Machine Learning for Social Good: Predicting Success of Fundraising Projects Using the DonorsChoose Data Set

Andrew J. Spencer
Princeton University
ajs9@princeton.edu

Jack A. Graham
Princeton University
jag6@princeton.edu

Yanjun Yang
Princeton University
yanjun.yang@princeton.edu

Kevin Tsao
Princeton University
ktsao@princeton.edu

Abstract

DonorsChoose is an online platform that was founded in 2000 and has already raised $685 million for classrooms in the United States of America. Given the importance of education and the number of people willing to donate to further the education of their communities, we are interested in discovering what factors make projects successful. In addition to the categorical data on location, resources requested, and others, we use bag of words and bigram representations for the textual information that teachers use to describe their requests. In our project, we first use PCA and LDA to find latent structure in the data set. We then use four different classification models on our data set to predict if a project will be successful and if the project will receive out-of-state donations. We also used an elastic net regressor to predict the number and amounts of out of state funding received by projects. We found it difficult to predict these outcomes in general, but among our classification models, a Random Forest model resulted in the best accuracy and precision.

1 Introduction

Every year hundreds of billions of dollars are donated to charities in the United States [1]. Despite this willingness of the American people to donate to charities, many schools still lack basic materials needed to deliver a quality education. This could be due to a failure of willing donors to be matched with school projects they may be willing to support. DonorsChoose began in 2000 to help bridge this gap between what donors are willing to contribute to and the causes that need their support.

The goal of our project is to further the mission of DonorsChoose to empower teachers and schools to provide a better education. Previous work on this data set [2] has found that donors tend to donate to schools from the same state as them. In this project, we will analyze this data set to look for what features commonly occur in projects that are fully funded, as well as what features of projects allow them to stand out to out-of-state donors, who in general are less likely to fund these projects. This information can be used by schools and teachers to create projects and generate descriptions and titles that are more likely to attract donors' attention. This information can be used by the platform to give feedback on projects to empower them to succeed. We believe that providing teachers with the knowledge of what can help to make projects succeed can improve the quality of experience for donors, schools, and the DonorsChoose platform.

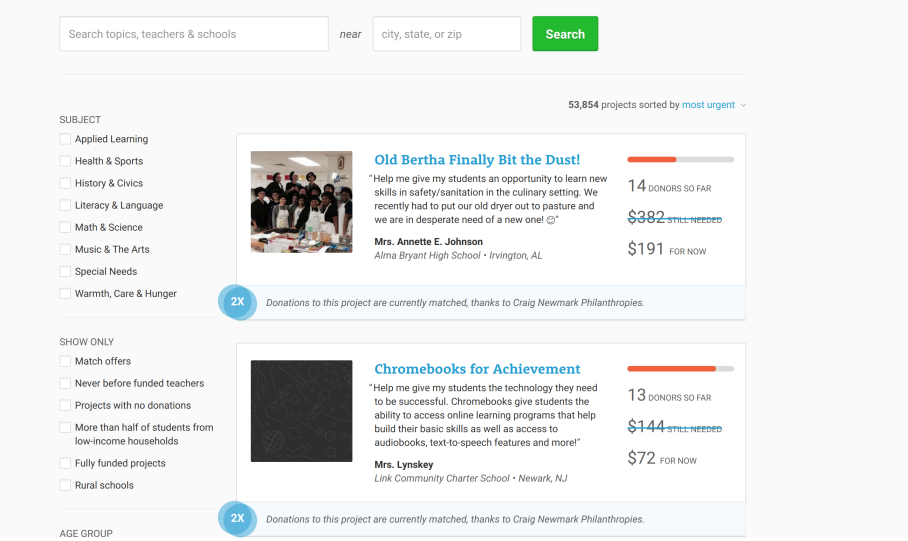

Figure 1: Choosing a project to fund from DonorsChoose

In our project we use two unsupervised learning models to discover latent features and attempt to find patterns that occur in these data. We then use four classification methods to predict whether a project is funded and whether it received out-of-state donations. Additionally, two regression methods are used to predict the amount of out of state donations.

2 Related Work

Our project builds on prior exploratory analysis from James Shepherd [2]. The exploratory analysis found a number of interesting trends in the data. DonorsChoose was started by a teacher from New York, but teachers from California have received more donations than any other state. Donors have a strong preference for donating to causes within their own states, so California's high number of donations could be largely explained by its large population. California and New York are also the two states where donors are most likely to donate to another state.

As a percentage of total donations from the state, Wyoming and South Dakota donate to other states the most often. California has the highest percent of in-state donations, showing that their high number of out-of-state donations is because of the state's large population. Within California, San Francisco has the largest amount of donations of any city. The vast majority of donors only donate a handful of times, and donors are more likely to donate to projects that are similar to projects they have already donated to before. Shepard defined similarity of projects using PCA analysis projected into three dimensions. Shepard's analysis did well at exploring the data set and providing interesting information, and our analysis will build on that exploration by using more sophisticated methods.

3 Methods

3.1 Data Processing and Feature Engineering

The data we used consisted of approximately 5 GB of records split into six CSV files posted on Kaggle [3]. The six files consist of data on donations, profiles of donors, projects that requested donation, resources that were requested by projects, teachers that created fundraisers looking for donations, and schools where the teachers who requested donations worked, respectively. The data originated from a relational database, and each CSV file was a separate table. For this project, we examined the donations, the projects, the profiles of donors, and the profiles of schools. These files contained 4687884 rows by 7 columns, 1110017 rows by 18 columns, 2122640 rows by 5 columns,

and 72993 rows by 9 columns, respectively. See the Appendix for the full names of each of the columns.

To construct the features for supervised learning, we examined the columns labeled "Project Type," "Project Grade Level Category," "Project Resource Category," "Project Subject Category Tree," "Project Subject Subcategory Tree," and "Project Essay." For "Project Type," "Project Grade Level Category," and "Project Resource Category," the data is categorical, mono-exponential (i.e. each data point belongs to exactly one category), and nullable. To construct features from this information, we simply converted them to dummy indicator variables. For "Project Subject Category Tree" and "Project Subject Subcategory Tree," the data is categorical, poly-exponential, and nullable. To construct features from this, we created indicator variables for each possible value of a category, and to represent a data point with multiple values, we set each corresponding indicator to 1. For "Project Essay," the data is textual. We turn this data into features by stemming and lemmatizing each word, removing all stopwords, dropping all words that did not appear often enough, and finally transforming it into a bag of words, as well as a bag of bigrams. For the results of this paper, we used 56,503 total features.

We also constructed 4 outcome variables for supervised learning. The first outcome variable is an indicator of whether a project is fully funded or not. To construct this, we examined the "Project Current Status" column and marked all instances of "Fully Funded" as a positive outcome and all instances of "Expired" as a negative outcome. Instances of "Live" were dropped. The other three outcome variables relate to projects receiving out-of-state donations: whether a project has received such a donation or not, how many of such donations, and how much money came from such donations. To construct these, we matched schools, projects, and donors to donations using their IDs. Fortunately, this was relatively simple, as the data was originally from a relational database.

For unsupervised learning, we used the same strategy as above to transform categorical data in indicator variables and textual data, which come from "Project Title" and "Project Short Description," into bag of words.

As the data set had over a million observations of projects, preprocessing and training our models on all of the data turned out to be unfeasible. Therefore, we only trained our models on a random sample of the data. For the unsupervised learning procedures, we took random samples of 10,000 projects. For classification, we randomly sampled 10000 data points with positive outcomes and 10000 data points with negative outcomes to obtain a balanced training set, and 2000 positive and 2000 negative outcomes to obtain a balanced test set. For regression, we randomly sampled 10000 data points that had at least one out-of-state donation for the training set, and 2000 more data points for the test set.

Furthermore, due to the sheer size of the data, we could not use our own computers to build the features and outcome variables, as they would run out of memory. Instead, we utilized various high performance machines rented through Amazon Web Services. We used a smaller machine with 64GB of RAM and 16 cores to perform unsupervised learning, while we used a larger machine with 384GB of RAM and 48 cores to perform supervised learning. We also parallelized a large portion of our preprocessing code to take full advantage of the high number of cores on these machines.

3.2 Unsupervised Learning Methods

We used unsupervised methods to attempt to understand latent structure in both the categorical and the textual data. First, we used Principal Component Analysis (PCA) to project the categorical data into lower dimensions and then applied K-Means Clustering to investigate the presence of clusters in lower dimensions. We also used Latent Dirichlet Allocation (LDA), which can find latent topics responsible for generating text, on the bag of words representation of project titles and project short descriptions to learn about the latent categories of projects that are not readily available in the data. We hoped to use the results of LDA to supplement the other features when doing supervised learning.

3.3 Supervised Learning Methods

We used supervised learning methods to classify whether a project would be funded and whether a project would receive out-of-state donations, as well as to predict how many out-of-state donors a project receives and the amount of funding from out-of-state donors. For the first two questions,

we used the following classification methods, all implemented in the Scikit-learn libraries [3]. We used the default parameters unless specified otherwise and fitted hyper-parameters via 5-fold cross-validation.

1. *Multinomial Naive Bayes* (MNB), hyper-parameters: regularization constant C
2. *Random Forest* (RF), hyper-parameters: number of decision trees
3. *Logistic Regression* (LR), hyper-parameters: penalty term (ℓ_1 vs. ℓ_2 vs. Elastic Net) and regularization constant C
4. *Support Vector Machine with Linear Kernel* (SVM), hyper-parameters: regularization constant C

We chose these methods to provide a balance between linear and non-linear models in order to capture different aspects of the data. The linear models (i.e. MNB, LR, SVM) are simpler, and as a result, the parameters that are learned by these methods are more interpretable. On the other hand, a non-linear model such as the random forest classifier is able to pick out higher-order relationships, which are very likely to exist in textual data, and may perform better, but in turn, the learned parameters are barely interpretable. Other non-linear models, such as a support vector machine with RBF kernel, were attempted, but required too much computational resources.

To predict the amount of out-of-state donors and funding a project received, we used the following two regression methods, also implemented in Scikit-learn. Again, we tune the hyper-parameters using 5-fold cross-validation.

1. *Ordinary Least Squares Linear Regression* (OLS), no hyper-parameters
2. *Elastic Net Linear Regression* (EN), hyper-parameters: regularization constant α and ratio of ℓ_1 to ℓ_2 penalty

We chose OLS as a simple baseline regression method, and used a regularized regression EN to stabilize the learned coefficients in order to control for overfitting, as we have constructed a large number of features to use. We also standardized each feature before fitting the regressors so that the absolute values of the learned coefficients can be used to rank the learned importance of the features.

3.4 Evaluation

To evaluate our PCA and LDA, we examine the Akaike Information Criterion (AIC) on held out data to determine whether we have a good fit or not. To evaluate the classification models, we test the models on held out data and examine predictive accuracy, precision, and recall metrics. We also examine the corresponding ROC curves. To evaluate the regression models, we test the models on held out data and examine the r^2 values of the predictions compared to the true labels.

4 Results

4.1 Unsupervised Learning: LDA and PCA

Before digging in to the prediction questions of interest, we looked to discern latent structure in the data with unsupervised learning methods. First, we used LDA on the text responses to "Project Title" and "Project Short Description." As discussed in the Methods section, we took a sample of 10,000 projects and ran LDA on bag of words representations, with a vocabulary size of 571 for the project titles and 2071 for the project descriptions.

For LDA, we fit the model with the number of topics ranging from 1 to 20, and found that the AIC on held-out data increased monotonically as the number of topics increased, although the absolute difference was fairly small. See Figure 2 for details. We then examined the latent topics learned by each of the LDA fittings and found that the model with five topics was the smallest model to give reasonable topics for both titles and short descriptions. We thus settled on five topics as a compromise between having a non-trivial number of topics and a decent AIC score. Examining the topics, we find that many of the top words seem fairly generic, such as "fun," "play," and "creative." The topics learned for the for project short descriptions were also fairly hard to interpret, but one

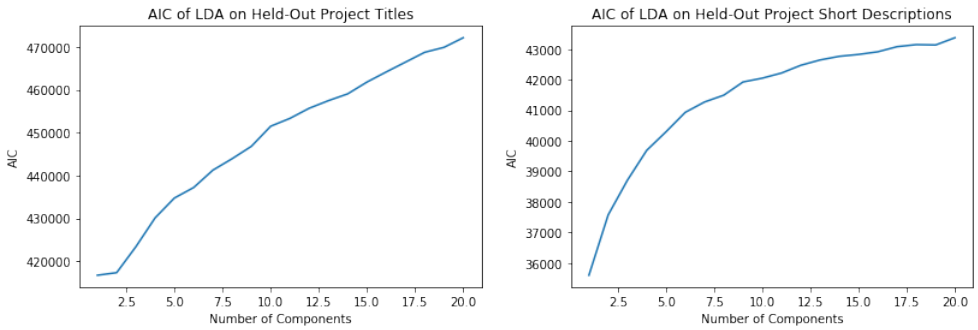

Figure 2: Akaike Information Criterion for LDA on Held-out Data

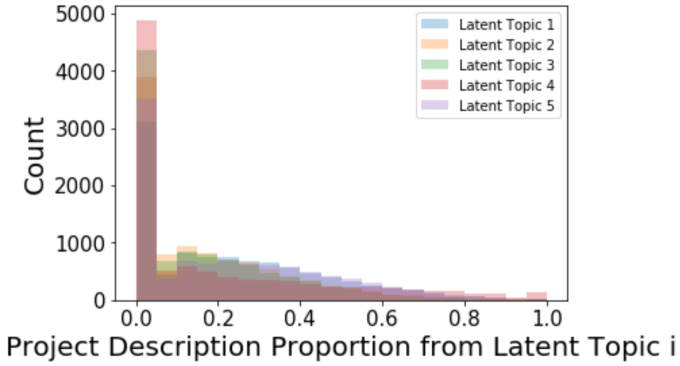

Figure 3: Proportions of topics contained in each document

topic seemed to refer to projects from lower-income backgrounds, with its top words including ”poverty,” ”free,” ”lunch,” ”low,” ”income,” and ”challenge.” Looking at Figure 3, it seems as though there’s a cluster of documents made almost entirely of this topic (Latent Topic 4). See the Appendix for the top 5 words in each topic.

For PCA, we attempted to fit the model with a number of components from 1 to 100, and found that the AIC on held-out data increased monotonically as the number of components increased. See the Appendix for the graph. We were not able to extract useful information from the categorical data projected down into component space, and while we were able to find nice-looking clusters in the reduced space with k-means, the cluster centers projected back into feature space were not interpretable. See the Appendix for the clustering.

We also attempted to use the proportions from each topic as learned by LDA as features in our supervised learning algorithms below, but this did not improve the predictive power of our methods.

4.2 Fully Funded vs. Expired Projects

Despite the amount of data and feature engineering, predicting whether a project will be funded turned out to be very difficult. On a training set of 10000 positive and 10000 negative results, we find that none of our classifiers was able to achieve an accuracy above 60%. In terms of accuracy and precision, the random forest classifier (RF) performed the best with values of 59.4% and 63.2% respectively. This is followed by MNB, SVM, and finally LR for both. In terms of recall, SVM performed the best with 58.6%, followed by LR, MNB, and finally RF. Overall, SVM appears to predict relatively equally as well for both positive and negative results, while the other three tend to predict positive results better. Furthermore, the fact that RF outperforms the three linear classifiers seems to suggest that the features correlate with the outcome variable non-linearly. See Table 1 for details. Surprisingly, we found that increasing the number of data points in the training set does not

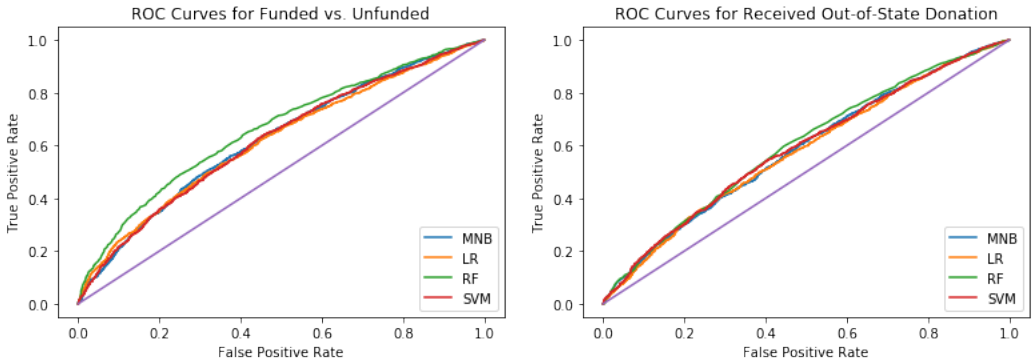

Figure 4: ROC Curves for the classifiers

	Accuracy	Precision (+)	Precision (-)	Recall (+)	Recall (-)
RF	0.594	0.632	0.567	0.525	0.670
LR	0.566	0.588	0.545	0.545	0.589
MNB	0.582	0.618	0.556	0.510	0.660
SVM	0.585	0.594	0.576	0.586	0.585

Table 1: Metrics for classifying whether a project receives out of state funding using 10000 positive samples and 10000 negative samples

improve the results by much. The highest accuracy that we achieved before running out of memory was with RF using 40000 positive and 40000 negative results, with an accuracy of 62.5%. We attempted some specifications with feature selection but found that it not did significantly improve our results, possibly due to there being relatively little signal in the data to begin with.

Furthermore, the ROC curves show that RF slightly dominates all other classifiers, which in turn show similarly-shaped curves. However, all four classifiers have relatively flat ROC curves, which is in line with our other metrics in suggesting that there is a lot of misclassification. See Figure 4 for details.

Examining the parameters learned by MNB in more detail, we find that the most influential features appear to be a mixture of monograms and bigrams related to natural disasters, such as Hurricane Harvey and Hurricane Irma, as well as certain project resource categories, such as "Educational Kits Games" and "Classroom Basics." On the other hand, mentioning technological resources, such as LCD projectors or MacBooks, correlates negatively with a project being funded. Additionally, the parameters learned by LR show that certain project resource categories, such as "Educational Kits & Games," "Reading Books," and "Desks and Storage" were positively correlated with a project being funded, and categories such as "Technology" and "Supplies" are negatively correlated with being funded. We found the fact that projects requesting technological resources are funded less often to be surprising, but after examining the cost of such projects, we found that they request $369.75 more on average, which may explain why they are funded less often.

Coefficients learned by SVM, however, imply that using certain buzzwords, such as "development," "exciting," "hope," "hands-on," "growth," etc. correlate positively with a project being funded, while the more cliché buzzwords, such as "great" and "motivating," correlate negatively. This suggests that word choice may have a noticeable impact on the outcome of a project. See the Appendix for a list of the top 5 features learned by each model.

4.3 Out-of-State Donations

For predicting whether a project receives out-of-state donations, we find that there appears to be an even weaker signal. Of the four classifiers that we tested on a training set with 10000 positive and 10000 negative results, RF performs the best in terms of accuracy, precision, and recall, with values of 58.4%, 56.9%, and 61.4%, respectively. SVM performs second best in terms of accuracy and

	Accuracy	Precision (+)	Precision (-)	Recall (+)	Recall (-)
RF	0.584	0.569	0.600	0.614	0.554
LR	0.546	0.535	0.557	0.552	0.539
MNB	0.552	0.560	0.546	0.497	0.608
SVM	0.572	0.559	0.587	0.599	0.546

Table 2: Metrics for classifying whether a project is funded or not using 10000 positive samples and 10000 negative samples

precision, with values of 58.5% and 59.4%, respectively. MNB and LR follow closely in accuracy with values of 58.2% and 56.6%. Again, it would appear that the features relate to the binary variable of receiving out-of-state donations in a non-linear fashion. See Table 2 for details. Again, increasing the number of data points in the training set does not appear to improve the predictive accuracy of any classifier by much. The highest accuracy that we achieved before running out of memory was with RF using 40000 positive and 40000 negative results, with a value of 60.7%. The ROC curves for these classifiers show that RF dominates the other classifiers for the most part, but all four curves are very close to one another, as well as to the base line. See Figure 4 for details.

Similarly, for predicting the number of out-of-state donations, as well as the total amount of out-of-state donations that a project receives, we see very weak correlation between the features and the outcome variables despite the large number of features and feature engineering that was attempted. The ordinary least squares regressor (OLS) was not very informative as it tended to overfit the data when the number of features was large, and gave poor predictions even when there were more data points than features. On the other hand, the Elastic Net regressor (EN) tended to behave better. On a training set of 10000 projects that received out of state donations, 100% ℓ_2 penalty was optimal according to cross validation during training for predicting both the number and amount of out-of-state donations. When predicting the number of out-of-state donors, EN learns very small coefficients for all of the features, resulting in predictions that differ very little from the average. The training r^2 value was 0.0749, while the test r^2 value was 0.0155. However, for predicting the amount of out-of-state donations, EN is able to pick up more signal with a training r^2 of 0.162 and a test r^2 of 0.131. Again, feature selection did not improve the results.

Examining the parameters learned by MNB in more detail, we find that the most influential features that correlate with receiving out-of-state donations were a mixture of bigrams which include terms related to STEM or special needs, such as "science," "system design," and "benefit hearing." The latter is in line with the parameters learned by LR, where the project category "Special Needs" is among the top most influential features, and with SVM, where the monograms "instrument" and "calculate" rank among the top features that correlate positively with the outcome variable. As for features that MNB, LR, and SVM learned to be correlated negatively with the outcome variable, we see that technology-related features, such as the words "laptop" and "computer," as well as the resource categories "Technology" and "Computers and Tablets" are among the top. This is in line with previous analysis of the negative correlation between technology and projects being fully funded.

However, when only examining projects that received out of state funding, there appears to be a positive correlation between technology and the number of out of state donations received. When examining the learned coefficients of EN, we find that the words "laptop," "ipad," and "chromebook," among others, now appear to be positively correlated with donations. This suggests that donors might favor projects that request funding for technology if given some other signal that the project is worthwhile. Also interesting to note is that when constrained to projects that receive out of state funding, we find that categorical data, such as what types of materials are needed and what subject of study the project is related to, are much less predictive compared to textual features. For example, 5 of the top 10 features as learned by LR for predicting out of state donations are categorical, while none of the top 10 features for EN are. This suggests that once a project becomes "trending" (we take receiving out of state donations as a sign of popularity, since donors tend not to donate to out of state projects), detailed descriptions appear to have a larger impact on getting people to donate. Furthermore, it appears that among the top features that correlate positively of out-of-state donors, the monograms "help" and "money" rank third and fifth, respectively. This

gives some support to the idea that it is advantageous to be direct and concise with one's needs when asking for donations.

This is in contrast to what EN learns about amounts of out-of-state funding received. In this scenario, the learned coefficient show that donation projects that mention buzzwords such as "model," "success," "opportunity," etc., as well as projects that are attached to high schools and/or mention words such as "high school," "trip," "life," etc., tend to bring in more out-of-state money overall. The former is similar to what we discovered for features that predict whether a project is funded. The latter, however, is new. Examining the cost of high school projects (those that belong to Project Grade Level Category of Grades 9-12), we in fact see that the mean cost is *lower* than the overall mean cost over all projects that receive out of state funding ($535.94 vs. $668.22). We also find that individual donations for these projects are on average higher than ($94.32 vs. $78.88). See the Appendix for a list of the top 5 features learned by each model.

4.4 Misclassifications and Large Residuals

Examining projects that are misclassified in terms of being funded or not, we do not see any consistent patterns in both the features used in training and in the original data. The same is true for projects that are misclassified when predicting receiving out-of-state donations, and for those with large residuals when predicted using EN for both real-valued outcome variables. Therefore, it is very likely that the data is simply not predictive of these outcomes.

5 Conclusion and Future Work

Unfortunately, none of the models that we generated performed particularly well in making predictions on the outcomes of projects given the information we had in the data set. We believe that the data that we used was insufficient to accurately predict the projects' results. A possible contributing factor is that the DonorsChoose website contains more information than just the type of project and the project's textual data, and therefore this data set is not fully representative of the information that a donor sees when they access the website. Additional work should be done to look to generate features based on this additional information to determine if these features are more predictive of a project's success.

For example, when viewing the website to decide what to donate to, each project also contains an image. This image is not accounted for in our data set despite likely being the first thing potential donors see. To account for the pictures for each project, we may look to use convolutional neural networks to classify images and convert them into features which we may then use as a feature in our classification models for successes and failures.

The website also displays information on the number of donors who have contributed to the project and the amount raised out of the total amount requested so far. For further work, we would like the time series data on the donations to try to model the probability of success for a project dependent on the amount of time that has passed since the project began and the donations that a project has already received.

6 Acknowledgements

We would like to thank Matthew Myers and Jonathan Lu for helpful comments in office hours. Some code for the LDA section was adapted from Jonathan's notebook for Assignment 3 and code one of us used for that assignment. Some code for the bag of words conversions was taken from the code provided for Assignment 1.

References

[1] Giving USA: Americans Donated an Estimated $358.38 Billion to Charity in 2014; Highest Total in Reports 60-year History https://givingusa.org/giving-usa-2015-press-release-giving-usa-americans-donated-an-estimated-358-38-billion-to-charity-in-2014-highest-total-in-reports-60-year-history/ Accessed: 04/21/2019

[2] Shepard, James. DonorsChoose: Matching donors to causes.

[3] https://www.kaggle.com/donorschoose/ioTeachers.csv

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel,P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher,M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2011

Appendix

1. List of columns from Projects data set: Project ID, School ID, Teacher ID, Project, Type, Project Title, Project Essay, Project Short Description, Project Need Statement, Project Subject Category Tree, Project Subject Subcategory Tree, Project Grade Level Category, Project Resource Category, Project Cost, Project Posted Date, Project Expiration Date, Project Current Status, Project Fully Funded Date

2. List of columns from Donations data set: Project ID, Donation ID, Donor ID, Donation Included Optional Donation, Donation Amount, Donor Cart Sequence, Donation Received Date

3. List of columns from Donor data set: Donor ID, Donor City, Donor State, Donor Is Teacher, Donor Zip

4. List of columns from Teacher data set: Teacher ID, Teacher Prefix, Teacher First Round Project Posted Date,

5. List of columns from School data set: School ID, School Metro Type, School Percentage Free Lunch, School State, School Zip, School City, School County, School District

6. List of columns from Resources data set: Project ID, Resource Item Name, Resource Quantity, Resource Unit Price, Resource Vendor Name

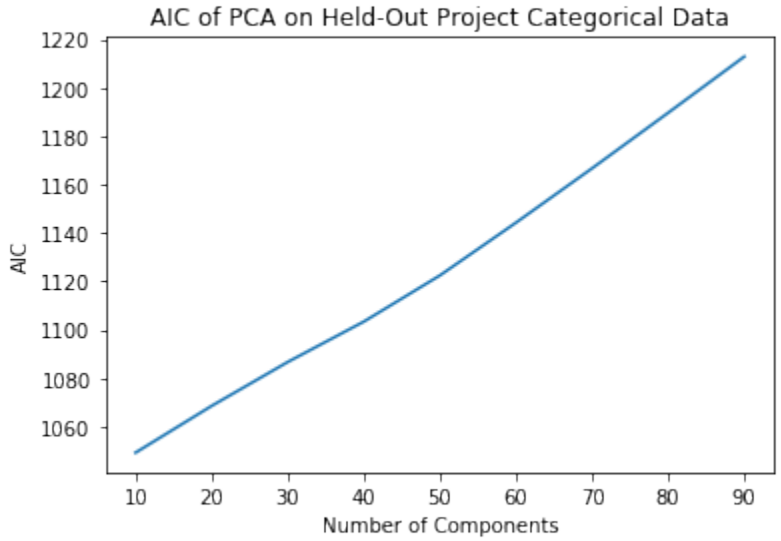

Figure 5: Akaike Information Criterion for PCA on Held-out Data

Project Titles					Project Short Descriptions				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
learner	keep	high	skill	music	aamaz	time	help	background	child
use	come	brain	second	educ	group	music	child	famili	life
play	languag	year	set	life	languag	start	skill	poverti	full
center	kindergarten	see	futur	stem	learn	abl	center	lunch	eager
listen	mind	hear	one	class	classroom	class	hand	challeng	first

Table 3: Top 5 Words in latent topics learned from LDA

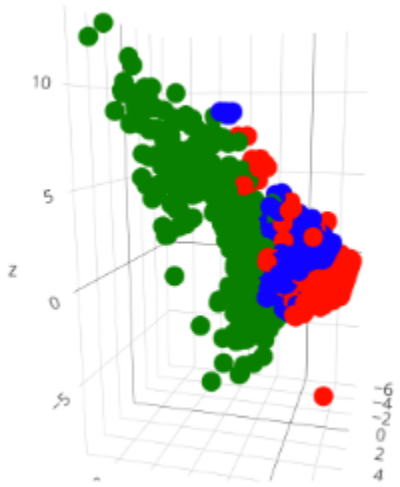

Figure 6: K Means Clustering of Project Categorical Data project categorical data project down to the first three principal components

LR	MNB	SVM
Technology (-)	harvey (+)	Technology (-)
Supplies (-)	hurrican harvey (+)	Supplies (-)
Educational Kits & Games (+)	Educational Kits & Games (+)	Books (+)
hurrican (+)	Classroom Basics (+)	chair (-)
Desks & Storage (+)	hurrican (+)	headphon (+)

Table 4: Top 5 most influential features learned by classifiers to predict whether a project is funded or not (+/- indicates sign of coefficient)

LR	MNB	SVM
make friend (+)	applic skill (+)	laptop (-)
drive (+)	better util (+)	instrument (+)
green (+)	read comput (-)	Supplies (+)
five (+)	scienc elect (+)	unit (+)
seat (-)	accomplish dream (-)	literaci (+)

Table 5: Top 5 most influential features learned by classifiers to predict whether a project receives out of state funding or not (+/- indicates sign of coefficient)

Number of OOS Donations	Amount of OOS Donations
bird (+)	model (+)
studi (+)	educ (+)
help (+)	high school (+)
instrument (+)	high (+)
money (+)	success (+)

Table 6: Top 5 most influential features learned by Elastic Net regressors (+/- indicates sign of coefficient)