Exploratory and Statistical Analysis of DonorsChoose.org

Nathan Liittschwager[1], Arielle Thibeault[1], Katherine Yamamoto[1], Alec Guthrie[1], Alex Liebscher[1], and Mike Ona[1]

***Abstract*— DonorsChoose.org, a website where donors can support and fund projects posted by teachers, is faced with the problem of cultivating and retaining donors. As the platform scales to millions of projects, there has been a push to bridge the gap between potential donors and real teachers. Since its founding in 2000, DonorsChoose has raised nearly $700 million, funding 1.2 million projects and helping over 28 million students across the country. However, the platform could be optimized to elicit as many donations as possible for the thousands of teachers in need of classroom resources and funding. We analyze six publicly available DonorsChoose datasets to answer a variety of questions regarding donation cultivation, retention, and conversion. We uncover demographic, geographical, and behavioral patterns underlying the donor; analyze time-dependent variables on a micro and macro scale; and examine donor and project characteristics critical to project funding. Ultimately, we wish to provide useful information and models to guide the platform toward a more effective email marketing strategy.**

I. INTRODUCTION

In 2015, the Education Market Association reported that, on average, most teachers spend about $500 out of pocket to pay for school and classroom supplies (White). The Treasury Department estimated that during the 2016 tax year, the federal government paid approximately $210 million in tax revenue stemming from classroom expense write-offs. But this figure is unrepresentative of the lengths to which teachers go to preserve the quality of their students' education as the tax code only allows teachers to write off up to $250 in classroom expenses (Figueroa). To combat this prevalent issue, Charles Best, a public high school teacher in the Bronx, created DonorsChoose.org, a website where teachers could post classroom project requests and donors could support and fund any project that inspires them. Since its founding in 2000, DonorsChoose has raised nearly $700 million, funding 1.2 million projects and helping over 28 million students across the country. As of now, 429,307 teachers have received funding from 3.2 million donors, most of whom had never donated to public schools before being exposed to the crowdfunding website. Reaching 78% of public schools in the US, DonorsChoose is the most impactful platform supporting public education (DonorsChoose.org).

While DonorsChooses impact is evident in the statistics, teachers across the nation, especially those in low-income school districts, are still struggling to raise the funds for necessary school supplies. Even after contributions from DonorsChoose.org, public school teachers assume the burden of $1.6 billion in school supply costs due to tighter budgets and the inability of low-income families to afford basic school supplies for their children (White). In order to further assist public school classrooms and teachers, the team at DonorsChoose.org needs to expand the reach of their marketing campaigns. This analysis is aimed at creating a method that connects donors with projects that most inspire them by analyzing the trends of previous donations. This matching method will enable DonorsChoose to build targeted email campaigns that recommend certain projects to prior donors, thus encouraging these donors to make additional donations. If even a fraction of prior donors are motivated to donate repeatedly, DonorsChoose can make an impact on more classrooms.

II. DATA

A. Source

The data is sourced from DonorsChoose.org and contains anonymized information about donations from the past five years. The provided data is split into six different files: projects, schools, teachers, resources, donations, and donors. The projects dataset consists of information (project name and description, school ID, teacher ID, number of projects posted by each teacher, grade level of students for which the projects are designed, project subjects, project costs, current funding status of each project, the posted date, and fully funded date) on all project requests posted between January 1, 2013 and May 1, 2018. The schools dataset contains information about each participating school, including the school ID, name, metropolitan type, percentage of students qualifying for free lunch, state, zip code, city, and district.

Additionally, the teachers dataset consists of the teacher ID, the teacher's prefix, and first project posted date of every participating teacher. The resources dataset

[1]University of California, Mathematics, San Diego, USA

contains the name, price, and quantity of each resource in the project request as well as the name of the vendor from which the resource is purchased. Lastly, and most importantly, the donations and donors datasets provide insight into who is donating and to what projects they are supporting. The donations dataset contains the donor ID, whether or not the donation included an optional donation, donation amount, and how many donations the donor has made while the donors dataset focuses on location of the donor and whether or not they are a teacher.

B. Data Summary

In total, the data consists of a combined dataset including school location and demographics, project statistics and qualitative information, and limited donor information. Of 2.12 million unique donor IDs, 10% of which are also teachers who have posted project requests on DonorsChoose, and 4.69 million unique donation IDs with average donation amount of $60.70. The donors stemmed from 1500 different cities and 1060 unique zip codes across the nation. It should be noted that these figures are not the same because only the first 3 digits of the zip code are provided by the data. Furthermore, donors were more heavily concentrated in large cities, mainly Chicago, New York, Brooklyn, Los Angeles, and San Francisco. As we discuss later though, population size of cities and states easily confounds these metrics.

Additionally, 28.4% of donors have donated more than once. In the last year of donations, the mean cumulative donation amount per donor was $129 (median of $50), with six unique donors donating in total over $200,000. 17 donors had over 1,000 individual donations, however the mean number of donations is 2.0 and the median is 1.0. Donors donate on average $58 (median of $38). The dataset is heavily skewed by the outlying donors and donations, thus medians are used when relevant as a more robust central tendency measure.

The schools dataset provides information on 7,300 unique school IDs. 31% of these schools are located in suburban areas, 31% in urban areas, 18% in rural areas, 8% in towns, and the metro type of 11% of the schools was not reported. 12% of the schools are located in California while other schools are heavily concentrated in the states Texas, New York, Florida, and Illinois. Additionally, 1100 different school districts are accounted for with more project requests in the New York City School District, the Los Angeles Unified School District, and the Miami-Dade County School District. Lastly, on average, 57.6% of the students at these participating schools qualify for a free lunch. Fig 1 illustrates the range of percent of students participating in the free lunch program per school metropolitan type. This demonstrates DonorsChoose's commitment to assisting low-income public schools.

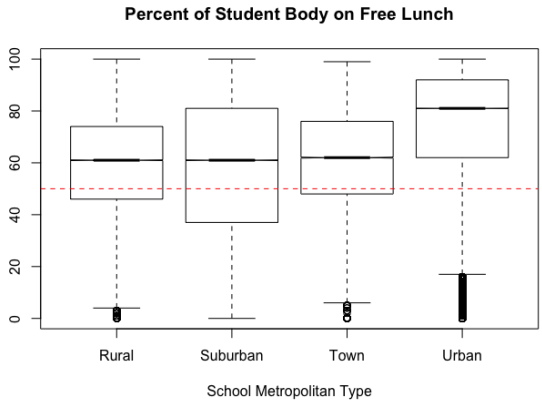

Fig. 1: The percent of the student body utilizing free lunch per metropolitan type from the DonorsChoose data. Note that the average for each metropolitan type is above 50% indicating that projects posted on DonorsChoose are typically of schools in low-income neighborhoods.

From the given datasets, we can immediately see distinct patterns. Of the 8 project subjects (Applied Learning, Health & Sports, History & Civics, Literacy & Language, Math & Science, Music & The Arts, Special Needs, and Warmth, Care & Hunger), the most and least common project types are Literacy & Language with 410,321 occurrences and Warmth, Care & Hunger with 10,039 occurrences respectively with Math & Science as the second most common project subject (217,896 occurrences). The proportion of projects with the subject Literacy & Language is 0.3842 with 95% confidence interval [0.3834, 0.3849] while the proportion of projects of the subject Math & Science is 0.2040 with a 95% confidence interval [0.2034, 0.2046]. Together, the top two subjects make up about 60% of the project population.

C. Text Summary

Text mining was used to summarize and visualize the projects text. The project titles, essays, descriptions, and need statements were each separated by funding status to visualize any differences between projects that were fully funded and those that expired. Each grouping was restructured into a tidy text format through tokenization to create single column tables with one token per row. Data in this format allows for the removal of punctuation as well as stop words (common English words that provide little explanation of the data). The text was

then converted to lower case for easier comparisons with other text datasets.

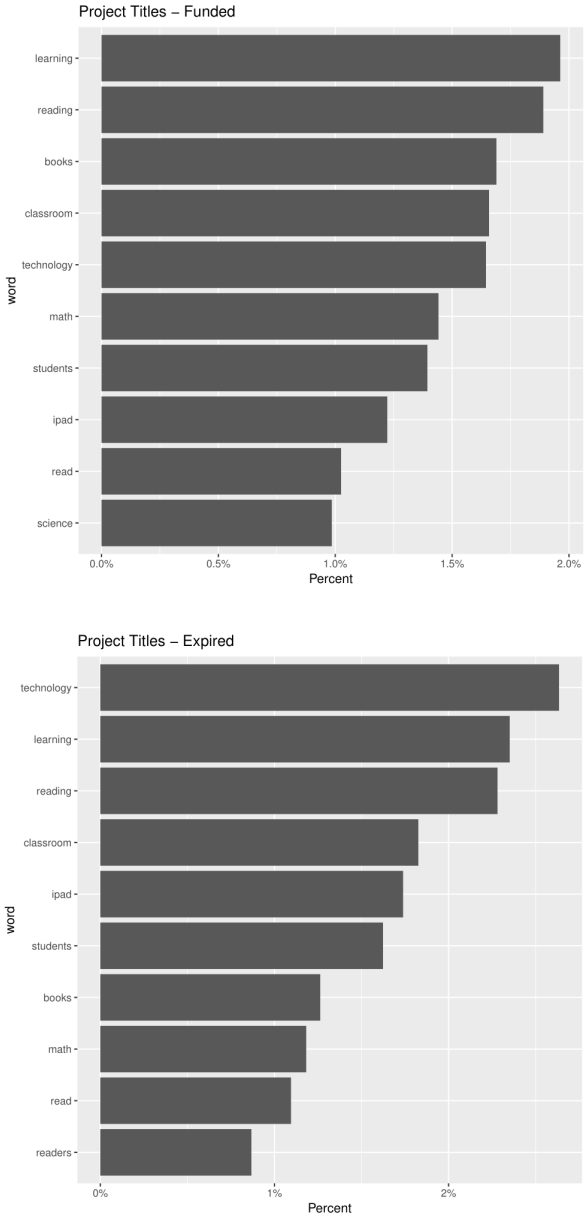

Fig. 2: The top ten most frequent words in funded project titles.

Fig 2 shows the top ten most frequent words in the project titles. Technology and iPad are more prevalent in projects that have expired. Fig 3 shows a word cloud of the top one hundred words in the titles of fully funded projects. Table I provides the frequencies and percents of the top ten words in this word cloud.

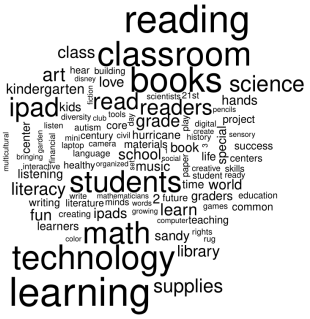

Fig. 3: Word cloud of the top one hundred words in the titles of funded projects.

TABLE I: Frequencies and percentages of the top 10 words in funded project titles.

Word	Frequency	Percent
learning	1488	1.96
reading	1433	1.89
books	1281	1.69
classroom	1257	1.66
technology	1247	1.64
math	1093	1.44
students	1057	1.39
ipad	927	1.22
read	777	1.02
science	747	0.98

In addition, sentiment analysis was performed to evaluate the emotion in text. Sentiment lexicons in R contain large collections of English words with negative or positive scores, and even emotions such as joy or sadness. These lexicons were created either through crowdsourcing or via an author whose assignment of scores was validated through crowdsourcing. With the text already in a tidy data structure, each word is compared to the lexicon and then assigned a negative or positive score. Fig 4 shows the top ten most negative and positive words in the project titles.

Overall, there does not appear to be a significant difference in diction between funded and expired projects.

III. Background

In 2016, the US Bureau of Labor Statistics reported that the median annual salary for public school teachers ranged from $55,800 to $58,030 with elementary teachers making the least and high school teachers earning the most (U.S. News). Considering how much of that is deducted through income taxes, teachers are barely able to sustain the average cost of living in the United States. That being said, public school teachers are still forced to pay out of pocket for school supplies for their students and basic classroom materials due to low budgets and the inability of low-income students'

Fig. 4: The top ten most frequent words in expired project titles.

parents to afford supplies. Recently, a new tax bill was passed which reduced some of the pressure of income taxes while also maintaining a $250 tax write-off for teachers to take advantage of. Unfortunately teachers spend, on average, $500 a year out of pocket so they are still unable to recover all of their expenditures.

To relieve some of the financial burden off of his fellow public school teachers, Charles Best created DonorsChoose.org, a crowdfunding website that connects teachers in need of funding to donors. The organization is totally self-sufficient as it maintains its operations by default-allocating 15% of donor contributions towards overhead. Because DonorsChoose is a nonprofit organization, there is no revenue or excess capital that can be allocated for marketing and research and development expenses. For this reason, nonprofits heavily depend on repeated donor contributions and word of mouth. Donor commitment has been proven to be positively influenced by share experiences, or receiving word of mouth (Sundermann). In fact, a blood donor retention study conducted in Oslo, Norway, found that the most important recruitment channel was the influence of active blood donors (Bosnes). Furthermore, the theory of planned behavior which explains all behaviors that people can exert self-control over, has been implemented to study the intentions and behavior of blood donors. One study in particular revealed that a donor's intention to donate is fueled by the desire to reap the benefits of the expected outcome of the donation (Hyde, et al.). These benefits include the strengthening of one's self-esteem and the maintenance of a positive image amongst your peers.

A. *Contextual Importance*

According to a survey conducted by Scholastic, public school teachers spend, on average, $530 of their own money on classroom supplies with those in low-income schools spending almost 40% more than that (Figueroa). While contributions through DonorsChoose.org alleviate some of the financial pressure that teachers face, the ability to pair donors with specific projects will enable the organization to fulfill more classroom requests. Building more targeted email campaigns will motivate donors to not only donate but contribute additional gifts.

B. *Approach*

Developing a strong marketing plan is a key aspect of building a successful nonprofit organization which relies heavily on donations and positive word-of-mouth. Relationship marketing, in particular, is extremely effective for donor retention as it focuses on building long-term relationships with key donors (Arnett, et al.). A survey conducted by Professor Adrian Sergeant, a leading authority on achieving philanthropic growth, showed that approximately one out of five donors stop donating due to changes in their financial circumstances while a similar proportion of donors lapsed simply due to them transferring their support to another organization (Sargeant, 2003).

In order to improve donor retention, DonorsChoose should implement a marketing campaign built upon factors that are known to effect donor commitment and loyalty. These factors include knowledge of a personal

link between the donor and the cause and the extent to which the donor shares the beliefs of the organization (Sargeant, et al.). While these factors cannot be influenced post hoc, knowing the motives for donors initial support should be taken into consideration when developing a marketing campaign aimed at encouraging past donors to donate more frequently. For example, this analysis studies the relationship between political affiliation and donations as surveys show that Democrats tend to donate slightly more than Republicans. This analysis also delves into the aspects of projects that best motivate donors to contribute. Segmenting donors into categories like average donation amount, number of contributions, and frequency of contributions is effective for building lasting relationships which will strengthen donor retention (Klein).

IV. Statistical Analysis

The following analysis takes the format of alphabetized subsections. We wish to start broad and narrow in toward individual donors and projects as the analysis continues. We begin by looking broadly at geographical and time-dependent characteristics, then examine topics regarding project characteristics, and lastly conclude with donor micro-characteristics and an introductory recommendation system.

A. Equality between States

1) Question: A naive approach to assessing the significance of certain states with regard to the amount their donating is to simply order states from highest donating to lowest donating. However, this neglects the fact that population easily confounds measurements of state characteristics, thus if we are truly interested in states that donate high amounts, we must normalize them. The same idea may be applied to any geographical characterization (i.e cities or zip codes). After this, we'd like to examine if states have a proportionally equal amount of donors? Is there a proportionally equal amount of donation money coming from donors within that state? Knowing the significance of certain states can influence the priority of states that are managed and marketed to.

2) Methods: Approaching these questions must involve some process of normalization. We utilize the U.S. Census Bureau's 2010 state demographic data to normalize the total number of donors per state and the total amount of donation money from all donors within a state. Our first hypothesis posits that each state has a proportionally equal amount of donors. The second hypothesis posits that each state has a proportionally equal amount of donation money coming from donors within a state.

To test our hypotheses, we conduct two χ^2 Goodness of Fit test for all ($N = 51$) states (District of Columbia included), assuming a uniform distribution for the null.

3) Results: Both tests result in p-values of 1.0, thus we cannot reject either hypothesis as significant. Interestingly though, a look at the standardized residuals (SR) of the second hypothesis tells us that Massachusetts and the District of Columbia have a large ratio of donation amount per individual, with SR values of 0.8 and 1.1, respectively. All other states have less than a 0.2 SR value.

B. Politics and Donations

1) Question: Can we prove that either Democrats donate more or Republicans donate more? Targeting email recipients based on political leaning could lead to a higher conversion rate. Fig 5 plots the probability of a randomly selected individual within a state being in one of two political parties given the average donation amount per person in that state. The plot is fit with a Least Squares linear regression model and indicates standard error in translucent regions.

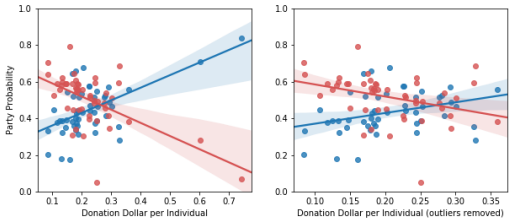

Fig. 5: The probability of a randomly selected individual within a state being in one of two political parties given the average donation amount per person in that state. Right: only donations per individual less than 0.5 to increase clarity.

Although there appears to be a clear linear relationship between how much individuals in a state will donate and the probability of choosing an individual from a certain party, this is not sufficient information to say whether Democrats and Republicans donate different amounts.

2) Methods: To answer the original question, Can we prove that one party donates more than the other?, we perform a nonparametric permutation test on the sets of total donation amount per individuals for Democratic states ($n = 19$) and Republican states ($n = 32$). Data is from the 2014 Federal Elections Committee results for the U.S. Senate and the U.S. House of Representatives.

3) Results: The resulting difference in means from the permutation test is about 0.09 and thus there does in fact exist a statistically significant difference between the two groups of ratios ($p = 0.002$). We can argue that a randomly selected individual from a Democratic state has donated about 10 cents more than a randomly selected individual from a Republican state. Although this seems negligible, the effect over millions of individuals is noteworthy.

C. Time Series Analysis and Forecast of Donation Amount

1) Question: What has been the trend across time of unique donors giving to projects, with Donorschoose.org as a vehicle? Is there a trendline? Seasonality? It has been noted in the donorschoose data set that some individuals donate extremely large amounts of money per donation, while some donate fairly little, however, with every unique donor giving to donorschoose.org, a project naturally receives a donation - and thus an increased likelihood of success. And since the donors are unique and come from a variety of backgrounds, they may donate to a variety of projects. In this section, we attempt to model the trend and seasonality of these unique donors as measured in the number of unique "hits" that donorschoose.org experienced per week, starting in 2013 up to present day. That is, the number of unique donors visiting (and giving money) through the site donorschoose.org, aggregated by week.

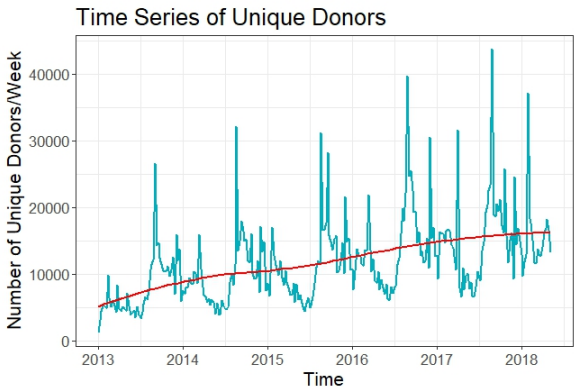

Fig. 6: Note the obvious seasonality of the data, with spikes of donor activity occurring approximately halfway between years, and just before the new year (seemingly around Christmas). The red trendline indicates a slow increase in the number of unique donors giving money.

2) Methods: Time Series Data may be seen as an empirical realization of a stochastic process that depends on a discrete valued integer of time. In our case, the dataset corresponding to donorschoose.org contained information on the total number of donations each project received from 2012 up until present day, approximately May of 2018, for every day of the week starting January 1st, 2013. In order to realize this information as a time series indexed by discrete, even spaced values of time, the number of donations received each day were summed up for the week (across all projects), for every week, starting from 01/01/2013 until 05/05/2018. Then, the donations were filtered by the unique Donor ID attached to each donation, so that the count of unique individual donors remained for the week. We hypothesize that the number of unique donors giving through donorschoose.org is imperative to the success of future projects. To this end, an Exponential Smoothing State Space Model (ETS) was fit to the time series. For comparison, an ARIMA model was fit as well. The two models are compared with a hold out set consisting of the first 19 weeks of 2018, against the test metric *mean absolute percentage error*, but we also include *root mean square error.* Both models were fit using the *forecast* package in R. Methodology comes from (Hyndman, 2018).

3) Results: Using the *forecast* package in R, we ended up fitting two models: an ARIMA model with drift, and an exponential smoothing state space model with seasonal and trend decomposition using locally weighted linear regression (STL+ETS, or just ETS). The two models are compared for accuracy, as can be seen in Table II, with the ETS model being selected for further analysis, due to the better test data fit.

TABLE II: The models appear comparable on the training data, but the difference is clear on the test set – ETS(A,Ad,N) is the superior model.

Model	Train RSME	Test RSME	Train MAPE	Test MAPE
ARIMA	3338.84	7163.15	13.84	32.80
ETS	2828.22	6229.50	15.75	24.66

The lower RSME and MAPE values for the ETS model indicate that the model is a better fit for the data, and that its forecasts will be more accurate. The test data set consisting of the first 19 weeks of 2018 was added back into the training set, and the model fit across the entire time series. Because ETS alone does not always produce accurate for data with long seasonality periods (which is present in our time series in Figure 6 – the seasonality seems to be a fixed period over a year), the time series data must be decomposed and de-seasonalized. Then an ETS model is fit, and the data is re-seasonalized using the most recent period. This composition was done using STL and may be seen in Figure 8.

The remainder components of the decomposition in-

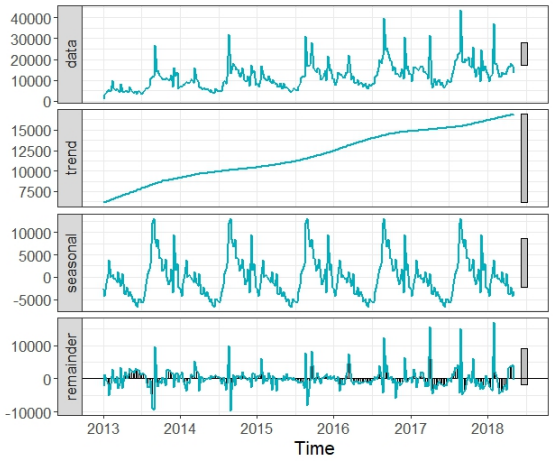

Fig. 7: Decomposition of the time series into it's trend, seasonal and remainder components. Decomposition allows the removal of the seasonal component, so that a model may be fit. Note the strong upward trend - indicating increases in the number of unique donors per week over time.

dicate the stochastic properties of the time series - these components are not well explained by either seasonality or trend, and may be interpreted as essentially random white noise added to the data. The remainder is approximately *homoskedastic*, but improvements in decomposition are likely possible. The slight change in variance near the end of the remainder decomposition is evident in the residual plot. Ideally, the residuals should appear to be perfect "white noise".

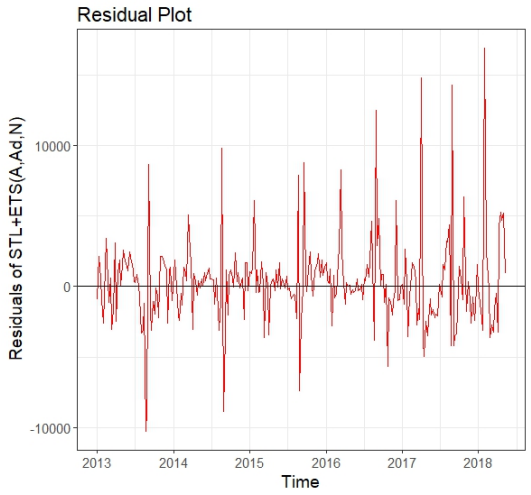

Fig. 8: The residuals aren't perfect, but are fairly close to appearing as "white noise". Model improvements are likely possible.

Upon fitting an ETS(A,Ad,N) (See methods and Theory for an interpretation of the (A,Ad,N) notation) model to the deseasonalized data, seasonality is then added back in. A forecast up to year 2020 is then produced and can be seen in figure 9, with the the quarterly point estimates and prediction intervals given in table III.

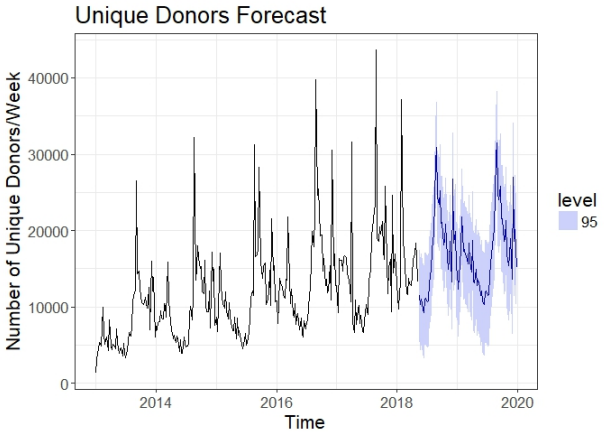

Fig. 9: An STL+ETS(A,Ad,N) model forecast of the unique donor count and total donation money given to donorschoose.org. Blue line indicates the forecast, and blue shading a 95% prediction interval.

TABLE III: The point estimates and prediction intervals for the number of unique donors to donorschoose.org from the STL+ ETS forecast model. The notation 2018.75 is interpreted as the point at which 75% of the year has passed– i.e., approximately 273 days into the year.

Year	Point Forecast	95% Prediction Interval
2018.500	10,634	(4,572, 16,695)
2018.75	21,904	(14,965, 27,243)
2019	14,548	(8,296, 20,800)
2019.5	11,504	(4,735, 18,273)
2019.75	21,592	(14,561, 28,622)
2020	15,173	(7,900, 22,447)

As can be seen in the forecast, large spikes of unique donors are expected to give to donorschoose.org near the 39th week of the year – September. It is likely possible that these spikes occur as teachers and parents are likely getting ready for the new school year - and thus become more cognizant of the availability of projects (donors) and the need to post them (teachers).

4) Conclusions: The time series data is fairly well fit by an STL decomposition and ETS(A,Ad,N) model, however the fit isn't perfect in that the residuals don't appear to have perfect constant variance. In real world data, it is difficult to achieve a perfect residual plot, but model improvements are likely possible using transformations of the data, such logarithmic or fourier transformations. Nevertheless, the model proved more accurate on a validation set of the time series than an ARIMA model.

Seasonality was detected in the series by inspection of the STL decomposition plots, with spikes of donor activity occurring around September, which is known to be the start of the academic year across the country. The activity then trends downwards with more surges in activity directly after the new year. The trend of the number of unique donors per week is upwards, indicating that the donorschoose.org is experiencing new growth.

We recommend choosing to fit email recommendations to potential new donors during the months of September and late January, for that is the period of time in which donorschoose.org experiences the most activity – likely as a function of parents becoming cognizant of the school year.

D. Project Feature Selection and Feature Optimality

1) Question: If the goal is to receive the most amount of money for a project, the logical origin of investigation would be to analyze projects that have done well in the past and look for similarities. What are the top qualities of a successfully funded project? Which factors have the most influence in predicting the monetary success of a project through DonorsChoose? In this section, we aim to determine the optimal project qualities for maximum project funding. Utilizing these results can help illuminate donor patterns and aid in formulating more successful project posts in the future.

2) Methods: Multiple regression is a tool used to quantify the relationship between multiple explanatory variables, or predictors, and the response variable. Using a weighted linear combination of the predictors, one can assess the extent of which each of the variables influences the outcome of the response variable. Logistic regression is a variant of linear regression in which the outcomes of the response variable are a dichotomy. In our analysis, this dichotomy is expressed by "Fully Funded" and "Expired". Using a multiple logistic regression model, we can analyze the weight of the project qualities, such as project cost and type, to determine the importance of each in the funding status of past projects.

The multiple logistic model was constructed using a combined data frame containing both project and school logistical information for 976,554 unique projects. Only past completed/ expired projects were considered in this analysis. Using a training data set of 488278 observations against a testing data set of 488276 observations, a generalized linear model was calculated with 10-fold cross validation for the logistic output to predict the status of a given project, using the convention that

$$\begin{cases} 1 & projectFullyFunded \\ 0 & projectExpired \end{cases}$$

The results of the logistic model were analyzed using the Z-statistic, as all the top predictors have $p < 0.0001$. For a two-sided hypothesis test, we compare the Z-statistic to a standard normal distribution. The *Null Hypothesis* states the weighted contribution of a predictor variable is due to chance, or negligible (α = 0.05). Variables whose Z-statistic lies outside the 95% confidence interval are considered significant in the analysis and have meaningful contribution to the outcome of a projects funding status.

TABLE IV: Top Five Project Status Predictors: The following predictors pertaining to the projects themselves are listed in decreasing significance.

Coefficient	Estimate	Std. Error	Z-Statistic
Project Cost	-0.0006393	0.000005505	-116.123
Teacher Project Sequence	0.004141	0.0001858	22.289
Warmth, Care & Hunger	1.323	0.06877	19.240
Project Funding Life Time	0.003391	0.0002373	14.286
Music & The Arts	0.2415	0.01747	13.820

3) Results: As anticipated, the cost variable was the most significant predictor in whether or not a project gets fully funded during its allotted funding time. Indeed, Fig 10 visualizes the distribution of costs for Fully Funded projects. An obvious trend shows that as the price increases, the number of projects that receive full funding decrease. Similarly, the funding life time of a project is a significant contributor to a projects status. Both results are intuitive, the more time a project is open to receive funding and the lower the project cost, the more likely it is to be fully funded before its funding expiration date. Additionally, the "Teacher Project Sequence" variable has a large contribution on the status outcome. In this analysis, "Teacher Project Sequence" refers to the number in the list of total project of a specific project that was donated to. The higher up in the list sequence a project is listed, the more likely it is to be seen by donors and therefore the more likely it is to receive donations. Table IV illustrates the significance of the aforementioned predictors.

In addition to understanding the most influential project qualities, the logistic model also analyzed variables that explain donor patterns as well. Among the top five predictors, the project subjects "Warmth, Care & Hunger" and "Music & The Arts" indicate that donors

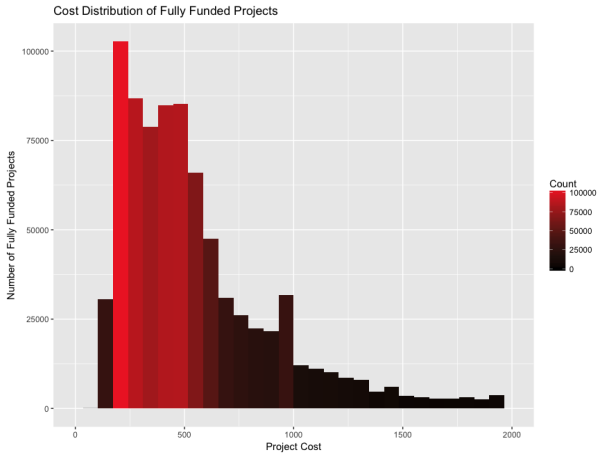

Fig. 10: The cost distribution for Fully Funded projects demonstrates the relationship between project cost and status: there are more Fully Funded projects that cost less than there are that cost more. The average Fully Funded project cost was 645 USD.

are more likely to donate to projects of these categories. This result is significant since projects of the "Warmth, Care & Hunger" were the least represented subject type. From this, one could hypothesize that donors respond more to projects with an emotional draw to them, rather than a technical one. Other significant predictors include, school metropolitan type, the time of year that the project is posted, project type, and grade levels of the recipient class(es). Analysis of the model indicates that donors are more likely to donate to Urban schools than other metropolitan types. Additionally, student-led projects for primary grades 9-12 are the most likely to be donated to. A limited review of time series indicates that projects posted in late November and in the month of December are the most significant for predicting a Fully Funded status. A more in-depth time series analysis is found above.

E. Donation Timing and Eliciting the Next Donation

1) Question: Timing is an essential component when targeting donors with emails in an attempt to elicit their next donation. It is in the organization's interest to send emails at a time that entices donors toward a donation without seeming pushy or overbearing. With this in mind, we ask: Given a donor's donation history, what is the best timeframe for sending an email with recommended projects?

2) Methods: We subset the data to only the last year of donations and group donations by their donor. For any prospective donor, they have β donations, where $\beta \geq 1$. We hope the prospective donor will make one more donation, so we filter the data subset to donors with only $\beta + 1$ donations. We assume the prospective donor has come back to donate with at least some amount of time t in hours between each donation. For example, an interesting subset of donors are those where $t = 168$, which describes donors who have a minimum of 7 days between their donations. This implies they were impelled to revisit the site to donate again.

From the aggregate subset of donors with $\beta + 1$ donations, we calculate the median amount of time (in days) between each donation, given a minimum t value. Based on this information, we can estimate the number of days given the number of hours between certain donation transitions.

3) Results: Depicted in Fig 11 are the median gaps between donations (in days) from the above methods with $\beta \in \{2, 3, 4, 5\}$ and $t \in \{1, 2, 6, 12, 24, 48, 72, 96, 120, 144, 168\}$. Perhaps one of the most prominent conclusions from the figure is that donors are quicker to donate after they have donated multiple times.

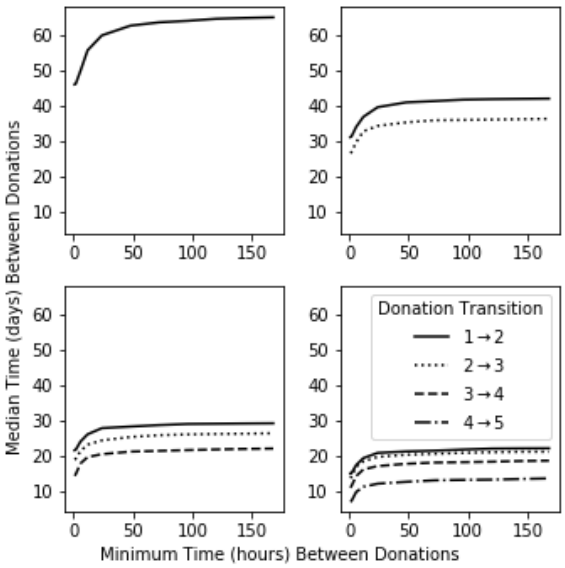

Fig. 11: The median duration (in days) between consecutive donations while controlling for the minimum amount of time t (in hours) between donations and the number of total donations β per donor.

Let us consider again the example from the Methods above. Suppose the organization is interested in recommending projects to a donor who has already made three donations, and the question now is what the most effective timing of the next promotional email is. Suppose this donor waited 168 hours, or 1 week,

between their second and third donation. Based on this information, we estimate that the most effective email should be sent about 22 days after the third donation if we wish to elicit a fourth donation.

F. Relation Between Donation Rate and Project Funding Status

1) Question: Inspired by Wash (2013), we wish to investigate the relation between a project's donation growth and the likelihood of it being funded. When choosing projects to market to potential donors, it might be advantageous to send projects which are at a certain point in their progression toward being funded. Wash (2013) showed that individuals are more likely to donate to projects they believe will be carried out and benefit the recipient. Thus it is important for the platform to recommend projects which are likely to get fully funded. What are progress indicators of a project deemed "suitable" (or "unsuitable") for recommendation?

2) Methods: We wish to estimate when a project is deemed "suitable" for email marketing based on its performance, where we "suitable" is defined as "likely to reach fully funded status." In this case, a project is likely to be completely funded if there is a 0.95 probability of it reaching fully funded status. From the dataset, we observe projects that have donations up to 120 days past the project inception date, are only expired or fully funded, and have complete information to describe the cost and donation amounts of that project. We fit a cubic polynomial on a 0.95±0.02 probability threshold using a least squares polynomial model, minimizing the squared error:

$$L = \sum_{i}^{n} [y_i - (a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3)]^2$$

We characterize "suitable" projects as those lying in the space above the polynomial fit.

3) Results: In Fig 12, we present a heatmap which visualizes the probability of a project getting funded as a function of days since the project inception and the portion of the project that has already been funded. The least squares polynomial model has been fit in white.

The implications of this model suggest that projects recommended to potential donors should be nascent and already gaining traction at a certain rate defined by the polynomial model. Donors will be more likely to donate to these projects due to their perceived (and often times, actual) success.

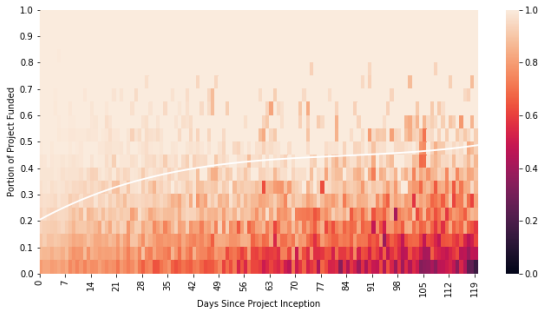

Fig. 12: Nascent projects with a high portion of the project donated to are highly likely to be fully funded. The 95% probability of becoming fully funded is delineated in white.

G. Finding a Predictive Model for Success of a Project

1) Question: According to the official Donorschooose.org help website (help.donorschoose.org), when a project fails to meet its goal for funding status (by hitting the expiration date before the monetary goal is met), a variety of potential events can unfold. An email is sent to the donor, who then may choose whether to forward the money to the teacher who set up the project anyway, for use in a future project, or to reallocate the money to a more urgent project in need. If no action is specified by the donor, then Donorschoose.org chooses the latter option automatically. Thus, in order to ensure that teachers actually receive the money they request, and in order to effectively market projects that are likely to succeed in order to cultivate donors, a predictive model for the success of a project may be of use. Our question is: Can we predict the eventual success of a donorschoose.org project, using only the publicly available information of the project itself?

2) Methods: Towards the end of predicting the eventual success of a project, the DonorsChoose "Projects" and "Schools" datasets were aggregated in order to attach the usual project information to the geographic location and financial status of the school. The data set was then engineered into a set of features in order to use as predictors for three possible models: a generalized logistic regression model, a gradient boosted tree model, and a logistic ridge regression. Our response variable was the "funded status" of the project: that is whether it was "Fully Funded" or "Expired", by encoding each status as 1, or 0, respectively (see Section C). Of the features chosen for the project, we chose the following (in no particular order): project cost, school metro type, percentage of student body on free lunch, school state, the subject of the project (say,

math or history, and so on), length of time the project was up for, the start date of the project, and the desired resource for the project.

Our choice of predictors for project success was decided on the relative uncertainty of donor behavior. In order to attempt to capture the best predictors of donor success, our model needs to encapsulate possible motives for donor behavior - including geographic location, affinity towards particular school subject, or desire to aid in providing a particular kind of resource. In section C., we discussed which features seemed most predictive of project success in a logistic regression setting. In this section, since Ridge Logistic Regression and Gradient Boosting have built in feature selection, we simply apply as many features as possible, and allow the model to select for the most relevant (in regards to the success of its own algorithm). All three models will be contrasted to each other, and the relative merits of each discussed.

In order to fit each model, the *caret, glmnet, gbm* packages in R were used in conjunction with each other. *Caret* in particular is a tuning algorithm, which performs k-fold cross validation in order to tune model parameters to maximize model outcome in regards to a specific metric - in our case, we opted for the *Reciever Operating Characteristic Curve* (ROC Curve), in order to find the best balance between predictive accuracy and specificity, and in order to avoid "trivial models". See Methods and Theory for a deeper discussion on the ROC Curve and it's use for model fitting.

The data, being large, was partitioned into a 60:40 split of training and hold out sets, where the training set consisted of $683,589$ data points, and the hold out set consisted of $292,965$ observations. The partition was created using *stratified sampling*, so that the approximate proportions of Expired to Fully Funded classes were about the same in the train set– a ratio of 0.290 expired over fully funded. A 5-fold cross validation was performed many times according to the *Caret* algorithm on the training set, and when the ROC Curve value was maximized, and the model fit, a final prediction was performed on the hold out set with a decision boundary cutoff of ≥ 50 being recorded as a prediction of "Fully Funded. Then, a confusion matrix of the results were recorded and the approximate accuracy of the model was taken.

3) Results: Using *caret* to optimize the tuning parameters for the three models in accordance with maximizing the Area Under the ROC curve (AUC), the following final models were chosen:

TABLE V: Final tuning parameters as selected by the *caret* algorithm – The algorithm randomly searches across various values for the tuning parameters and selects for the one that maximizes the AUC during 5-fold cross-validation.

Model	Tune 1	Tune 2	Tune 3	Tune 4
Logistic Regression	N\A	N\A	N\A	N\A
Gradient Boosting Machine	Iterations: 500	Shrinkage: 0.1	Tree Depth: 8	Minimum Observations per T.Node: 10
Ridge Regression	$\alpha = 0$ (ridge)	$\lambda = 0.0063$	N\A	N\A

It should be noted that 500 iterations was chosen in the final Gradient Boosting Machine model because we found that the model performance did not improve with more iterations, but fitting time increased dramatically. In fact, when initially fitting the model with $n = 1000$ iterations, the model took well over 24 hours to fit, with no model performance improvement observed after 500 iterations. Model performance for GBM improved with larger tree depth, but was computationally expensive as well. Shrinkage parameters had a great effect on the learning and performance of the GBM machine, as can be seen in Fig 13.

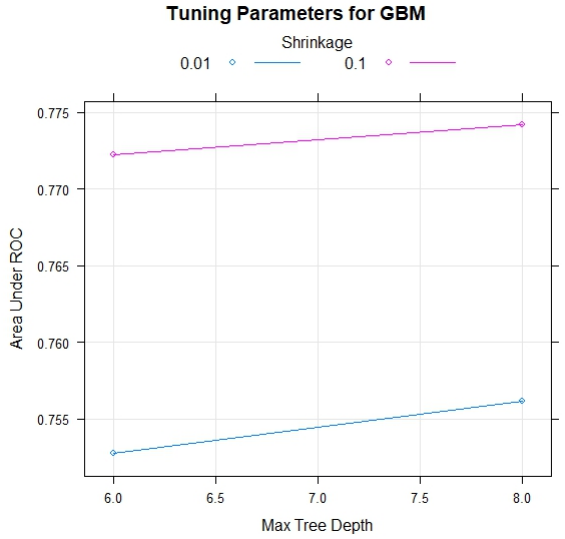

Fig. 13: A plot of the tuning parameters for the Gradient Boosting Machine. The y-axis is the AUC for the ROC curve, which is used to optimize the model. The x-axis is the maximum tree depth, and each line the model learning rate.

Shrinkage may be interpreted as how quickly the model learns – smaller shrinkage levels denote a slower learning rate. As can be seen Fig 13, a slower learning rate evidently benefits more quickly from tree complexity, but it unknown whether the two models would converge. Since the higher shrinkage rate does produce a model with a higher AUC, that model was

selected.

For penalized regression to be classified as a Ridge Regression, the tuning parameter α is fixed at 0 (see methods and theory). Thus, the only parameter to optimize is λ, essentially the penalty function to the coefficients on the model. The smallest possible λ is selected such that

$$arg\min_{\beta}\{\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{P}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{P}B_j^2\}. \quad (1)$$

holds. The final selection for λ was 0.0063. Holding α at 0 ensures that our model is purely a ridge regression, as opposed to an elastic net or a LASSO.

There were no parameters to select for in a generalized logistic regression.

TABLE VI: The performance of each final model at the 50/50 decision boundary for classification. Note that the accuracy between all models is comparable – however, the Ridge Regression and Logistic Regression have very low Specificity: the models are essentially trivial. This will be explored further.

Model	Area Under ROC Curve	Accuracy	Sensitivity	Specificity
Logistic Regression	0.689	0.776	0.986	0.048
Gradient Boosting Machine	0.774	0.790	0.963	0.207
Ridge Regression	0.706	0.778	0.988	0.06

The ROC curves for each model was drawn (Fig 15 and Fig 16) and the respective Area Under the Curve scores given in Table VI. The gradient boosting machine massively outperformed the logistic and ridge regression models, with an AUC of ≈ 0.77 (1 indicates a perfect model, 0.50 indicates model is equivalent to random guessing) after 5-fold cross validation. Ridge followed behind with an AUC of 0.705 which was only slightly better than the logistic regression, which had an AUC of 0.69. Moreover, at the 50/50 cut off between classes, the GBM had more balanced Sensitivity and Specificity, while the ridge and logistic regression both had extremely high Sensitivity scores, but ≤ 0.07 Specificity scores - indicating that the resulting models were essentially trivial - nearly all results were predicted to be in the positive class - "Fully Funded". This may be an after effect of the imbalanced classes in the data, for there was a nearly 4:1 ratio of Fully Funded to Expired classes in the training and testing set of data.

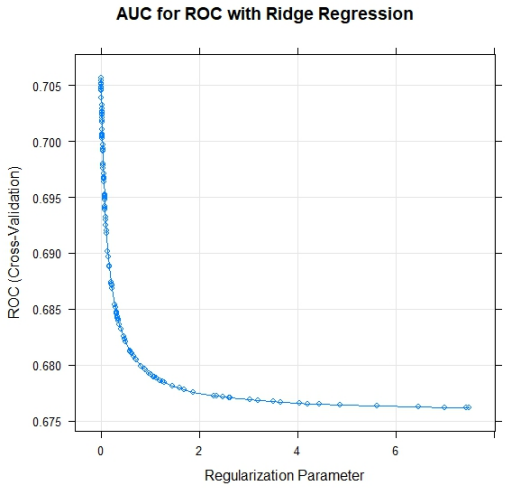

Fig. 14: λ is the penalty function attached to the parameters. As can be seen by the curve, small penalties actually fit the model best, with the AUC bottoming out around $\lambda = 1$.

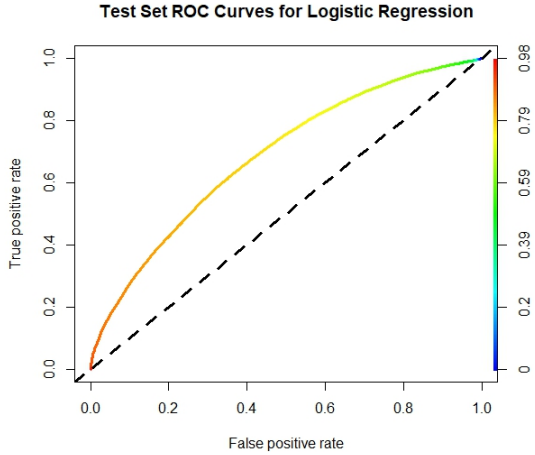

Fig. 15: ROC Curve for logistic regression. Noticeably more shallow than the curves for GBM and Ridge.

4) Conclusion: The GBM out performed the Ridge and Logistic regression with a traditional $\geq 50\%$ cut-off between class predictions. However, the model was extremely computationally expensive and took well over 24 hours to fit – GBM tuning parameters had to be heavily curtailed to prevent exponentially long fitting time. On extremely large data sets,the model may be prove problematic to utilize, despite the overall stronger performance. Class imbalance present in the data may be the primary driver of producing a trivial model for the Logit and Ridge regression models. In the next section, we give a discussion on choosing decision boundary cutoffs to alleviate the problem of class imbalance: using information from Receiver Operating Characteristic curve.

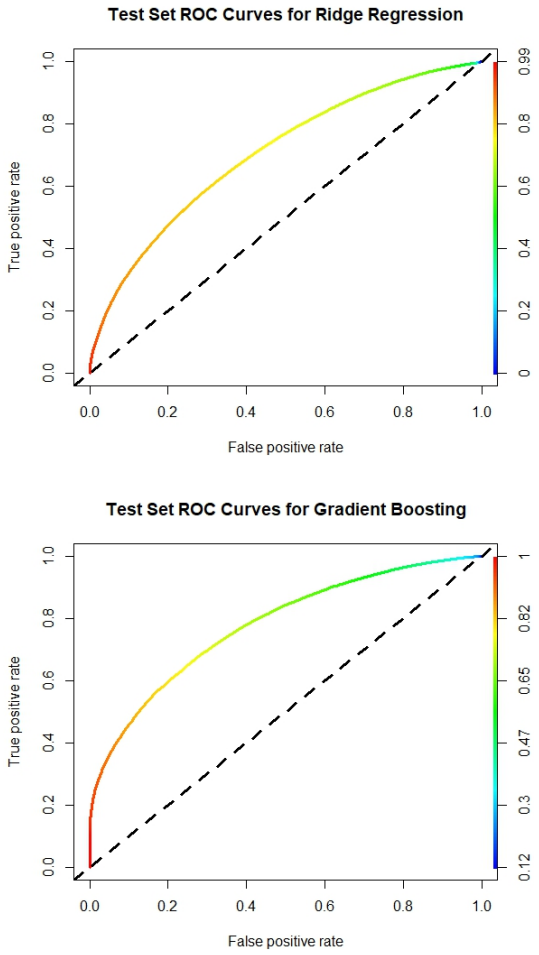

Fig. 16: The ROC curves for the Ridge regression and the Gradient Boosting Machine, respectively. While the GBM has an AUC score of 0.77 and the Ridge has a score of 0.705, the curves appear fairly similar.

We believe the use of a predictive model such as the ones described here may be useful in creation of a recommender system for donor cultivation - for then successful projects can be pushed in order to cultivate donors. By facilitating a positive experience with donorschoose.org, donors may be more likely to continue donating.

H. Prediction of Funded Status with Imbalanced Data: A remedy via Decision Boundary Optimization.

1) Question: As was shown in the preceding section, the models fit to the data have numerous costs and benefits. The Gradient Boosting Machine (GMB) proved to be the most effect model not only with respect to raw predictive power, but also sensitivity and specificity: As a result, it had the highest Area Under the ROC Curve (AUC) value from the cross-validation procedure, with an AUC of 0.774 - indicating a fairly good model on a data set in which predictive power of the features may be low. Moreover, it had the highest Accuracy, Sensitivity, and Specificity scores when performing predictions on the hold out set (TABLE III). However, the model was extremely time intensive to fit, with a run time of *nearly 24 hours* – while more powerful computers do alleviate this cost somewhat, it should be noted that the Logistic Ridge Regression was not significantly behind the GBM, with an AUC score of 0.706 from 5-fold cross validation on the hold-out set, while being significantly easier to implement, both in terms of computational requirements, and interpretation. However, the Sensitivity and Specificity of the Logistic Ridge Regression was dreadfully unbalanced: 0.988, and 0.06 respectively at the usual 50/50 decision boundary cutoff for prediction between Expired and Fully Funded classes. This begs the question: Can the issue of low Specificity be improved by the manipulation of decision boundary cutoffs?

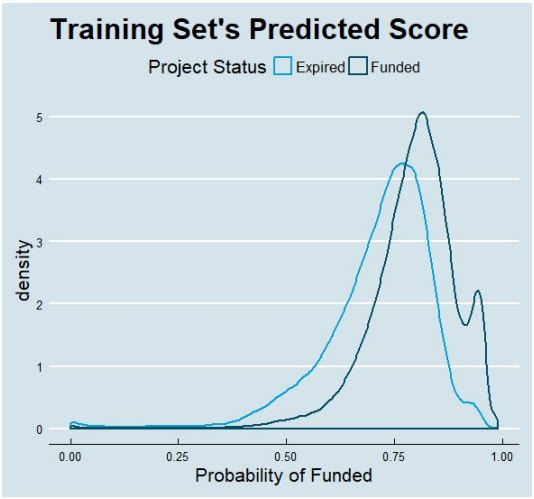

Fig. 17: A density curve of the predicted probabilities of the training set for logistic ridge regression. The lack of separation in the peaks indicates the model is attaching high probabilities to data points that have true labels of both ”Expired” and ”Fully Funded” – a result of class imbalance in the data set. Our task is to salvage the model by selecting an optimal decision boundary.

2) Methods: In order to improve a simple logistic regression, ridge regression was applied. However, ridge regression performs feature selection not by shrinking the coefficients of the features to zero, but instead just penalizing them. However, with such imbalanced data, our model became essentially trivial, with the majority of cases being predicted as ”Fully Funded”. To address the issues of Sensitivity and Specificity (and hopefully improve model accuracy), we attempt the address the

issue of class imbalanced data by exploring a variety of probability cut offs for predicting a particular project as ”Fully Funded”. To this end, we will attempt to optimize the decision boundary cut-off for for the prediction between the two classes of ”Expired” and ”Fully Funded”. Since ”Fully Funded” is the positive class, the cut offs are designed so that if a probability is above a given threshold, we will mark it as ”Fully Funded” by an indicator function. Below the cutoff, the class will be marked ”Expired”.

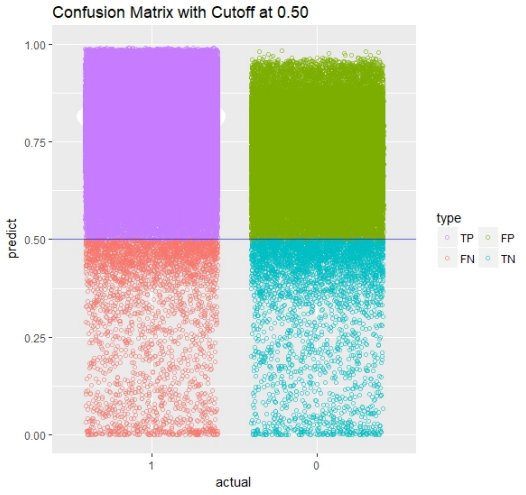

Fig. 18: This rather barbaric graph is a visual representation of the problem logistic ridge regression’s current decision boundary, at the 50/50 cut off. A perfect model would consist of only purple and blue colors. Note the massive amount of data points being predicted into the positive class! (denoted with 1).

3) Results: Using the fitted logistic model, multiple predictions were made using the hold-out set of projects data, with different boundary cutoffs. By using the ROC curve in the previous section, we selected the so-called ”optimal” decision boundary of 0.7786 probability, as well as decision boundaries of 0.05 increments. They are summarized in table VII.

TABLE VII: The probability cut offs and corresponding accuracy and sensitivity/specificity values. If a predicted probability was above the value in ”Decision Boundary”, it was entered as a member of the ”Fully Funded” class.

Decision Boundary	Accuracy	Sens.	Spec.
.65	.772	.926	.234
.70	.774	.854	.375
.75	.686	.723	.554
0.7786 (”Optimal” by ROC)	.627	.612	.680
0.80	.572	.518	.769
0.85	.424	.279	.922

As can be seen in the table, arbitrarily increasing the decision boundary results in a model that eventually becomes worse than random guessing – ie Accuracy below 0.50. However, a very modest increase in the decision boundary threshold gives a substantial boost to to model Sensitivity and Specificity, without a drastic loss in Accuracy. With a 0.65 decision boundary, the model is nearly as good as a boosted tree model with a typical 50/50 cut off! Further increases lower the accuracy, but give big gains in specificity. A visual representation of the 0.70 cutoff can be seen in Fig 19.

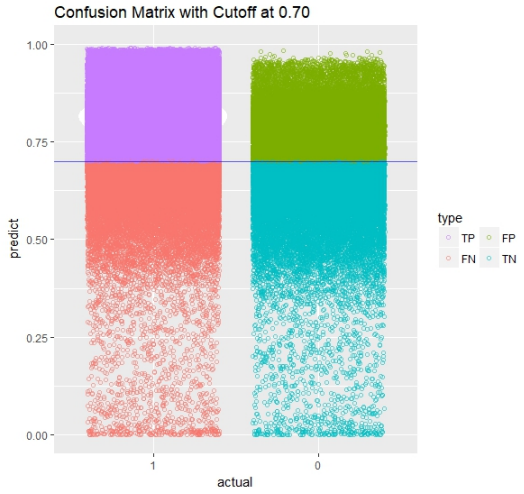

Fig. 19: A much more balanced confusion table. While sheer number of classes obfuscates the interpretation, the visual interpretation of the boundary cut off is usual.

4) Conclusions: Very moderate increases to the decision boundary results in significant model performance in terms of Sensitivity and Specificity, at the cost of a minimal loss in accuracy. While the AUC score of the logistic ridge regression is still lower of that than the GBM, the model is easier to fit and interpret.

Sensitivity is the proportion of predictions that were correct, and seeing as we are primarily interested in predicting project success, maximizing this value without too heavy a penalty to the Specificity may be more important. But it can clearly be seen that for imbalanced class data, optimizing a decision boundary can lead to improved model performance.

Such changes to the decision boundary may prove useful in the prediction of project status when the data is imbalanced.

I. Donation Characteristics of Teachers and Non-Teachers

1) Question: Donors in the dataset are characterized as teachers or non-teachers. We are also interested in the differences in donation habits between these two groups. Do teachers donate more than non-teachers, in terms of total amount, average amount, and frequency?

Significant differences may motivate the organization to compose emails to each group that represent their behavior.

2) Methods: We approach the above question by first examining the number of donations in two empirical distributions: the donation count for teachers and the donation count for non-teachers. We create two distributions for the average amount donated by each group, and lastly we create two distributions for the total amount donated for teachers and non-teachers. For each distribution set, our null hypotheses posits that there does not exist a difference between the teacher distribution ($n = 495,257$) and the non-teacher distribution ($n = 91,572$). We apply a Welch's t-Test to each set of distributions, with $p < 0.01$.

3) Results: In all three tests, p is nearly zero, thus we cannot reject the null for any of the three tests. Therefore, we cannot say that teachers' donation habits are significantly different than non-teachers'. However, reporting the actual differences may also be of interest to the organization.

During their lifetime as a donor, the median of total donation amounts for teachers is \$50 more than non-teachers. However, the median average donation for teachers is about \$2.60 less than non-teachers. From this, we can infer than teachers donate less per donation but more frequently than non-teachers.

J. Similarity Between Projects: A Recommendation System

1) Question: Large entertainment businesses such as Spotify and Netflix have recently begun to rely more heavily on recommendation systems to retain their users. Similarly, we will proceed to suggest two approaches to recommendation systems which DonorsChoose may implement to increase personalization in their marketing emails. We are faced with the challenge of suggestion projects to donors given a very limited foundation of content to work off: odds are, a donor has only ever donated to a project once. The question becomes, Which projects should the platform recommend to a donor to elicit the largest donation possible? We make the assumption that a donor is more likely to donate to a project which is similar to their donation history.

2) Methods: We approach this question from two methodologies: Cosine Similarity and Jaccard Similarity (See Methods and Theory). Jaccard Similarity is chosen because it is a *negative match exclusive* measure which operates well over asymmetric binary vectors. With each of these metrics, we compute the metric distance between n projects with m features, creating an $n \times n$ matrix of distances between the i^{th} and j^{th} elements.

The features of the $n \times m$ input matrix of the Cosine Similarity algorithm are project grade, cost, and start date as well as school metro type, percent lunch, and teacher session. These features must first be encoded as numeric values. In this particular case of feature engineering, a numeric value is assigned to qualitative data in the most sensible way possible. The most important property of these numeric values are their uniform spacing and order. The choice of ordered numeric values for school metro type are justified by following the conventions of the National Center for Education Studies (NCES, 2016). The choices for project grade are justified by maintaining order of the grades. Project start date was converted from a date to a unique integer that preserves order. The resulting numeric data may be slightly altered even though careful effort was made to create the most sensible encodings. After encoding, the features are re-centered, and normalized to a range in [-1,1] by using the *preProcess* method from the *caret* training library. It is because of this normalizing that the order was most important during encoding and not the particular values.

Each project's features can be though of as a vector representing a point in $\mathbb{R}^6$ with each basis corresponding to a single feature contributing equal weight. Vector pairs are passed into the cosine similarity formula which computes their inner angle. For prediction a project is chosen, such as one already supported by a target donor, and its feature is compared pairwise to every other project in order to find the smallest inner angle. We evaluate projects for content-based recommendations rather than the collaborative filtering method, given our data had a limited proportion of repeat donors.

The features of the $n \times m$ input matrix of the Jaccard Similarity algorithm are all binary features regarding the project funding status, teacher gender, project subcategory, school metropolitan type, project class grade, project cost, percentage of school on free lunch, and rough school geographical location (north, west, midwest, northwest, northeast, southeast). Nominal variables (such as subcategory), were vectorized and then made binary. After sorting by most similar for each project, we extract the most similar project for each project.

3) Results: A simple cosine similarity is demonstrated to illustrate the possibilities of prediction. Below 200 projects are pairwise compared for their cosine

angle similarity across only 6 features. Purple represents dissimilar projects while blue represents very similar projects. Notice how there are clear pockets of similar projects.

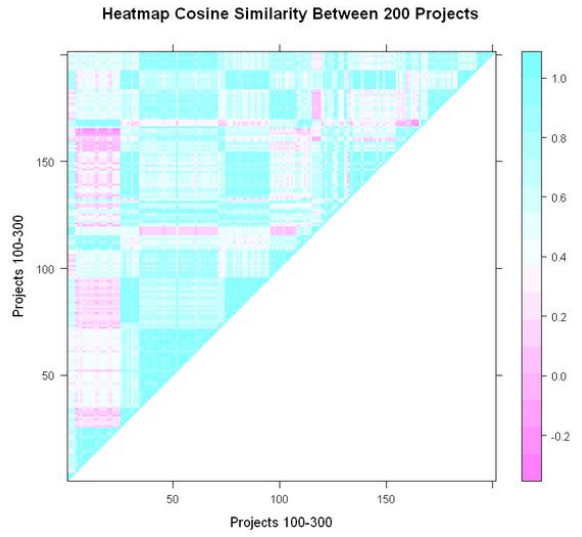

Fig. 20: A Cosine Similarity metric of 200 projects from #100 to #300.

This heatmap serves as a useful tool for visually predicting similar projects among a small set. However, a visual model is unrealistic for discerning which projects are the *most* similar. Therefore two functions and a wrapper were written. The wrapper called *cosingle* takes an input project, a range of projects to search, and returns the top three most similar projects. The table below demonstrates the top three most similar projects for project #200.

```
[1] "top 3 projects in row 100-300"
```

	project.id	project.grade	project.cost	school.metro.type	school.percent.lunch	teacher.proj.seq	project.start.date
	8e1543957ad6755c9853751ec7e2a2e3	2	427.2	2	95	4	15955

	project.id	project.grade	project.cost	school.metro.type	school.percent.lunch	teacher.proj.seq	project.start.date
	21f11abb0106d849dfbef0cc78ef5299	2	586.75	2	95	1	15983

	project.id	project.grade	project.cost	school.metro.type	school.percent.lunch	teacher.proj.seq	project.start.date
	5989ad2acd8d5dba1465341bfac791ba	2	399.4	2	95	1	16025

```
[1] "that are similar to target project in row 200"
```

	project.id	project.grade	project.cost	school.metro.type	school.percent.lunch	teacher.proj.seq	project.start.date
	b289a99d90052f71d543ff80d6b6a737	2	690.56	2	95	2	15920

Fig. 21: Project #200's top 3 Similar Projects within Projects #100 to #300.

Notice in this particular example how of the 6 features in the analysis, 3 of them were matched perfectly by the algorithm (project grade, school metropolitan type, and percentage of school on free lunch). The cost variable had the only variation, the furthest value being $197.62 away from the value in question, but given the project cost standard deviation of 1082.457, these values are marginally close to the original value. For a more accurate system, one would implement more features on a larger subset but the above example provides a generic example.

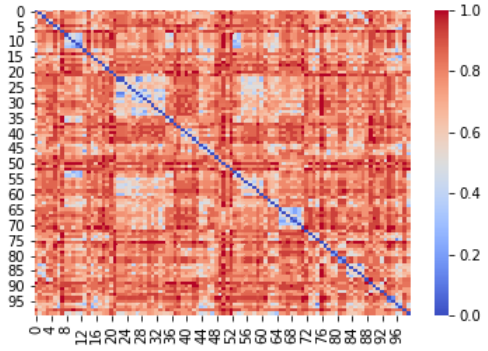

Fig. 22: The Jaccard Similarity metrics of the first 100 projects. The diagonal elements will always have a similarity of 0.0 since the j^{th} project is exactly equivalent to the j^{th} project.

The mean Jaccard Similarity over 31,000 randomly selected "similar" projects is 0.152 ($SD = 0.099$), where 0.0 indicates completely similar, and 1.0 indicates completely dissimilar. This is strong evidence that translating nominal and continuous data into binary vectors may also lead to a reliable recommendation system. In Fig 22, we see the similarities of the first 100 projects. The diagonal will always have a Jaccard Similarity of 0.0, since the j^{th} project is exactly equivalent to the j^{th} project. The overwhelming amount of red is a positive indicator that the algorithm has found similar projects beyond the first 100, however it not abnormal to see some blue (similarity) in the heatmap.

Should DonorsChoose wish to implement a recommender system into their email marketing platform, both Cosine and Jaccard Similarity could provide reliable results capable of providing similar projects to a donor's history. With better, more personalized recommendations, the organization is more likely to elicit the next donation from a donor.

V. DISCUSSION

Marketing research is extremely susceptible to the effects of confounding variables. While this analysis does control for many variables, such as whether or not the donor is a teacher, how many projects a donor has contributed to, and the socioeconomic status of the participating school, there exist many uncontrollable environmental variables that play an important role in quality and quantity of donations. For one, the state of the economy can have a great impact on donor behavior. The Stanford Center on Poverty and

Inequality conducted a study which reinforced the assumption that charitable giving declines during times of economic turmoil using the Great Recession of 2008 as a reference (Reich, Wimer). In fact, in 2007, the year prior to the recession, total charitable giving in the United States reached an all-time high at approximately $320 billion. Yet after the Great Recession, total giving dropped by 7% in 2008 and by another 6.2% the following year, marking the first decline in giving since 1987. Furthermore, donations towards education decreased by 5.5% after the recession. Because the economy is so unpredictable, there is no way to control for its effects on donor behavior.

Another possible confounding variable is the socioeconomic standing of the donors as this effects the amount and frequency of donations. While the city, state, and first three digits of the zip code of the donor are known variables, it is difficult to accurately predict their income based on this information and impossible to predict variables such as number of children and religion which have been proven to have significant effects on the frequency of donations (Yao). Furthermore, it is unknown whether or not the donors have children who directly benefit from fully funded project requests. It is possible that prior donors with children, especially those who have only made one donation through DonorsChoose.org, were solely motivated to contribute to a project request because it was for their childs classroom.

In order to improve this analysis, it would be beneficial to control for these confounding variables. One way to do this is to determine the average income and economic diversity of every zip code listed in the donor dataset. By doing so, the donors income levels can be predicted based on the city, state, or zip code that they provided when making a contribution to a project. A study conducted in 2012 by the Chronicle of Philanthropy found that people who earned under $100,000 were more likely to donate than those who earned more. The same study also revealed that high-income people who live in economically diverse areas give more on average than those who live in wealthier neighborhoods. For example, in the 11206 zip code (Brooklyn, N.Y.), residents who earn more than $200,000 only makeup 0.3% of the population, yet they account for 38.3% of all donations that stem from the zip code. On the other hand, 55.5% of the population living in the 55144 zip code (Minneapolis, Minn.) is considered wealthy but only contribute 1.5% of total donations. With this information, the team at DonorsChoose.org will be able to target email campaigns at prior donors who live in more economically diverse zip codes as they are more likely to donate additionally due to increased exposure to struggling schools in their area.

VI. CONCLUSIONS

This analysis reveals that each state had a proportionally equal amount of donors and donation money coming from donors living within the state. Also, donors living in the District of Columbia were found to donate more per individual than those in any other state. This result is constant with previous surveys which attribute D.C. residents' generosity to their wealth and the strong presence of nonprofit organizations which are based in the area (SmartAsset). Moreover, contrary to past surveys which point to Republicans as the more charitable party, donors living in states considered to be Democratic were found to contribute approximately $0.10 more to DonorsChoose project requests than those living in Republican states. Given these conclusions, the team at DonorsChoose.org should target their email campaign at prior donors who live in Democratic states and in areas with high concentrations of nonprofit organizations.

The statistical analysis also revealed that the most significant predictors of a fully funded project were cost (as cost increased, number of fully funded projects decreased), funding lifetime (projects with longer lifetimes were more likely to be fully funded), and project list sequence (the higher up on the list, the more likely the project was to receive donations). Donors were also found to be more likely to donate to the project subjects "Warmth, Care & Hunger" and "Music & The Arts". In regards to other significant predictors of funding, the analysis indicates that donors are more likely to donate to schools in urban areas and to projects that benefit public high school students. Therefore, in order to increase the effectiveness of the email campaigns, DonorsChoose should recommend project requests that cost less and benefit high schools in urban neighborhoods. Recognizing donor motivation, like project subjects that speak to the donor the most, is extremely important as it allows DonorsChoose to recommend requests to prior donors who are more willing to make an additional contribution when connected to projects in subject areas that interest them.

It may be recommended to fit and optimize a model for the prediction of the eventual success of a project, using the work done here as a jumping off point. For when constructing an email recommender system in an attempt to cultivate donors, it may be imperative that only the most successful projects are recommended initially, in order to instantiate a positive affect in the donor towards the idea of donation to projects.

Logistic Ridge Regression performs fairly well, and is fairly simple to implement. However, if computing power is available, we recommend a Gradient Boosting Machine, which outclassed all other models that were fit.

Time series analysis on donor activity indicates that spikes of activity occur around September, during the beginning of the academic year, as well as (starting in 2016) just after the new year, likely when potential donors are feeling generous. We recommend that cultivation of donors by email correspond to these periods of increased activity, when more potential donors have their minds turned towards the academics of children.

Moreover, the analysis concluded that nascent projects with a high portion of funding complete at a certain point in the project's lifetime are more likely to receive donations due to the donors' perceived success of the project. Thus, to maximize the frequency of donations from prior donors, the email campaigns should recommend projects that have already gained traction at the rate specified by the polynomial model introduced in section E. These projects should also complement the specific aspects of the donors which motivate them to contribute to project requests.

Lastly, DonorsChoose may wish to implement a recommendation system to personalize emails even further. Donations are likely to be elicited by recommending similar projects found using Cosine or Jaccard Similarity.

VII. METHODS AND THEORY

In this section, the various procedures and theoretical tools used in this analysis will be stated and summarized. While there is insufficient space to provide fully rigorous proofs, some proofs might be sketched, and the reader will be directed to the appropriate source material.

In this section, Y_i will always denote a *response variable* while X_i will always denote an *explanatory variable* or a random variable related to a data point. In this paper, "explanatory variable", "predictor" and "features" will be used interchangeably, but they all mean the same thing.

OLS Linear Regression.. Consider n data points, $(x_1, y_1), ..., (x_n, y_n)$, where x_i is the *explanatory* variable and y_i is the *response* variable. The *linear method of least squares* is the straight line

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2}$$

that minimizes the function

$$L = \sum_i^n [y_i - (\beta_0 + \beta_1 x_i)]^2. \tag{3}$$

Here, ϵ refers to the error within the model. Now, $y = \beta_0 + \beta_1 x$ has slope

$$\beta_1 = \frac{n\sum_i^n x_i y_i - \big(\sum_i^n x_i\big)\big(\sum_i^n y_i\big)}{n\big(\sum_i^n x_i^2\big) - \big(\sum_i^n x_i\big)^2} \tag{4}$$

and y-intercept

$$\beta_0 = \frac{\sum_i^n y_i - \beta_1 \sum_i^n x_i}{n} \tag{5}$$

$$= \hat{y} - \beta_1 \hat{x} + \hat{\epsilon}. \tag{6}$$

Here L is a function that gives the squared residuals. The line $y = \beta_0 + \beta_1 x + \epsilon$ that satisfies this theorem is called the "best fitting" model. †

The analysis lead to the rejection of the linear regression in favor of exponential regression. We have that if $y = \beta_0 e^{\beta_1 x}$, then it follows that $ln(y) = ln(\beta_0) + \beta_1 x + \epsilon$. The fact that $ln(y)$ and x can be applied to the above Theorem to yield

$$\beta_1 = \frac{n\sum_i^n x_i ln(y_i) - \big(\sum_i^n x_i\big)\big(\sum_i^n y_i\big)}{n\big(\sum_i^n x_i^2\big) - \big(\sum_i^n x_i\big)^2} \tag{7}$$

$$ln(\beta_0) = \frac{\sum_i^n ln y_i - \beta_1 \sum_i^n x_i}{n} \tag{8}$$

for the slope and intercept, respectively. (Larsen & Marx).

Logistic Regression. Suppose that the responses $Y_1, ..., Y_n$ are independent and bernoulli distributed with some unknown probability p_i. Then, $EY_i = p_i = P(Y_i = 1)$. Suppose that p_i is related to a fixed data point x_i by

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i. \tag{9}$$

The left hand side is the log of the odds of success for Y_i, note that it takes exactly the form of a linear model as seen in regular Ordinary Least Squares (OLS) regression. Say that the *link function* is the function $g(p) = \log(p/(1-p))$, then inverting the link function by taking the exponential function on both sides gives

$$p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}. \tag{10}$$

Here $0 < p_i < 1$. In this manner, OLS regression may be specialized to the prediction (and inference) of bernoulli trials, or "classification" problems by the use of a link function such as g. The link function given here in (8) is the *Logit* link function - which was used in the logisitic modeling procedure in this paper (Casella& Berger).

Ridge Regression. *Ridge Regression* is a linear regression model that opts to modulate the effects of each explanatory variable by means of *shrinkage.* Instead of removing explanatory variables entirely, their influence on the model will be penalized. The penalty is chosen in such a way as to minimize a penalized residual sum. This is given as

$$arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{P} B_j^2\} \quad (11)$$

where λ is a complexity parameter that controls shrinkage, P is the number of features and N is the sample size.. The residual sum (10) is canonically denoted $\hat{\beta}_{ridge}$. The larger the value of λ, the higher the level of shrinkage. In matrix form, (10) may be expressed as

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta \quad (12)$$

so that the ridge solutions are given as

$$\hat{\beta}_{ridge} = (\mathbf{X}^{\mathbf{X}} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad (13)$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. More detailed exposition may be found in *Elements of Statistical Learning*, from which this primer was taken.

Sensitivity and Specificity. Let TN and TP denote the number of *true* negative and positive predictions. Let FN and FP denote the number of false negative and false positive predictions, respectively. Then,

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (14)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (15)$$

Exponential Smoothing State Space Model. We provide an explanation of the STL + ETS(A,Ad,N) model, and discuss the notation (A,Ad,N). The 3-tuple ETS(A,Ad,N) can be interpreted as an exponential smoothing model, with Additive Errors, Additive Dampening, No Seasonality (A,Ad,N) respectively. (The seasonality is obtained by the STL decomposition.)

An exponential smoothing model is a forecasting model for time series in which past observations are weighted, with older observation weights decaying exponentially as observations get older. Thus, more weight is placed towards more recent observations – this gives the name *exponential smoothing.* The smoothing process may be described more formally as follows:

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + ... \quad (16)$$

where $\alpha \in (0, 1)$ is the smoothing parameter. So the forecast for time $T + 1$ is a weighted sum of the previous points $y_{T-k}, k \in \{0, ..., T\}$ controlled by the parameter α. For larger values of α, the model gives more weighting towards more recent observations.

Our model uses an *additive damped trend* (Ad), so that the trend-line in the model is flattened over time (implying the model reaches some state of equilibrium. This may be represented as the system of equations

$$\hat{y}_{t+h|t} = l_t + (\phi + \phi^2 + ... + \phi^h)b_t \quad (17)$$

$$l_t = \alpha y_t + (1-\alpha)(l_{t-1} + \phi b_{t-1}) \quad (18)$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)\phi b_{t-1} \quad (19)$$

where $\phi \in (0, 1)$ and $\beta^* \in (0, 1)$ are smoothing parameters for the trend line. Here $\phi_h = b_t$ is an estimate of the slope at time t, and l_t is an estimate of the level of the series at time t.

The following are the initial values of the model:

$$l_0 = y_0 \quad (20)$$

$$b_0 = y_2 - y_1 \quad (21)$$

Since the errors are additive, the preceding system of equations can be written as

$$y_t = l_t + \phi b_{t-1} + \epsilon_t \quad (22)$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha\epsilon_t \quad (23)$$

$$b_t = \phi b_{t-1} + \beta\epsilon_t \quad (24)$$

where $\epsilon \sim iidN(0, 1)$. (Hyndman, 2018).

STL Decomposition. STL stands for "Seasonal, and Trend decomposition using Loess". All information present taken from (Hyndman, 2018). Under the assumption of an additive time series model, we can write

$$y_t = S_t + T_t + E_t \quad (25)$$

where y_t is the is the data point at time t, and S_t, T_t, E_t are the seasonal, trend and remainder components at a time t. Essentially, a time series may be decomposed into several elements that make it up. For example:

figure 23 is the decomposition of our time series data. The first picture is the raw data, and it has been decomposed into its trend line, seasonality and the remaining elements (which are not explained by trend or seasonality). Seasonality simply refers to a known and fixed period present in the data. In essence, time series analysis involves the decomposition of the data into deterministic and non-deterministic parts, where seasonality is taken as a fixed period, and thus deterministic. STL + ETS time series modeling works by first decomposing the time series and controlling the seasonality - then an ETS model is fit and seasonality is added back into the data.

It should be noted that this is not a decomposition in the typical sense. Rather it is a smoothing method in which Locally Weighted Regression (LOESS) is utilized.

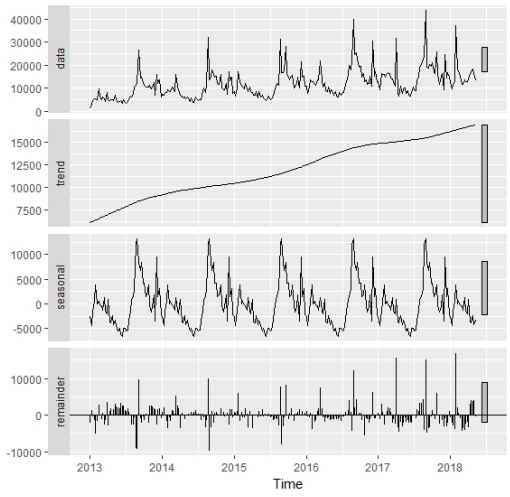

Fig. 23: STL decomposition of our time series.

For a more complete treatment on this process, see (Cleveland, et al, 1990). We give a description of LOESS below.

Locally Weighted Linear Least Squares (LOESS). We merely provide a summary of the framework of LOESS. Please see (Cleveland & Devlin, 1979) for a wonderful treatise on the subject.

Consider a set of response variables $(Y_1, ..., Y_n)$ and a set of p predictors $x_i = (X_{i_1}, ..., X_{i_p})$, for $i = 1, ..., n$. Define a general as

$$Y_i = g(x_i) + \epsilon_i \tag{26}$$

where ϵ_i are normally distributed errors with mean 0 and variance σ^2. Suppose $\hat{g}$ is our estimate for g. Then we have the relationship

$$\hat{y}_i = \hat{g}(x) \tag{27}$$

at the fitted value x. Define a δ-neighborhood around x in the predictor space. The estimator $\hat{g}$ then applies this neighborhood at x and forms the response $\hat{y}_i$ by weighting the x_i near x in the neighborhood than the x_i away from x in the neighborhood. Hence "locally weighted", as values closer to x are given heavier weights. The weights are, usually, defined as

$$W(u) = (1 - u^3)^3, \quad u \in [0, 1) \tag{28}$$

and 0 otherwise. Suppose there are q points in the neighborhood around x. Now, let $d(x)$ be the distance between the qth nearest x_i to x. Then, the weight of the observation (y_i, x_i) is

$$w_i(x) = W(\rho(x, x_i)/d(x)) \tag{29}$$

where rho is the euclidean distance between x and x_i.

Thus $w(x)$ as a function of i is a maximum for x_i close to x, decreases as the x_i increase in distance from x, and becomes 0 for the qth-nearest x_i to x.

Classification Trees. A classification tree is a regression tree that is used to predict a qualitative response by predicting that each observation belongs to the most commonly occuring class of training observations in the region to which it belongs. This is accomplished by means of a binary splitting algorithm. The classification error rate is defined as the fraction of training observations in that region that do not belong to the most common class:

$$E = 1 - \max_k(\hat{p}_{mk}) \tag{30}$$

where $\hat{p}_{mk}$ is the proportion of training observations in the m^{th} region that are from the k^{th} class. Note that $0 \leq \hat{p}_{mk} \leq 1$.

The *Gini index* defined by

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{31}$$

is the measure of total variance across the K classes. The Gini index is referred to as a measure of node *purity*, a small value indicates that a node contains predominantly observations from a single class.

More detailed exposition may be found in *Elements of Statistical Learning*, from which this primer was taken.

Boosted Classification Trees. Boosting is a method for improving the predictions resulting from a decision tree in which trees are grown sequentially with each tree utilizing information from previously grown trees. Cross validation is used to select the optimal number of trees, B, while the skrinkage parameter λ controls the rate at which boosting learns. The *interaction depth*, d controls the complexity of the boosted ensemble. Generally, d controlls the interaction order of the boosted model. We present the Gradient Boosting Algorithm. Let $(y_i, x_i)_1^n$ be a set of data points, f an unknown model such that $y_i = f(x_i)$, and $L(y_i, f(x))$ a differentiable loss function. Then the algorithm for the gradient boosting machine may be described by the following algorithm, taken from the *Elements of Statistical Learning*, by Hastie, Tibshirani, and Friedman.

As can be seen from the algorithm, the model estimate $\hat{f}$ is computed by iteratively minimizing the function

$$\gamma_{jm} = arg\min_{\gamma} = \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma). \tag{32}$$

In the case of our Gradient Boosting Machine, the model estimates $f_{m-i}, i = 0, ..., m$ were classification trees.

Cosine Similarity. Cosine similarity is a method for determining similarity between two objects using the

Algorithm 10.3 *Gradient Tree Boosting Algorithm.*

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to M:
   (a) For $i = 1, 2, \ldots, N$ compute
   $$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
   (b) Fit a regression tree to the targets r_{im} giving terminal regions R_{jm}, $j = 1, 2, \ldots, J_m$.
   (c) For $j = 1, 2, \ldots, J_m$ compute
   $$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L\left(y_i, f_{m-1}(x_i) + \gamma\right).$$
   (d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
3. Output $\hat{f}(x) = f_M(x)$.

Fig. 24: Gradient Boosting Machine Algorithm.

angle of linear distance over a Euclidean space. Qualatiative values for an object are enumerated into a non-zero vector and then run through the following formula:

Let A and B be non-zero vectors with n elements.

$$Similarity = cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}} \quad (33)$$

or equivalently

$$"angular similarity" = 1 - \frac{cos^{-1}("cosine similarity")}{\pi} \quad (34)$$

The resulting similarity values range from [-1,1] where 1 indicates perfect similarity, 0 indicates orthogonailty, and -1 indicates perfect dissimilarity.

Jaccard Similarity. The Jaccard Similarity is a metric for measuring the similarity of sets, defined as the size of the intersection divided by the size of the union between the sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The resulting distance is a characterized as $0 \leq J(A, B) \leq 1$. As elaborated on in Choi *et al.*, the Jaccard Similarity distance is a *negative match exclusive* measure. These account for the combined co-attributes to the total attributes of two vectors. Performance over aymmetric binary vectors is subsequently unsaturated and accurately reflects true similarity, opposed to other popular binary metrics such as Russell-Rao or the simple matching coefficient.

VIII. ACKNOWLEDGMENTS

REFERENCES

[1] Bosnes, V., Gasdal, O., Heier H.E., and Misje A.H. (2005). "Motivation, recruitment and retention of voluntary non-remunerated blood donors: a survey-based questionnaire study. *Vox Sanguinis, Volume 89, Issue 4*.

[2] Choi, S.S., Cha, S.H., and Tappert, C.C (2010). "A Survey of Binary Similarity and Distance Measures." *Systemics, Cybernetics and Informatics*.

[3] Figueroa, Ariana (2017), "How much do teachers spend on classroom supplies?" *National Public Radio*.

[4] "The Geography of Charitable Giving" (2012). *The Chronicle of Philanthropy*.

[5] How much do elementary school teachers make? (2016). *U.S. News: Money*.

[6] How much do high school teachers make? (2016).*U.S. News: Money*.

[7] How much do middle school teachers make? (2016). *U.S. News: Money*.

[8] Hyde, M.K., Masser, B.M., Robinson, N.G., Terry, D.J., and White, K.M. (2009), "Predicting blood donation intentions and behavior among Australian blood donors: testing an extended theory of planned behavior model." *The University of Queensland, Australia*.

[9] Hyndman, R.J. and Athanasopoulos, G. (2013). "Forecasting: principles and practice". *OTexts: Melbourne, Australia*. http://otexts.org/fpp/. Accessed on ¡05/20/2018¿

[10] Klein, Kim (2016), Strengthening Relationships by Creating Categories of Donors. *Fundraising for Social Change, 7th Edition*.

[11] Miller, Derek (2017). "Americas Most Charitable States 2017 Edition". *Smart Asset*.

[12] Reich, Rob and Wimer, Christopher (2012). "Charitable Giving and the Great Recession." *The Russell Sage Foundation and The Stanford Center on Poverty and Inequality*.

[13] Sargeant, Adrian (2003), Relationship Fundraising: How to Keep Donors Loyal. *Nonprofit Management and Leadership, Volume 12, Issue 2*.

[14] Sargeant, A. and Woodliffe, L. (2007), Building Donor Loyalty: The Antecedents and Role of Commitment in the Context of Charity Giving. *Journal of Nonprofit and Public Sector Marketing*.

[15] Sundermann, L.M. (2017), "Share experiences: receiving word of mouth and its effect on relationships with donors. *Journal of Services Marketing, Vol. 32 Issue 3, pp.322-333*.

[16] Wash, Rick. (2013), "The Value of Completing Crowdfunding Projects." *Association for the Advancement of Artificial Intelligence*.

[17] White, Martha C. (2016), "Heres How Much Your Kids Teacher Is Shelling Out for School Supplies." *Time Magazine*.

[18] Yao, Kimberley (2015), "Who Gives? The Determinants of Charitable Giving, Volunteering, and Their Relationship." *University of Pennsylvania: Wharton Research Scholars*.

[19] Cleveland, Robert B; Cleveland, William S; Terpenning, Irma. Journal of Official Statistics; Stockholm Vol. 6, Iss. 1, (Mar 1990): 3.

[20] William S. Cleveland and Susan J. Devlin Journal of the American Statistical Association Vol. 83, No. 403 (Sep., 1988), pp. 596-610

[21] Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc., 2001.