we had to condense some information down to just one number or category for each entity. For instance, we had to use only the preferred contact information to determine the entity’s region as opposed to using all data on file for the entity’s past or possibly multiple current locations. We also filtered out entities with unreasonable information. The following is a complete list of filtering we used to create the final dataset to be used in modeling:

- Removed entities belonging to more than three Greek organizations
- Removed observations with no associated entity ID
- Removed data on transactions before 1987
- Only kept data on transactions during or after the year of graduation
- Used only preferred contact information
- Created a category *unknown* to be used for entities with missing information for predictors like region

The list and descriptions of variables used in our study is found in Table A.1 of the appendix.

For possible categorical predictors with many categories, we created new variables with fewer categories. For instance there were over 50 different types of athletic activities, so we reduced this into two categories, *athlete* and *non-athlete*. We chose to reduce the number of categories for statistical modeling purposes. Reducing the number of categories increases the sample size within categories and, thus, within intersections of categories from multiple variables. This allows us to model interactions between multiple categorical predictors. When possible, rather than removing entities in a category with few observations, we combined them with another category, but if we had no *other* or *unknown* category that we could combine the category with, we had to filter out entities within the category. Details on the collapsing of categories and removal of entities follows:

- Region - Canada and military into *unknown*
- Gender - Remove those not identifying as male or female
- Degree Category - Associate into *other*

Other variables that might be useful in helping predict a response were also created. Some of these variables were GradToGive, GaveLastYear, GaveLastYear.Amt, LastGave, etc. A full description of these are found in Table A.1 of the appendix.

Exploratory Data Analysis

Before formally analyzing any data it is important to conduct an Exploratory Data Analysis (EDA) to give preliminary insights that are useful in their own right as well as informing subsequent modeling efforts. By exploring the data via visualizations and summary tables, we see what variables have relationships with donation amount and likelihood. We decided to use appropriate heat maps, bar charts and time series plots depending on the variable or variables of interest.

Region and State

Figure 1 below indicates that the US state with the highest average amount of total donations per transaction, since 1986, is California with just over $3,750 (1011 total transactions). This is true despite California not being one of the states with the oldest donors. Other states worth mentioning are Massachusetts being second highest with just under $3,310 (384 total transactions) and Connecticut with just under $3,310 (268 total transactions). There were quite a few states that had small average total donations, however these states had less than 100 total transactions. The three lowest giving states with transactions greater than 100 were donors without a state with an average of just under $350 (2183 total transactions), Washington with just under $580 (248 total transactions) and Oregon with just over $680 (125 total transactions). The highest three median donations by state, of those which had a reasonably large number of donations, were Kansas, Michigan and Missouri with a median donation of $373 (102 total transactions), $245 (846 total transactions) and $215 (370 total transactions), respectively.

When looking specifically at regions, both the Northeast and West have the highest average total donations by Alumni with $2,013 and $1,942, respectively (Table A.3). However, the highest median total donation was the South with $175. This seems to indicate that ‘big’ donations have a large influence when aggregating by region and state.

Figure 1: Average amount of donations Heat Map by State.

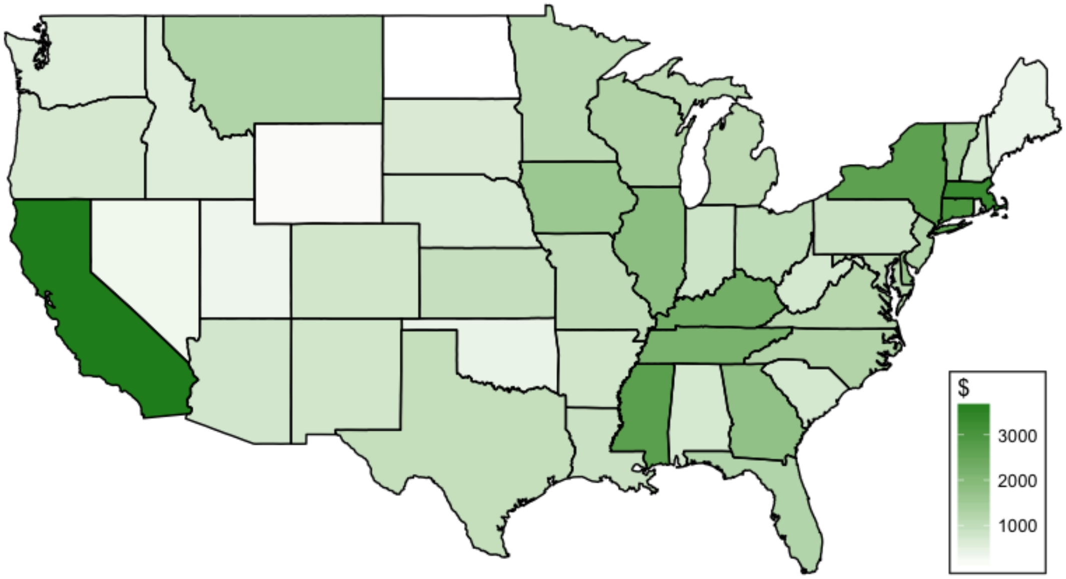
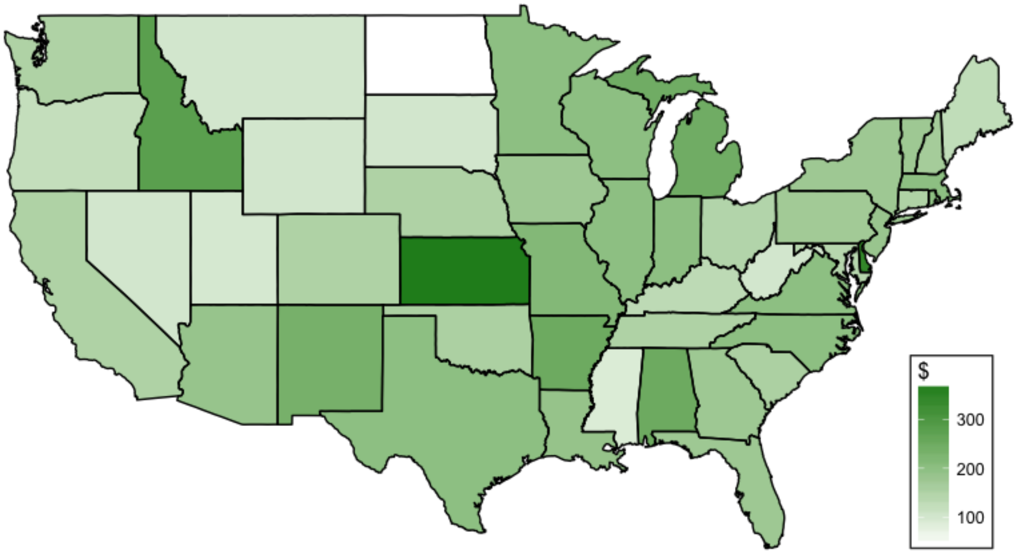


Figure 2: Median amount of donations Heat Map by State.



Athlete Status

Figure 3 highlights the difference in total amount of donations between athlete and non-athlete alumni from Miami. We can see that just over 85% of non-athletes have given less than $1000, just over 10 percent more than athletes. There seems to be a higher proportion of athletes that donate larger amounts as Alumni. We can see that if Miami receives a large donation it is more likely to come from an ex-student-athlete since just over 4% of athletes have given more than $10,000 compared to just under 1.5% of non-athletes. Figure 4 shows the median donations of Alumni since they have graduated. Again this graph supports the claim that Athletes seem to be giving to the University more than Non-Athletes on a regular basis.

Figure 3: Conditional Frequencies on Total Giving comparing Athletes and Non-Athletes

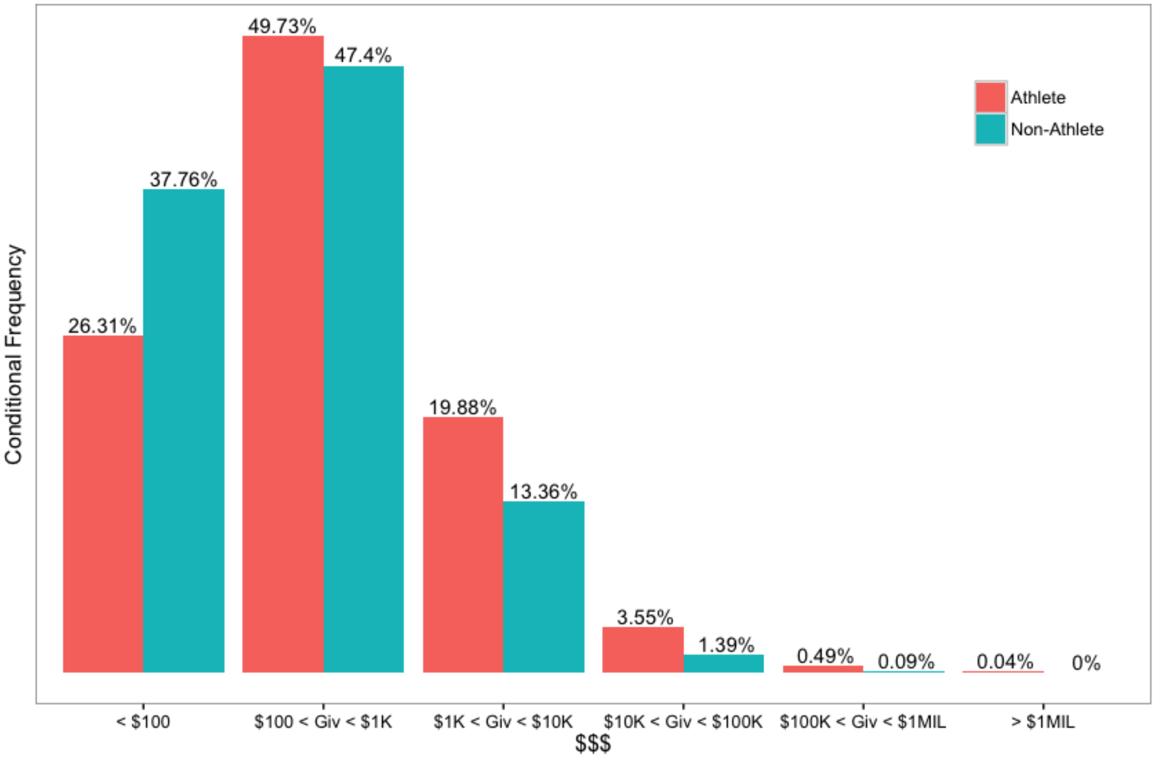


Figure 4: Median Donations since graduating from Miami by Athlete Status.

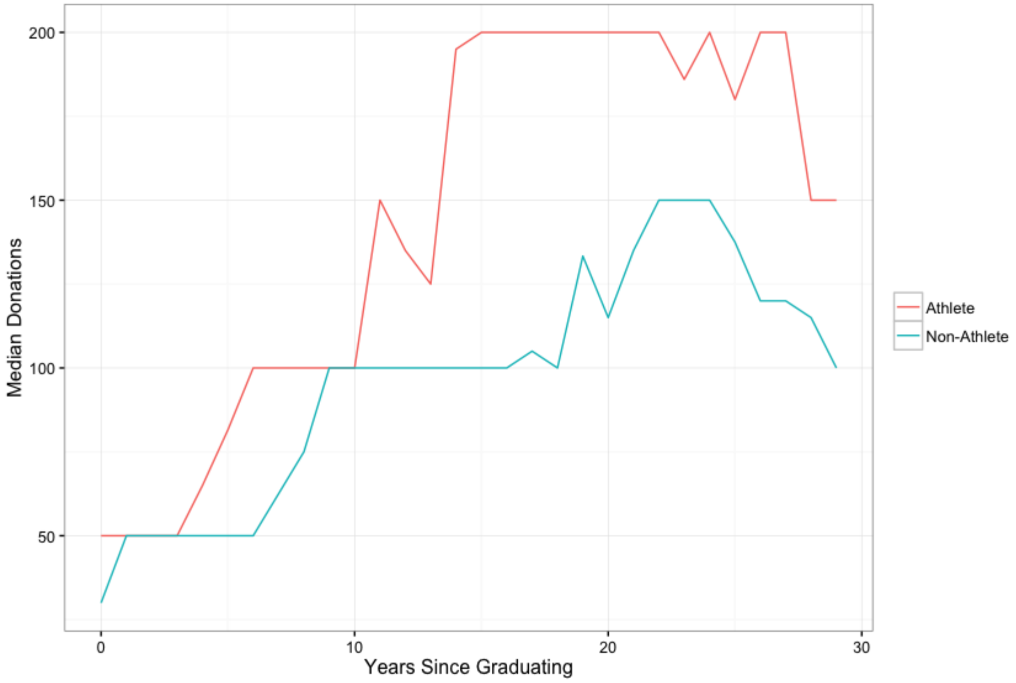
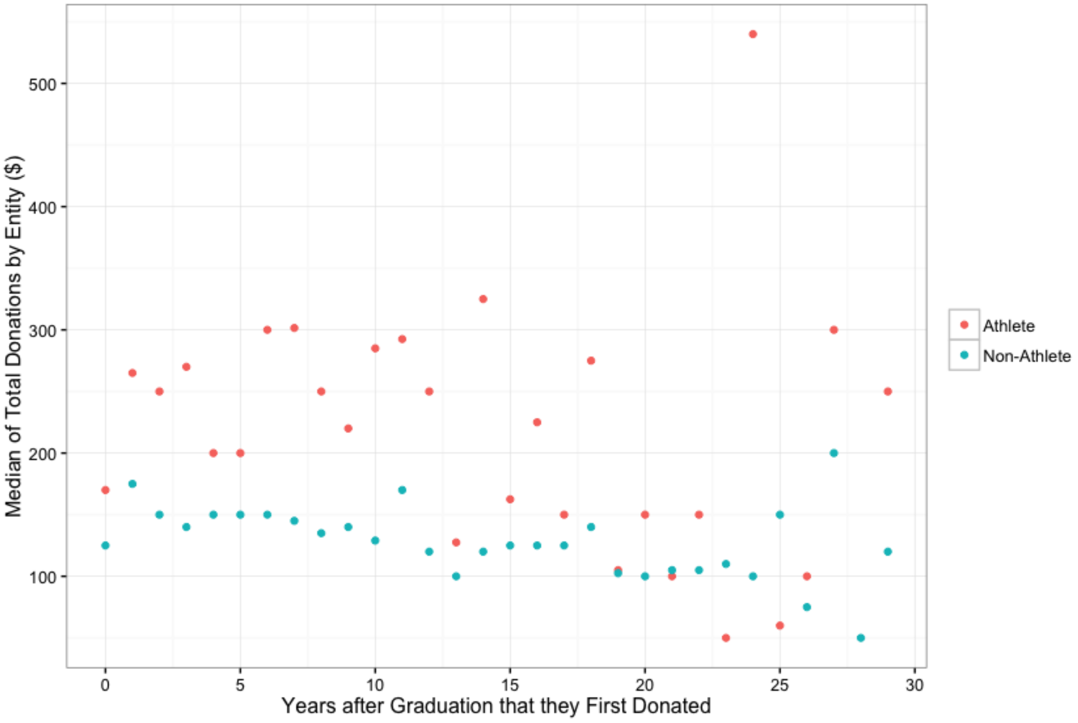


Figure 4 shows that donors give more the older they become. What is noticeable is the difference between athletes and non-athletes. Athletes seem to be giving higher median donations than non-athletes not only close to graduation, but also considerably more after a 10 year period and later giving. This is also evident when looking at Figure 5 where we can see that Athletes tend to be giving more in total. However, there were less than 10 athletes for first time donors who donated at 20 years or later, which would explain the volatility after 20 years.

Figure 5: Median of Total Donations versus first year of donating.



School, Greek Life and Marital Status

In Figure 6 we can see that Farmer Business school seems to bring in more and larger donations than the other schools once you get past 20 years as being as an alumni. Since this school is very reputable and a lot of 'big' name employers are attracted by the school, it makes sense that successful students will be willing to give more money. Figure 7 and Figure 8 look similar and give some compelling insight that Alumni who participated in Greek life are more likely to give more and additionally married alumni are more likely to give more, especially 12-15 years after graduating.

Figure 6: Median Donations since graduating from Miami by School.

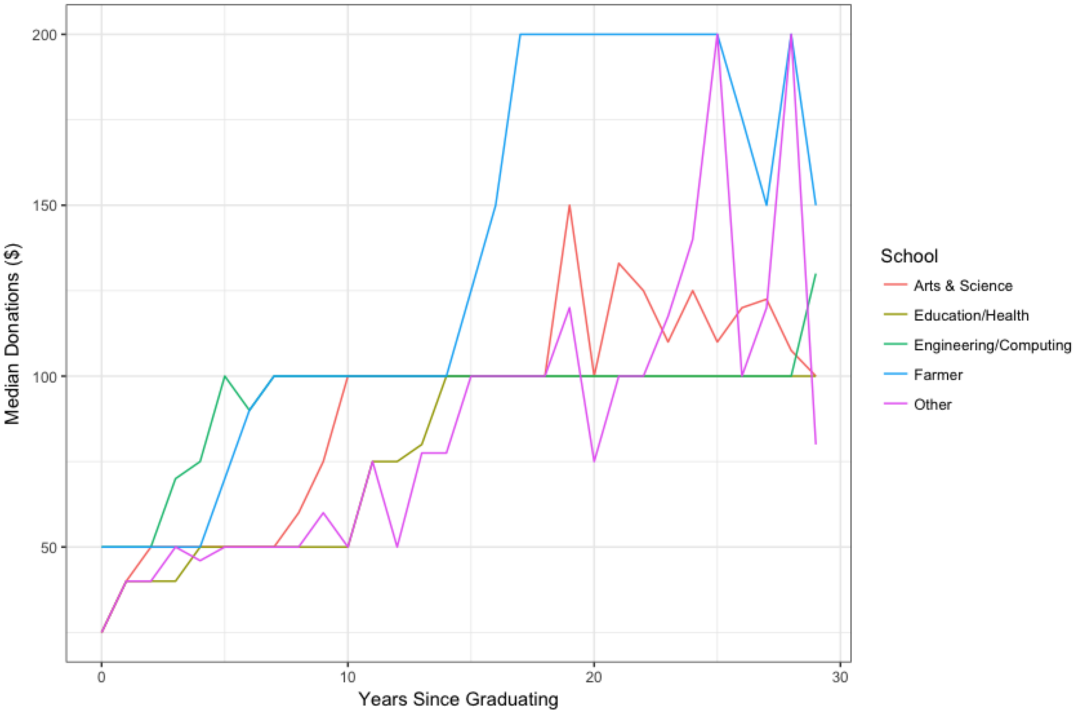


Figure 7: Median Donations since graduating from Miami by Participation in Greek life.

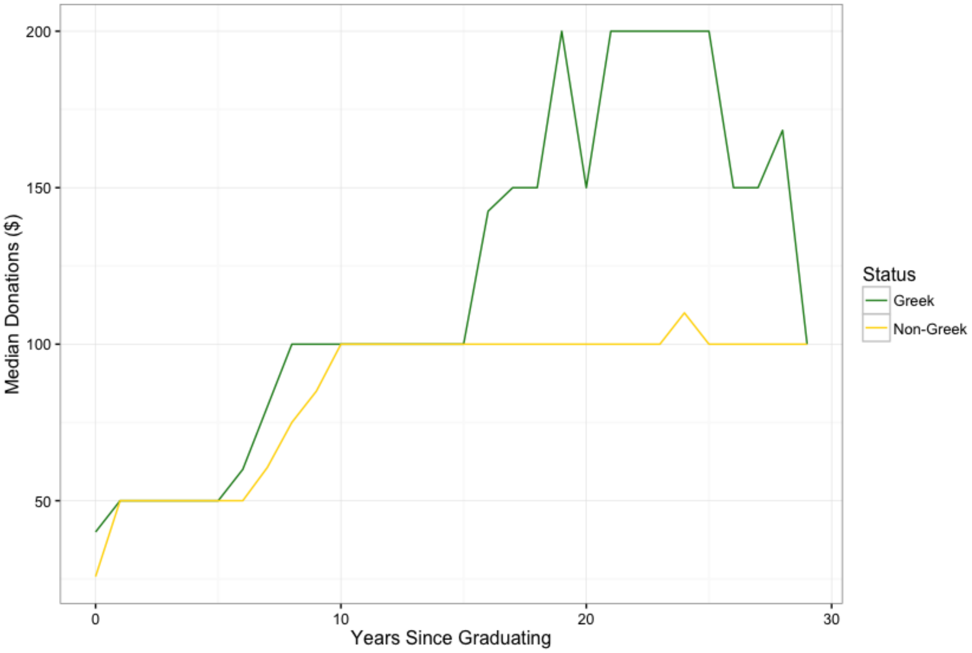
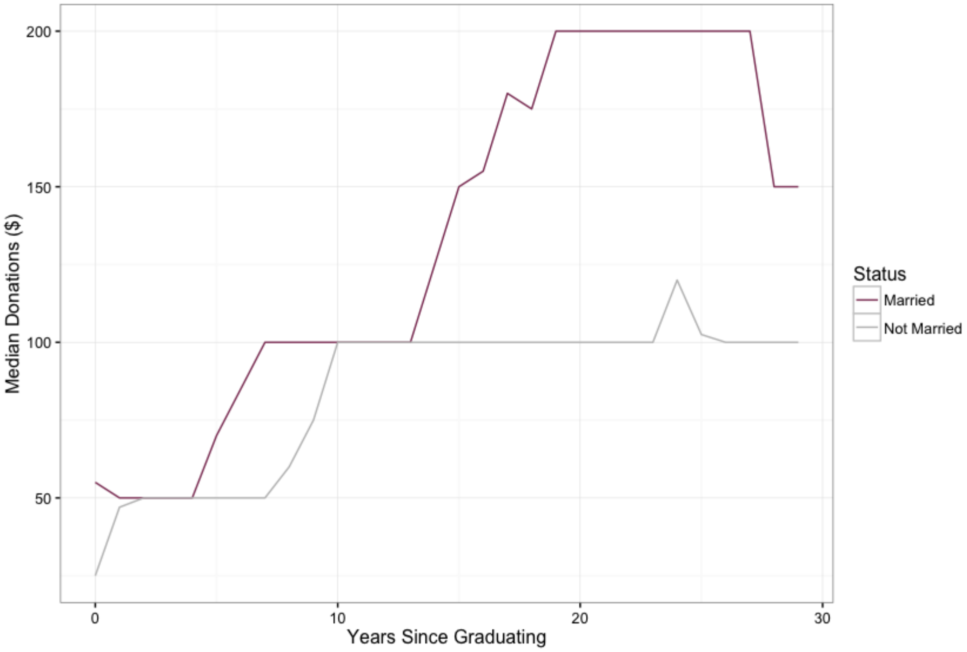


Figure 8: Median Donations since graduating from Miami by Marital Status.



Methods & Results

In order to facilitate the construction of predictive models, the data was partitioned into two sets. The first, called the training set, was used to fit models, and the second, called the validation or test or holdout set, was used to evaluate the quality of the models to predict new data. We randomly selected half of the entity IDs, and those selected were used in the training set (roughly split in half). The rest constituted the holdout sample, which was used to test the fitted models on new data. The predictions on this new data are evaluated according to the square root of the mean squared error (RMSE) of prediction if the response is numeric (i.e. “the amount of next gift”), or the area under the receiver operating characteristic (ROC) curve if the response is binary (i.e. “will the entity give next year?”). The models with the lowest RMSE or largest area under the ROC curve indicate better predictive ability. (The ROC curve illustrates the performance of a binary classifier--in this case, a logistic regression model--by assessing how accurately our fitted model predicts whether an entity gives in a particular year or not.)

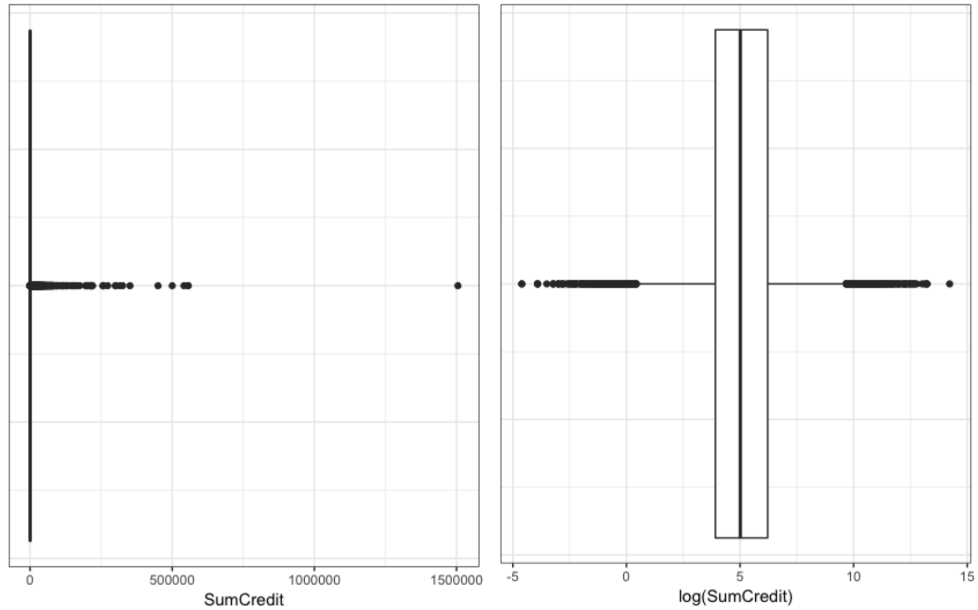
To select a final predictive model, we generated a variety of models, then tested them on the holdout sample. For the two models with a numeric response (Models 1 and 2 below), we used a statistical method called forward selection to train models of varying sizes. That is, we tried to find the one-predictor model that explained the most variation in the training set, then did the same thing for two-predictor models, three-predictor models, etc. The best model for each size was then compared using the RMSE based on the holdout set. The model with the lowest RMSE was the model size we would use. Once the best-sized model and coefficients were selected using the training dataset, the prediction of the holdout dataset was computed. The predictions were then analysed to gain insight about characteristics of the donors.

Model 3 is a logistic regression, used because the response is binary (“Did the entity make a donation this year?”). This statistical model is more computationally demanding to fit, so we generated a model using a simplified procedure. Specifically, we generated a large model with all predictors of interest as well as many interactions. Then, we fit this full model on the training set, and pared it down by retaining those predictors that appeared to have the largest association with the response. The various models that were generated were compared based on the performance of the model with respect to the validation set, using the area under the ROC curve as the criterion.

Model 1: Response = log(total amount given)

For this model we looked at the total amount that each entity has given. As can be seen in the left part of Figure 9, the distribution of total donations by entity is heavily right-skewed. Since modeling typically works better for response variables that are more symmetric, we took the log of total donations and constructed a model to predict this quantity. Many of the derived predictors having to do with giving behavior of an entity over time (e.g. GradToGive and GaveLastYear) cannot be used in this model. However, one predictor that may be important is the length of time between an entity's graduation and the year they first gave. Thus, "Year.1st.Give" is considered in this model.

Figure 9: Distribution of Total Amount Given (left) and Distribution of log(Total Amount Given) (right).

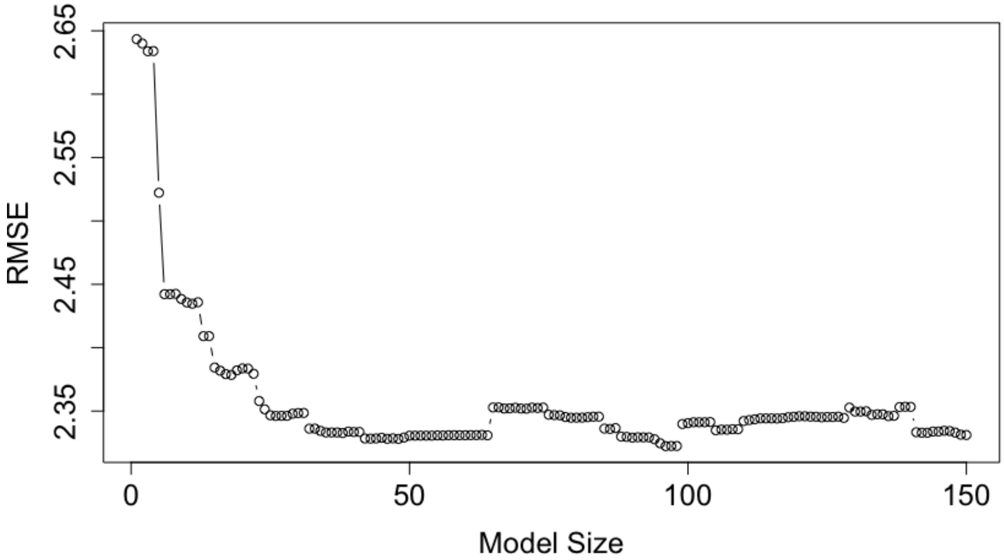


In Table A.1 in the appendix, we include a column that specifies all of the predictors considered for this model. We then used the forward selection procedure described above to choose a final model. Figure 10 shows the plot of the performance of the best models of each size (based on the training dataset) and their performance in terms of RMSE based on the validation set. It is clear from this plot that very small models (i.e. models with less than 10 or 12 predictors) are inferior in terms of their predictive capability. However, once you get to about 40 predictors, there is little to be gained by additional predictors. We chose 40 as the size of our model.

This 40-variable model was applied to the test dataset and Table A.4 shows the predictors that were selected to be included in the final model. On average, our predictions missed the actual values by a factor of about 4 or 5. This suggests the difficulty in predicting the total donations of an entity in the aggregate.

Using this model to predict the outcomes of how much a participant is likely to donate, and to have a better understanding of the model, we decided to look specifically at profiling 'large donors' and 'small donors' using the 40 parameter model's predicted values. To do this we defined large donors as being someone predicted to donate more than $2,000 and small donors as someone predicted to donate less than $50. Using the final model we were able to see what levels of variables were important when predicting who would give a lot of money and who wouldn't give much.

Figure 10: Out-of-sample root mean squared error of prediction of log(total amount given) for the best models of varying sizes.



We found that 'large donors' were more likely to have graduated from Farmer Business School followed by Arts and Science, were from the Midwest and 70.12% of these large donors were involved in Greek life. Roughly 27.4% of these donors were former athletes. Additionally, 78.65% of these donors were male and 71.34% of these donors were married. It was also apparent that donors who gave more were likely to be involved in Service Events, Alumni Events and generally were more involved during their time at Miami as well as an Alumni. Of the 164 donors we predicted to be large, 97.56% had graduated before 1994.

In contrast, we found that 'small donors' were more likely to have graduated from Education and Health or Arts and Science , were from Midwest and only 29.47% were involved in Greek life. Around 2.57% of these donors were former athletes. Largely contrasting to 'big' donors, 64.24% of the predicted small donors were female and only 0.1% of these predicted small donors were married. Of the 4,119 donors we predicted to be small, 93.81% had graduated within the last 10 years; this may shed some light on the small proportion of married entities predicted to be small donors. Contrasting to 'large donors', 'small donors' were less likely to be involved in Service Events, Alumni Events and generally were less involved during their time at Miami as well as an Alumni. Table 1 summarizes these observations, and includes a comparison with the larger population of donors as represented in the holdout set (which, if you recall, is randomly chosen and roughly half of the full dataset).

Model 2: Modelling the log(Next Yearly Donation)

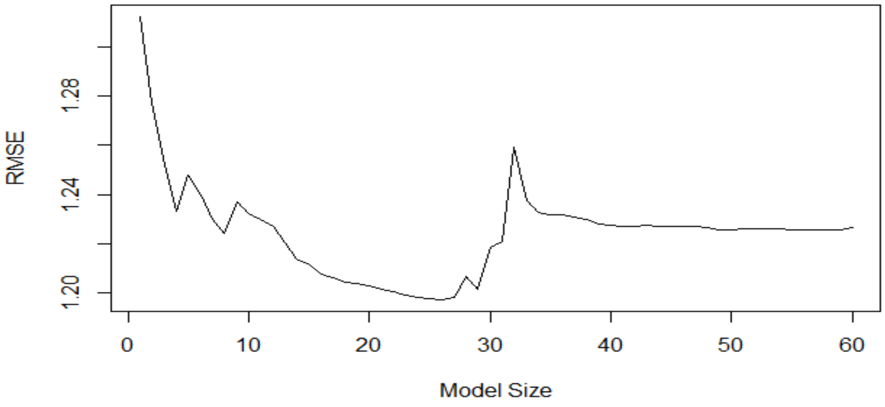
We now turn to modeling the size of the next yearly donation for a particular entity. That is, the response is the total amount given in the next year that a donation is made. For the same reasons as for Model 1, we model the log of the response, and as before we trained models on the training dataset and then made predictions on the hold-out dataset. To simplify the modeling, any years that a particular entity did not donate were excluded from the data; that is, only years that included non-zero donations were used to fit a model.

Table 1. Summary characteristics of predicted large and small donors from Model 1.

	Big Donors (n=164)	**Small Donors (n=4,119)**	**Full Holdout Dataset (n=34,607)**
School	Farmer (65.85%)	A&S (42.12%), Educ/Health (22.58%)	Farmer (35.29%), A&S (37.99%), Educ/Health (15.71%)
Region	Midwest (70.12%)	Midwest (71.69%)	Midwest (63%)
Athlete	Athlete (27.4%)	Athlete (2.57%)	Athlete (6.51%)
Greek	Yes (52.43%)	Yes (29.47%)	Yes (40.47%)
Gender	Male (78.65%)	Male (35.76%)	Male (43.98%)
Marriage	Married (71.34%)	Not Married (99.9%)	Married (24.03%)
Year of Graduation	<2006 (97.56%)	>2006 (93.81%)	<2006 (78.65%)

Due to the number of variables and the size of data set, we could not explore all possible two-way interactions between predictors. Hence we started with a "full model" that included all predictors specified as "Model 2" in Table A.1, along with some interactions that we believed might be important. Once we established this set of possible predictors, we performed forward selection in a similar way to the development of Model 1. Figure 11 shows the results of this procedure, and we see that the 26-predictor model has the lowest out-of-sample prediction error. This prediction error is around 1.2 and this means that, very roughly, we have an average error predicting log(Next Yearly Donation) of about 1.2. Empirically, we found that in terms of Next Yearly Donation, our predictions were in error, on average, by a factor of about 1.0563.

Figure 11: Plot of root mean squared errors of prediction for models of various sizes.



Using this model to predict how much a participant is likely to donate, and to develop a better understanding of the model, we also looked at large donations (predicted to exceed $1,000) and small donations (predicted to be less than $50). Using the final model, we were able to see what levels of variables were important when predicting who would give a lot of money (large donor) and who wouldn't give much money (small donor) during the next year they donated.

We found that large donations were more likely to have come from individuals who graduated from Farmer Business School, Arts and Science or Education and Health, were from the Midwest and 50% were involved in Greek life. Roughly 35% of these donations were made by former athletes. Additionally, 58.2% of these donations were made by males and 44.5% of these donations were made by married people. It was also apparent that donors who gave more were likely to be involved in Service Events, Alumni Events and generally were more involved during their time at Miami and as Alumni. Large predicted donations come from donors who were very likely to have given in the previous year, and whose last donation was quite large on average.

In contrast, we found that small donations were more likely to have come from individuals who graduated from Art & Science and Education and Health, most of these donors were from Midwest, South, Northeast, West and other areas, and 36.4% of the donations were made by people involved in Greek life. Around 1.7% of these donations were made by former athletes, 73% made by females, and only 22.5% of the donations were made by married people. Contrasting to large donors, small donors were less likely to be involved in Service Events, Alumni Events and generally were less involved during their time at Miami and as Alumni. Small predicted donations come from donors who were relatively unlikely to have given in the previous year, and whose last donation was small on average. Table 1 provides some comparative summaries and useful statistics. It can be seen that the big and small donations are clearly distinguished from the more general population ("Full Holdout Dataset") in many of the characteristics.

Table 2. Summary characteristics of predicted large and small donations from Model 2.

	Big Donations (n=98)	**Small Donations (n=11,814)**	**Full Holdout Dataset (n=69,896)**
School	Farmer (68.36%)	A&S (45.90%), Educ/Health (26.63%)	Farmer (40.25%), A&S (35.11%), Educ/Health (14.65%)
Region	Midwest (61.22%)	Midwest (65.78%)	Midwest (66%)
Athlete	Athlete (25.51%)	Athlete (1.71%)	Athlete (7.0%)
Greek	Yes (53.06%)	Yes (36.64%)	Yes (44.56%)
Gender	Male (59.18%)	Male (27.09%)	Male (44.52%)
Marriage	Married (40.82%)	Not Married (77.49%)	Married (33.82%)
Year of Graduation	<2006 (100%)	>2006 (70.37%)	<2006 (91.87%)
Last Donation	Mean: $38,313.08 Median: $21,750	Mean: $29.38 Median: $0	Mean: $238.27 Median: $50
Gave Last Year?	Yes: 95.9%	Yes: 16.92%	Yes: 53.06%

As an illustration, in our fitted model consider the giving profile of a selected predicted big donor with the following characteristics: female, not in the Greek system, athlete, graduated in 1987 with an undergraduate degree from FSB, living and unmarried from the midwest, involved in five service events, three alumni events, and nine total events. Based on this profile, Figure 12 shows the predicted amount given by using our model and varying the years after graduation over time. Similarly, Figure 13 shows the giving profile of a selected predicted small donor with the following characteristics: non-athlete unmarried female from the Northeast, one degree, graduated in 1989 from Other school category with an undergraduate degree, with no service event and a total number of four events or activities.

Figure 12. Profile plot of selected big donor, based on Model 2 predictions.

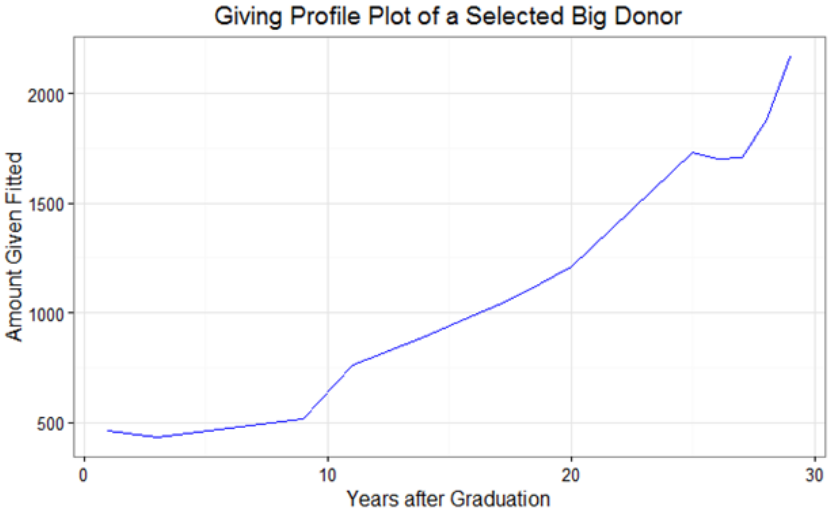
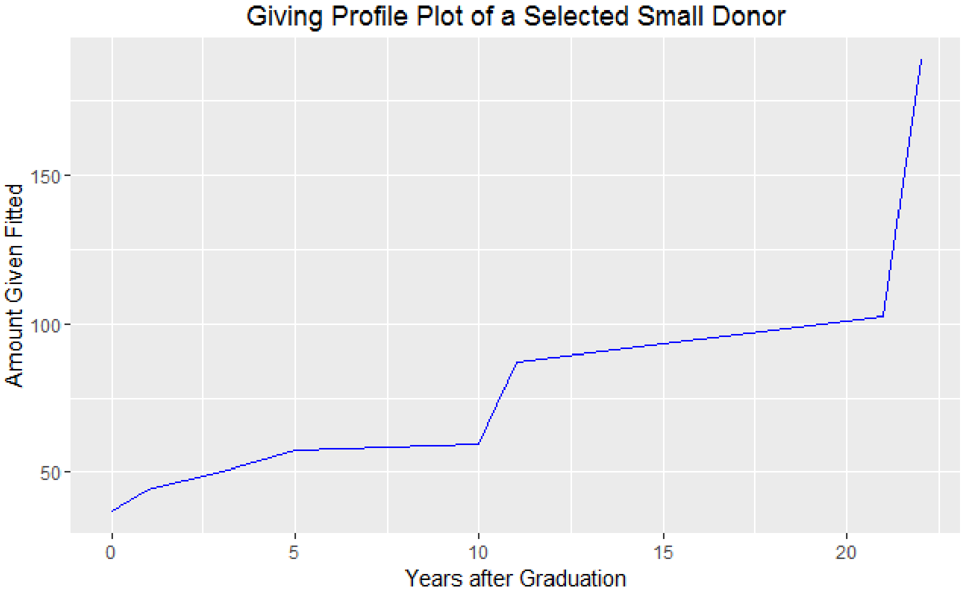


Figure 13. Profile plot of selected small donor, based on Model 2 predictions.



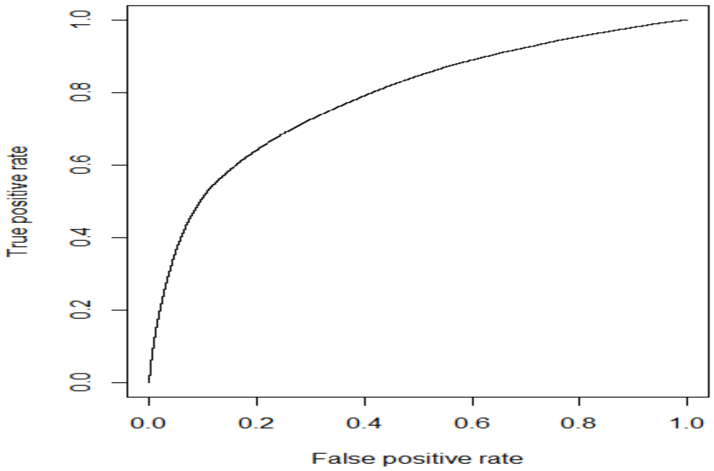
Model 3: Logistic Regression Model to Predict Probability of Giving

We wish to predict the probability of a person giving during a given year based on their characteristics, giving history, and participation in activities. To select a model, we first fit a “full model” consisting of all the predictors in the last column of Table A.1, along with some interactions that we thought might be important. We then iteratively removed predictors that did not appear strongly associated with the probability of giving. We evaluated these models using the area under the ROC curve when applied to the validation dataset. A final model was chosen as the one with the largest area under the curve. The final model can be seen in Table A.6 in the appendix.

Interpreting logistic regression coefficients here is somewhat complicated because predictors may be included in interactions. An interaction means that an association between a predictor and the probability of giving may change depending on the value of another predictor. Because of this, we will not attempt to interpret particular coefficients, but instead focus on the predictions our model makes.

Figure 14 gives the ROC curve, which displays the tradeoff between false negatives and false positives for different cutoff values. That is, if you choose a particular predicted probability as the cutoff between classifying an entity as "donor next year" or "not donor next year", you will make a certain number of correct and incorrect classifications. (One reasonable choice for a cutoff would be 0.5 - any predicted probability larger than that would predict that the entity would give next year, otherwise predict that they would not give.) The ROC curve evaluates the ability of the model to correctly classify for many different cutoffs, and displays the results. The area under the ROC curve for the final model was around 0.79 which suggests that the model is reasonably good at classifying entities into give/not give.

Figure 14: ROC curve for final logistic regression model.



Using this final model to predict how likely a participant is to donate, and to develop a better understanding of the model, we decided to look specifically at entity/year combinations for which predictions were relatively likely and unlikely. To do this we defined highly likely donors as being someone who has a predicted probability of 70% or more to give in a particular year and 'unlikely donors' as someone who has a probability of less than 10% to give in a particular year. Using the final model we were able to see what levels of variables were important when predicting who is most likely to give and who is unlikely to give in a particular year.

We found that highly likely donations were from people more likely to have graduated from Farmer School of Business, Arts and Science or Education and Health, from the Midwest. Of the years predicted to have more than 70% likelihood of a gift, 50% were involved in Greek life, roughly 7.84% of them were former athletes, 49% of were male and 57% were married. It was also apparent that a relatively large predicted probability of giving were associated with those who were most likely to be involved in Service Events, Alumni Events and generally were more involved during their time at Miami and as Alumni. For those predicted likely givers, we found that their last donation, on average, was larger than those in the full population, and they were much more likely to have given previously.

In contrast, we found that less likely donations were from people more likely to have graduated from Arts & Science, Education and Health and the Business School, most of these donors were from Midwest, South, West, Northeast, and other areas 37% were involved in Greek life. Around 5% of these donors were former athletes. 43% of unlikely donors were male and only 57% of these small donors were unmarried. Contrasting to 'highly likely donors', 'unlikely donors' were less likely to be involved in Service Events, Alumni Events and generally were less involved during their time at Miami and as Alumni. For those predicted unlikely givers, we found that their last donation, on average, was smaller than those in the full population, and they were less likely to have given previously. Table 3 summarizes this information and includes comparison with the full population.

Table 3. Summary characteristics of highly likely and less likely donations.

	More Likely to Give (n=25,982)	**Less Likely to Give (n=247,827)**	**Total (n=677,716)**
Probability	Greater than or equal to 0.70	Less than 0.10	
School	Farmer (50.1%)	A&S (39.95%), Farmer (32.65%)	A&S (37.92%) Farmer (36.23%)
Region	Midwest (73.32%)	Midwest (55.55%)	Midwest (62.02%)
Athlete	Athlete (7.84%)	Athlete (5.21%)	Athlete (6.5%)
Greek	Yes (50.0%)	Yes (36.5%)	Yes (41.3%)
Gender	Male (48.63%)	Male (43.15%)	Male(43.85%)
Marriage	Married (57.15%)	Not Married (56.85%)	Married (26.26%)
Year of Graduation	<2006 (98.77%)	>2006 (0.47%)	>2006 (5.62%)
Last Donation	Mean: $867.40 Median: $200	Mean: $73.91 Median: $40	Mean: $118 Median: $30
Gave Last Year?	Yes: 99.9%	Yes: 0.00%	Yes: 19.78%
Proportion of Previous Years Donated?	Mean: 0.78	Mean: 0.098	Mean: 0.21

In Figure 15, consider an entity from the data with following profile: non-athlete, unmarried female with one degree from the Farmer School of Business, graduated in 1988, from New Jersey (Northeast), consistently gives yearly, an undergraduate, no Greek experience, and not active in service or alumni activities. In Figure 16, we consider a different, but quite similar entity. In fact, the characteristics are the same, except graduation is in 1987 instead of 1990. However, the giving profiles are quite different. The main reason is that the Figure 15 donor gave much more often during the first 16 years post-graduation. After 16 years, she stopped, which caused her predicted probability of giving to decrease substantially. However, the Figure 16 individual did not give for the most part which is why the model predicts such a small probability of giving that decreases over the years.

Figure 15: Profile plot of selected highly likely donor.

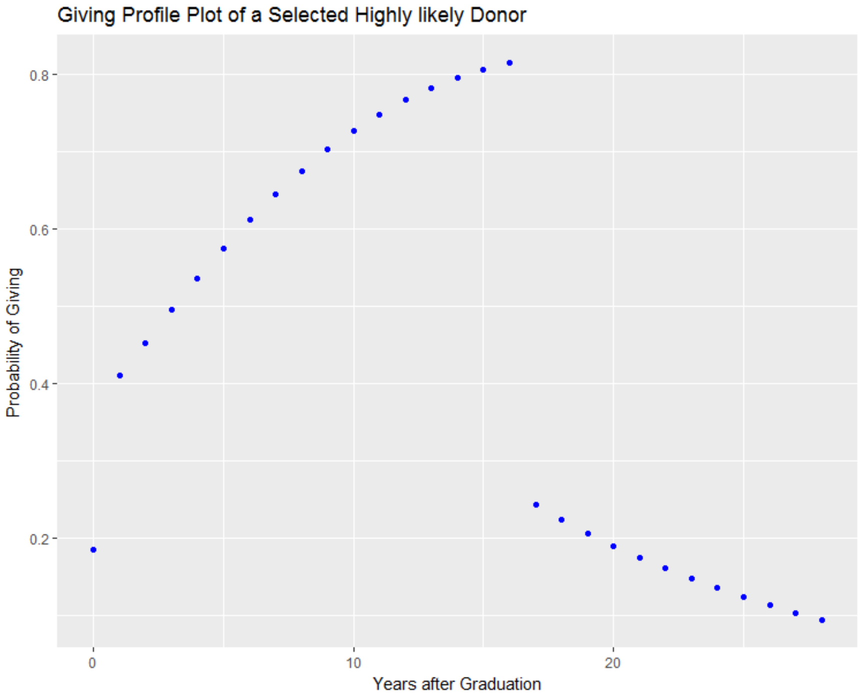
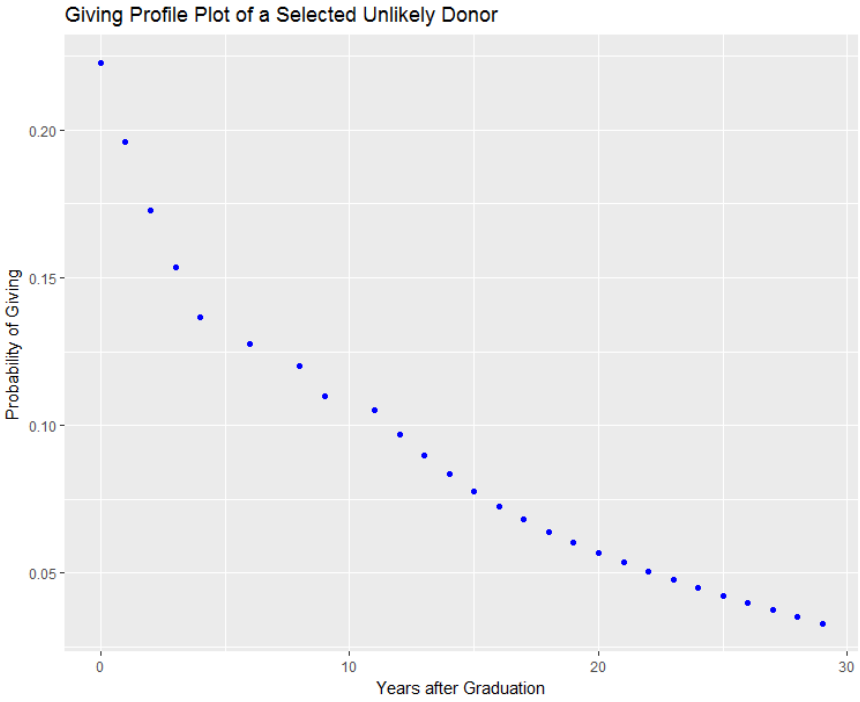


Figure 16: Profile plot of selected less likely donor.



Conclusions

In our results above we have attempted to model and answer 3 major questions:

1. What characteristics determine the total amount given by an individual donor?
2. What characteristics determine the yearly amount given by an individual donor?
3. What's the probability of someone giving in a particular year?

All three models seem to produce similar results when it comes to profiling who is giving back to Miami. As reasonably anticipated, donors that were giving more frequently were noticeably more active when it came to participation in various alumni events. These were highly predictive factors for all three models. Donors who gave more were more likely to have graduated from Farmer Business school or the Arts & Science college. This seems to make sense as Farmer Business school has an excellent reputation for having a competitive program, which could mean students graduating from this school or more likely to earn larger salaries once they enter the job force. The other models however didn’t really profile any difference between the schools. Also, participating in Greek life seemed to be an important factor too for all of the models. Overall, when it comes down to likelihood of giving and how much is being given it is very apparent that the alumni who participated in more activities when a student at Miami and participated more afterwards (i.e. coming to alumni events) were more likely to give more money and give more often. A final important observation, from Models 2 and 3: *the size and frequency of previous gifts play a critical role in the size and likelihood of future gifts to the institution*.

Since our focus was on prediction rather than interpretable models, there was a correlation structure within the data that we ignored. That is, it is likely that the giving patterns of a given entity will be correlated from year to year. We may have captured this to some extent by using predictors like “GaveLastYear”, but the modeling and computational challenges imposed by accounting for this correlation were heavy, especially since predictive accuracy may not have been improved.

Much of the effort for this project was spent in data handling and clean-up. Furthermore, the size of the data made model-fitting and selection a challenge. Because of these things and the time constraints imposed by the semester, we were not able to address many of the specific questions posed by the client at the outset. We hope that this report does not represent the end of this project, but only a midpoint. Perhaps a future STA 660 class could pick it up, or an undergraduate or M.S. student in the statistics department could use this as a launching point for a project of their own. Here are several ideas for future work: (1) the analysis of the predictions from our models focused on entity characteristics, but more could be understood about the giving characteristics of large/likely and small/unlikely donors (for instance, how likely were likely donors to have given in the previous year? Or how much was a large donor’s last gift, on average?); (2) more focused consideration of specific questions that the Advancement Office has; (3) additional predictive modeling methods; and (4) an interactive software application to allow personnel at the Advancement Office to provide inputs regarding characteristics of an entity and receive predictions that would calibrate expectations for the giving profile of that donor.

Appendix

Table A.1: Table and List of variables

Variable	Description	log(Total amount given)	Log (Next Yearly Donation)	Logistic model
Entity	Unique ID of a donor			
State	State that the Entity resides in			
Year	Year of Transaction			
Credit	The amount donated for a single transaction.		Model 2	
Gave	Indicator variable (1 if a person gave in a year and 0 otherwise)			
Athlete	The Alumni's status of being a former athlete or not.	Model 1	Model 2	Model 3
Degree.Count	The number of degrees that the entity has completed.	Model 1		Model 3
Degree.Year.1	The graduation year of the entities first degree.	Model 1	Model 2	Model 3
School	The school that the Entity belongs to. (Farmer, A&S, Health, Computing etc.)			
Relationship.Status	Whether the donor is living or deceased.	Model 1	Model 2	Model 3
Primary.Relationship	What type of degree the entity has.	Model 1	Model 2	Model 3
Gender	Male or female.	Model 1	Model 2	Model 3
Married	An indicator variable (1 if married and 0 if not)	Model 1	Model 2	Model 3
Country	Country the entity is from.			
City	City the entity is from.			
Region	US Region the entity is from (South, Midwest Northeast, West)		Model 2	Model 3
School.Category	Collapsed smaller schools together for school variable.	Model 1	Model 2	Model 3
GradToGive	The number of years from graduation to giving/transaction		Model 2	Model 3
GaveLastYear	Indicator variable (Yes or No)		Model 2	Model 3
GaveLastYear.Amt	Amount given the previous year prior to year under consideration			
LastGave	Amount given in the last donation			Model 3
Degree.Category	Classified as Graduate, Undergraduate or Others		Model 2	Model 3
SRVCE	The number of service events entity has been to	Model 1	Model 2	Model 3

REUN	The number of reunion events entity has been to	Model 1	Model 2	Model 3
ALUEV	The number of alumni events entity has been to	Model 1	Model 2	Model 3
CHAPT	The number of chapter events entity has been to	Model 1		Model 3
FRTTY	Indicator variable (1 = participated in a Fraternity, 0 = no participation)	Model 1		
SOROR	Indicator variable (1 = participated in a Sorority, 0 = no participation)	Model 1		
Greek	Indicator variable (1 = participated in a Greek Culture, 0 = no participation)	Model 1	Model 2	Model 3
Part.Level	The total number of events/activities that the entity has been involved with.	Model 1	Model 2	Model 3
GradToGive.2	The number of years from graduation to giving/transaction squared			
GaveLastYear.Amt.Log	Natural log of the amount given the previous year prior to year under consideration			
Year.1st.Give	The number of years between when an entity graduated and when they first donated money.	Model 1		
PropGave	The proportion of years the individual gave out of the total number of years since 1987.			Model 3

Table A.2 : List of Data files

Entity	Contains information about individual entities
athletics	Contain information about athletic activities of individual entities
giving	Contains transactional information on giving of individuals
degree	Contains a description of degrees and the codes
contact information	Contains contact information of various entities and individuals
relationship type	Contains a description and codes of relation type
participation history	Contains information about entities participation in various activities
customer specific	Contains information on score and dates of individual entities
appeals	Contains information about various types of appeals made to entities
contact restrictions	Contains information on the various restrictions on individual entities
translation codes	Contains a description of various abbreviation

Table A.3 : Summary of Donations by Region

Region	Donors	Total ($)	Mean ($)	Median ($)	Max ($)
Midwest	21,803	24,602,296.1	1,128.39	150	351,397.4
Northeast	2,436	4,904,368.3	2,013.29	170	500,050
South	5,531	7,621,074.7	1,377.88	175	555,970
West	2,470	4,797,554.5	1,942.33	150	1,504,614.6
Other	2,367	903,382.2	381.66	100	44,105

Table A.4. Coefficients for the best model (40 predictors) Model 1: log(total amount given)

```
                             (Intercept)                             Degree.Year.1 
                            1.681526e+02                             -8.140213e-02
                      AthleteNon-athlete                    DegreeCatUndergraduate
                           -6.749164e-01                             -5.447800e-01
                                 Service                                 PartLevel
                            4.544431e+01                              1.142094e-01
                        Married1:GenderM      Married1:SchoolEngineering/Computing
                           -2.871519e-01                              6.998734e-02
                  Married1:Degree.Year.1               Married1:AthleteNon-athlete
                            3.825524e-04                              3.664956e-01
                  Married1:Year.1st.Give            GenderM:SchoolEducation/Health
                           -6.411829e-02                              8.221173e-02
                   GenderM:Degree.Year.1                         GenderM:AlumEvent
                            6.685365e-05                              5.255663e-01
                       GenderM:ChapEvent                     GenderM:Year.1st.Give
                            1.625884e+00                             -2.201812e-03
     SchoolEducation/Health:Degree.Count                  SchoolOther:Degree.Count
                           -1.909578e-01                             -2.798255e-01
         SchoolFarmer:AthleteNon-athlete       SchoolFarmer:DegreeCatUndergraduate
                            5.381531e-02                              3.138354e-01
   SchoolEducation/Health:RegionNortheast          SchoolEducation/Health:AlumEvent
                            4.309470e-02                              2.875855e-01
SchoolEngineering/Computing:Year.1st.Give                 SchoolOther:Year.1st.Give
                            2.501007e-04                              2.596656e-03
             Degree.Count:DegreeCatOther       Degree.Count:DegreeCatUndergraduate
                           -4.497754e-01                              1.688907e-01
               Degree.Year.1:RegionSouth                     Degree.Year.1:Service
                           -8.856583e-05                             -2.266155e-02
          AthleteNon-athlete:RegionSouth                DegreeCatOther:RegionOther
                            2.477007e-01                             -3.693096e-01
                  DegreeCatOther:Service         Relationship.Statusliving:Service
                            3.224352e+00                              2.180066e-01
                    RegionWest:AlumEvent                          RegionWest:Greek
                            2.256876e-02                             -9.614029e-02
                           Service:Greek                     RegionOther:ChapEvent
                            5.534381e-02                              0.000000e+00
                   RegionSouth:ChapEvent                      RegionWest:ChapEvent
                           -2.757573e-01                             -3.508914e+00
                     RegionSouth:Reunion                         Service:ChapEvent
                           -1.955308e-02                              6.513234e-01
                   Reunion:Year.1st.Give
                            7.568357e-02
```

Table A:5. Results for Model 2: Modeling the Yearly log(Credit given)

```
                              (Intercept)                                      GenderM
                             3.757764e+00                                 1.181218e-01
                               GradToGive                              GaveLastYearYes
                            -5.658791e+00                                 1.023052e-01
        School.CategoryEngineering/Computing                     School.CategoryFarmer
                             4.028406e-01                                 1.974599e-01
                    School.CategoryOther                            AthleteNon-athlete
                            -1.446251e-01                                 3.795265e+01
                                    SRVCE                                   Part.Level
                             1.280088e-01                                 3.095853e-02
                     Degree.Year.1:ALUEV                  LastGave:AthleteNon-athlete
                             1.543895e-04                                 4.443927e-05
                     LastGave:GradToGive                          LastGave:RegionOther
                             2.702403e-06                                 3.470537e-04
                    LastGave:RegionSouth               LastGave:School.CategoryOther
                            -6.238480e-05                                -3.102162e-05
                GradToGive:RegionNortheast                      GradToGive:RegionSouth
                             1.271628e-02                                 7.459908e-03
                    GradToGive:RegionWest                  GradToGive:GaveLastYearYes
                             1.127206e-02                                 2.079209e-02
     GradToGive:School.CategoryEducation/Health GradToGive:School.CategoryEngineering/Computing
                            -1.070415e-02                                -2.290349e-02
             GradToGive:Degree.CategoryOther                     GradToGive:Degree.Year.1
                            -2.645987e-02                                 2.862641e-03
            Degree.Year.1:AthleteNon-athlete            School.CategoryFarmer:Part.Level
                            -1.914553e-02                                 2.040425e-02
                    AthleteNon-athlete:ALUEV
                            -2.094049e-01
```

Table A.6. Model 3: Logistic Model to Predict Probability of Giving

```
> summary(modFinal)

Call:
glm(formula = Gave ~ GradToGive + GaveLastYear + Gender + Athlete + 
    Relationship.Status + Married + Region + School.Category + 
    LastGave + Degree.Category + SRVCE + ALUEV + CHAPT + REUN + 
    Greek + Part.Level + GradToGive.2 + propGave * (GradToGive + 
    GaveLastYear + Gender + Athlete + Relationship.Status + Married + 
    Region + Degree.Category + SRVCE + ALUEV + CHAPT + Greek + 
    Part.Level + GradToGive.2) + GradToGive * (GaveLastYear + 
    Gender + Athlete + Relationship.Status + Region + School.Category + 
    LastGave + Degree.Category + SRVCE + REUN + Greek + Part.Level + 
    GradToGive.2) + GaveLastYear * (Gender + Athlete + +Region + 
    School.Category + ALUEV + CHAPT + Greek + Part.Level + GradToGive.2), 
    family = "binomial", data = Jones)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7845  -0.5831  -0.4363  -0.3131   3.0491  

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                                 -1.455e+00  9.709e-02 -14.984  < 2e-16 ***
GradToGive                                  -1.566e-01  9.263e-03 -16.905  < 2e-16 ***
GaveLastYearYes                              3.731e-01  5.077e-02   7.350 1.99e-13 ***
GenderM                                     -9.431e-02  1.325e-02  -7.117 1.10e-12 ***
AthleteNon-athlete                           9.745e-03  2.565e-02   0.380 0.703962    
Relationship.Statusliving                   -1.465e-01  8.861e-02  -1.654 0.098227 .  
Married1                                     1.327e-01  1.097e-02  12.090  < 2e-16 ***
RegionNortheast                             -3.156e-02  2.524e-02  -1.250 0.211154    
RegionOther                                  1.927e-01  2.886e-02   6.676 2.45e-11 ***
RegionSouth                                 -4.743e-02  1.769e-02  -2.681 0.007330 ** 
RegionWest                                  -1.421e-01  2.564e-02  -5.544 2.96e-08 ***
```

```
School.CategoryEducation/Health                          6.965e-02  1.863e-02   3.738 0.000186 ***
School.CategoryEngineering/Computing                     3.523e-01  2.711e-02  12.997  < 2e-16 ***
School.CategoryFarmer                                    2.315e-01  1.415e-02  16.358  < 2e-16 ***
School.CategoryOther                                    -1.666e-03  2.968e-02  -0.056 0.955245
LastGave                                                 5.625e-06  4.106e-06   1.370 0.170666
Degree.CategoryOther                                    -6.606e-01  8.297e-02  -7.963 1.69e-15 ***
Degree.CategoryUndergraduate                             1.421e-01  2.628e-02   5.406 6.44e-08 ***
SRVCE                                                    9.073e-02  1.741e-02   5.212 1.87e-07 ***
ALUEV                                                    9.341e-02  1.428e-02   6.543 6.05e-11 ***
CHAPT                                                    2.303e-02  1.968e-01   0.117 0.906874
REUN                                                    -3.532e-01  2.008e-01  -1.759 0.078635 .
Greek                                                    8.588e-03  1.437e-02   0.598 0.550152
Part.Level                                               4.071e-02  5.681e-03   7.165 7.79e-13 ***
GradToGive.2                                             5.321e-03  3.443e-04  15.457  < 2e-16 ***
propGave                                                 1.621e+00  1.937e-01   8.366  < 2e-16 ***
GradToGive:propGave                                      1.813e-01  7.973e-03  22.746  < 2e-16 ***
GaveLastYearYes:propGave                                 4.964e-01  3.362e-02  14.764  < 2e-16 ***
GenderM:propGave                                        -3.533e-02  3.011e-02  -1.173 0.240687
AthleteNon-athlete:propGave                             -1.663e-01  6.046e-02  -2.750 0.005958 **
Relationship.Statusliving:propGave                      -1.061e+00  1.683e-01  -6.304 2.91e-10 ***
Married1:propGave                                        2.334e-01  2.688e-02   8.682  < 2e-16 ***
RegionNortheast:propGave                                 1.988e-02  5.868e-02   0.339 0.734789
RegionOther:propGave                                    -8.055e-01  7.553e-02 -10.665  < 2e-16 ***
RegionSouth:propGave                                     1.026e-01  4.108e-02   2.497 0.012510 *
RegionWest:propGave                                      1.824e-01  6.072e-02   3.003 0.002672 **
Degree.CategoryOther:propGave                           -2.171e-01  1.792e-01  -1.211 0.225763
Degree.CategoryUndergraduate:propGave                   -3.889e-01  5.482e-02  -7.095 1.29e-12 ***
SRVCE:propGave                                           1.932e-01  3.212e-02   6.015 1.80e-09 ***
ALUEV:propGave                                           4.243e-01  4.029e-02  10.530  < 2e-16 ***
CHAPT:propGave                                           8.882e-01  5.224e-01   1.700 0.089070 .
Greek:propGave                                           1.893e-01  3.253e-02   5.818 5.94e-09 ***
Part.Level:propGave                                     -8.485e-02  1.301e-02  -6.520 7.02e-11 ***
GradToGive.2:propGave                                   -3.997e-03  3.136e-04 -12.744  < 2e-16 ***
GradToGive:GaveLastYearYes                               1.563e-01  5.373e-03  29.098  < 2e-16 ***
GradToGive:GenderM                                       8.494e-03  1.004e-03   8.459  < 2e-16 ***
GradToGive:AthleteNon-athlete                           -1.089e-02  1.916e-03  -5.684 1.31e-08 ***
GradToGive:Relationship.Statusliving                     2.515e-02  7.910e-03   3.180 0.001473 **
GradToGive:RegionNortheast                               4.968e-04  1.863e-03   0.267 0.789735
GradToGive:RegionOther                                  -5.113e-02  2.802e-03 -18.248  < 2e-16 ***
GradToGive:RegionSouth                                  -3.498e-03  1.319e-03  -2.653 0.007978 **
GradToGive:RegionWest                                   -5.303e-04  1.925e-03  -0.275 0.782938
GradToGive:School.CategoryEducation/Health              -5.433e-03  1.510e-03  -3.597 0.000321 ***
GradToGive:School.CategoryEngineering/Computing         -3.007e-02  2.226e-03 -13.510  < 2e-16 ***
GradToGive:School.CategoryFarmer                        -1.681e-02  1.127e-03 -14.921  < 2e-16 ***
GradToGive:School.CategoryOther                         -4.012e-03  2.461e-03  -1.630 0.103014
GradToGive:LastGave                                      1.345e-07  2.549e-07   0.528 0.597813
GradToGive:Degree.CategoryOther                          5.105e-02  6.362e-03   8.024 1.02e-15 ***
GradToGive:Degree.CategoryUndergraduate                 -9.527e-03  2.017e-03  -4.723 2.32e-06 ***
GradToGive:SRVCE                                        -4.662e-03  1.385e-03  -3.367 0.000760 ***
GradToGive:REUN                                          2.250e-02  1.459e-02   1.542 0.123077
GradToGive:Greek                                        -3.587e-03  1.089e-03  -3.293 0.000993 ***
GradToGive:Part.Level                                    3.308e-03  4.436e-04   7.458 8.77e-14 ***
GradToGive:GradToGive.2                                 -7.917e-05  9.071e-06  -8.728  < 2e-16 ***
GaveLastYearYes:GenderM                                  5.751e-02  1.827e-02   3.148 0.001643 **
GaveLastYearYes:AthleteNon-athlete                       1.765e-01  3.492e-02   5.055 4.31e-07 ***
GaveLastYearYes:RegionNortheast                         -6.156e-03  3.480e-02  -0.177 0.859606
GaveLastYearYes:RegionOther                              1.762e-01  4.551e-02   3.872 0.000108 ***
GaveLastYearYes:RegionSouth                              9.344e-03  2.421e-02   0.386 0.699518
GaveLastYearYes:RegionWest                              -3.394e-02  3.573e-02  -0.950 0.342096
GaveLastYearYes:School.CategoryEducation/Health         -3.311e-02  2.275e-02  -1.456 0.145477
GaveLastYearYes:School.CategoryEngineering/Computing     8.936e-02  3.248e-02   2.751 0.005940 **
GaveLastYearYes:School.CategoryFarmer                    6.240e-02  1.703e-02   3.664 0.000248 ***
GaveLastYearYes:School.CategoryOther                    -1.279e-01  3.774e-02  -3.389 0.000701 ***
GaveLastYearYes:ALUEV                                   -1.639e-01  2.408e-02  -6.806 1.01e-11 ***
```

```
GaveLastYearYes:CHAPT              -              1.584e-01  3.156e-01  -0.502 0.615710
GaveLastYearYes:Greek                            -1.191e-01  1.924e-02  -6.188 6.10e-10 ***
GaveLastYearYes:Part.Level                       -2.299e-02  7.081e-03  -3.247 0.001168 **
GaveLastYearYes:GradToGive.2                      -3.523e-03  1.935e-04 -18.211  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 688253  on 677715  degrees of freedom
Residual deviance: 551606  on 677637  degrees of freedom
AIC: 551764

Number of Fisher Scoring iterations: 5

roc.curve(modFinal, Jones)
[1] 0.7882253
```