Get More Donations for Classrooms in Need

Yi Wan
New York University
New York City, USA
yw1840@nyu.edu

ChiaLing Wang
New York University
New York City, USA
cw2189@nyu.edu

***Abstract*—**

We intent to analyze online donation behavior, the chance a request being funded, and resource distribution in education industry. Our data comes from DonorChoose.org, an online donation platform. We applied statistical methods over 600,000 data entries to analyze online donation behaviors and what factors influence a project's possibility to get funded. Natural Languages Processing Techniques are utilized to extract from project essays the most powerful describing words to increase funding odds. Furthermore, we built a Random Forest model to predict the funding status of a project. We then tested and validated our statistical observations using the prediction model. We will conclude with some observations we made and the testing results of the prediction model and try to explain the senses behind them.

Keywords—Education, Resource distribution, donation, Hadoop, Map-reduce, Hive, NLP, random forest classification.

I. Introduction

DonorsChoose.org is an online charity where a teacher can submit a project request and get donations to help students in need.

Data Source. The first one is Donorschoose.org [1] Open Dataset. It includes: Project dataset: 640,000 classroom projects that have been posted, including school locations, school types, teacher attributes, project categories, project pricing and impact, project donations, projects status. Donation dataset: 3,000,000 donations records, including donor city, state, and partial-zip, payment methods, donation types, donation times and amounts. Project written requests / essays: 640,000 text records of the teacher-written requests accompanying all classroom projects. The essay includes four paragraphs, introducing the classrooms, describing the situation and solutions, and empowering the donors. The second data source is School & Education Data from National Center for Education Statistics (NCES) data via Factual.

Our project focuses on 1) online donation behaviors. What donors care most when choosing the classroom to help? What increases the odds for a project to get funded? Does the teacher gender matter? Does people really would like help kids in high poverty area? What words should teachers use to describe their projects if their want to get more attention and donation.

2) Analyzing education resource distribution via online donation platform. For example, which part of the United States have more poor school that relies on teachers to beg online? What resource are they begging for? We will first analyze project proposal distribution and fully funded project distribution over United Stated to detect if there are some trend in the resource distribution. We then consider total projects / fully funded projects for different poverty level areas to find the correlation between poverty rate and funding rate [2] .

Finally, we build a Random Forest model to predict whether a given project will get fully funded or not. The random forest model is also used to verify our statistical analysis and hypotheses.

II. Motivation

Education resource distribution is not always fair so far. Not every school has advanced technology equipment or even basic supplies and books to let their students enjoy a better study environment and also foster their creativities and skills.

With the recent very popular online education donation platform DonorsChoose.org and Reddit Gifts, teachers post request for teaching facilities, such as books, technologies, and traveling programs and they receive the requested resource if their project receive full funding. From 2000 to 2014, DonorChoose.org has over 1,598,629 who donates $293,006,793 to help 215,028 teachers and 13,432,721 students. These huge numbers trigger people to think the reason why such lack in education resources exist in countries, such as the United States. Yet another situation in reality is that those education resources are never distributed evenly among communities. Some wealthier schools build a “supply center” where teachers can order fancy classroom facilities, while students in poorer district can only rely on their teacher to beg online. The gap made us to reason whether donators' choice will favor to poor community. Therefore, in our project we use the poverty data of the United States to see if the location and its poverty status of the classroom in need have something to do with the donations it received.

We would like to analyze what are the important factors that affect sponsors' decisions. Then we are able to recommend teacher what should they put in their request if they want to receive more funding, or predict the donation it will receive based on the donor behavior and history records. In addition, we are interested to analyze the resource requested distribution together with education budget and poverty data to give government suggestions on batter allocate education budget.

III. Related Work

With the emergences of online charities such as DonorsChoose.org and Reddit and their successes, many researchers have been focusing on mining the online donation behavior.

Heather Long [3] analyzed why teachers should have to be beg online for funding for classroom supplies. The KDD 2014 Cup [4] asks participates to analyze which projects are more "exciting" in terms of being more important to business. The projects requested by teachers are in some degree needed by students, regardless that they are books, technologies or suppliers. However certain projects deserves more attention and thus could be recommended to big donors. Greg Laughlin [5] analyzed whether gender has impact on the number of donation a project received. The author at first assumes that the number of donation a project received is gender-dependent basing on the two statistical observation: (1) there are only about 8,100 male teachers and 560,000 female teachers but the donors per project for projects requested by male teachers is 4.6% higher than those requested by female teachers; (2) The proportion of successful projects for projects requested by male teachers are 3% higher than female teachers. Greg Laughlin's conclusion is that teacher gender is not necessarily correlated with number of donations. Male teachers just happen to teach in places that tend to have more projects being funded successfully. For example, man are more likely to teach in charter school where the projects requested by those school draws more attention than non-charter school. Vlad Dubovskiy [6] analyzed what factors will give more impact on whether a project will get funding from the teachers' perspective. Vlad used statistical tools to find the likelihood of getting funded based on a grade level, project type, school's poverty level, and the describing words. Vlad's experiments show that projects requested from urban and high poverty school and more likely to get funded, and donors also favored to sponsor high school students. The majority of donors prefers to donate on books and suppliers even though technologies are considered to be important for modern classrooms. Among the projects that are requesting for the same resource, literacy and languages projects tend to receive much lower donation compared to other projects. Projects with negative and depress words have higher odds to be funded than those with positive words.

IV. Design

Our design flow is described in Figure 1. Basically, we mainly divided into 4 aspects, the input data, and data process by hive, analysis part and the final output. Below we will introduce more details for each part.

Data Preprocessing. We downloaded the raw datasets from DonorChoose.org, including project list, donation list, and project essays. We also reference other data sources, such as poverty rate by state and National Center for Education Statistics (NCES) [7]. Then we preprocessed and organize raw data using Hive. The downloaded dataset is a CSV file which needs to be transformed into Hive tables. To correctly interpret these data into hive table. We use CSV Serde to serialize and de-serialize the csv file because the raw data field may contain comma so if we directly load CSV into Hive table, several data fields will be misplaced. We also wrote Linux script. mainly uses "sed" commands to remove and replace illegal characters in the dataset because Mahout Random Forest library will encounter parsing error when illegal characters present. There are also many missing values in the dataset which will cause some of our algorithm failure, our Linux script also fills in some of the missing values. Another preprocessing involves add additional variables into existing dataset using Hive. The most important variable added is fully_funded attribute for the project list [8]. Fully funded status is true if a project received equal or more donation than it initially requested, and is false otherwise. Notice we eliminated projects such that their funding status is "alive" because we don't know whether an in-progress project will receive fully funding or not.

Data Analysis. Our major analysis consists of three modules: statistical analysis, word classification, and prediction model. All of the three modules are implemented in map-reduce mode [9] to improve performances Statistical analysis show our observations on online donation behavior and we build our hypotheses on them. Word classification finds out the most powerful words to describe a project so that it will get a higher funding chance. Prediction model is built to further test and verify our hypotheses regarding online donation behaviors.

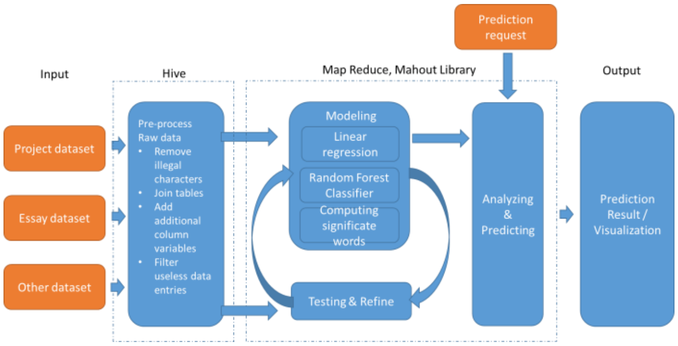


Figure 1 Software Architecture

(1) Statistical Analysis. We use Hive for this part to characterize the overall project distribution and fully funded project distribution. Furthermore, we observe funding status over single variable. Variables includes resource type requested by the teacher, school location, teacher gender.

(2) Word Classification. We coded an map-reduce Natural Language Processing algorithm in Java to extract the most powerful words to describe a project. The procedure is shown in Figure 2. Since we are extracting the powerful words that could lead to be project being fully funded, the input data is therefore essays of fully funded projects.

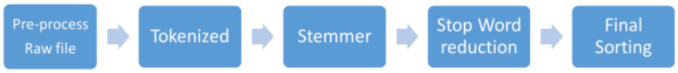


Figure 2 Text classification techniques flow

We first select all the essay from the project table using Hive, and our task is to analysis these words. We used text classification techniques to extract feature from project description [10]. After the preprocessing stage as previously introduced, we need to tokenize each word in the essay data. The challenge is to correctly interpret punctuations. Take single quote ‘’ as an example. The single quote appearing in we’re should be treated to be part of word “are”, while the single quote appearing in ‘pig’ is just used to emphasis, and therefore should not be considered as a part of a word. In implementation, we applied regular expression to tackle this problem when tokenizing the essay, for example to replace the pattern “^'*(\\w+)'*$” with the only word part.

After tokenization, we need to stem the words [11][12]. We use Porter stemming algorithm [14]. The basic idea of Porter’s stemming algorithm is to transform verbs in different tenses into their present tense form. Then our mapper program group together all tokenized and stemmed words and count the occurrence of each word and reducer program sum up the frequency for each word. After calculating the frequency of each word from the 640,000 project essays, we discovered too many high-frequency low trivial words (e.g. for, the). We eliminated those words that have no contribution to the content by deleting them if they appear on our stop list[15]. The stop list is the distinct word lists provided by some online resources [16][17][18]. And finally, we according to the frequency of term appearances in the essays to build the word dictionary by sorting, here we adopt a frequency-based weighting method to find the number of word occurrences in each essay description by using Hive to give the order.

In addition to extracting the most powerful words from all project essays, we are also interested in detecting the most powerful words to describe projects of different topics (e.g. Science, Arts). The subsequent topic related word classification problem are performed using Hive.

(3) Prediction Model. We use Mahout library [19] to build a Random Forest to predict whether a project will get fully funded or not. Our prediction model was also used to test and validate our statistical analysis in the statistical analysis part. We define the funding status as fully funded if the donation received by the project is more than the money requested. To analyze and make predictions about the funding status, we need to choose attributes that are relevant to the funding status from the dataset and a prediction model carefully, the process is shown in Figure 3.

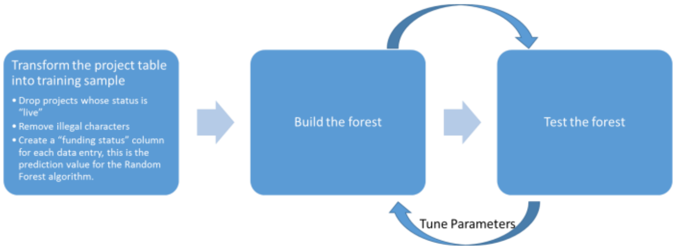

Transform the project table into training sample
• Drop projects whose status is “live”
• Remove illegal characters
• Create a “funding status” column for each data entry, this is the prediction value for the Random Forest algorithm.
Build the forest
Test the forest
Tune Parameters

Figure 3 Predicting funding status for a project

Training Sample. The data we used to infer the funding status of a project includes the location of a school, information of the teacher, resource requested by the project (e.g. book), poverty level of the school, information of the students (e.g. grade level), and funding status. Those attributes are selected as features because we observed their significance in our statistical analysis. To transform the raw data downloaded from DonorChoose into training sample, necessary process includes: (1) drop all the projects whose status are “live”, because we don’t know whether an in-progress project will be fully funded or not; (2) create the prediction value – funding status – for each entry of the training sample; and (3) delete some illegal characters appearing in the field of an attribute. All of the operations above are processed with Hive.

Random Forest Model. We choose Mahout as our data mining library mainly because it able to process large dataset in parallel. Mahout is an Apache open source machine learning library which primarily focuses on recommender engines, clustering, and classification. For classification problems, mahout provides SVM, Naïve Bays, and Random Forest. Random Forest performs well for complex classifier with large training set (less than tens of millions of training examples) and multiple predicator variables [11]. The algorithm is suitable for consecutive, categorical, or text-like predicator variables and deals with non-linear and conditional relationships better than other classification algorithms. In our dataset, there are categorical features (e.g. school_state, resource_type) as well as numerical features (e.g. total_money_requested). Therefore Random Forest turns out to be a reasonable choice.

V. Results

In the following section, we analyze the results of our experiments. We focus on the statistical phenomena of online donation behavior. We try to explain and infer from the statistical results the important factors that will increase the possibility of a project to receive funding. We then evaluate our inferences using the Random Forest prediction model we built.

Experimental setup. We use two datasets in our experiments: projects and donations dataset, and essay dataset. Projects and donations dataset includes 3 million records. Essay dataset includes 620K records. All of our experiments runs on NYU HPC Dumbo cluster. There are 68 compute nodes, each of which uses Linux (Centos 6.5), has 16GB memory, 2x4-core Intel "Harpertown" (c 2008) CPUs, and 1x1TB disk. Visualization tools we used are CartoDB, PowerMap, and Excel.

Statistical Analysis.

Figure 4 shows the projects proposal distribution over the United States. The east coast has more project proposal than the west coast, than the rest of the United States. This observation can be explained by that DonorChoose.org was originated in New York City and expanded towards the west. Despite this Figure 4 also revealed the uneven education resource distribution: classrooms in the southeastern states (e.g. SC, NC) rely more heavily on their teachers to beg online implies the education resources is scarcer for them.

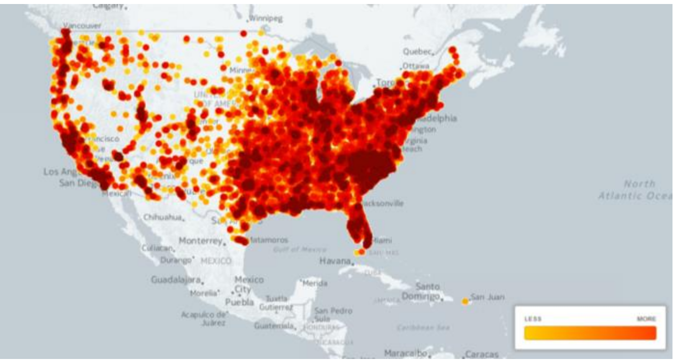
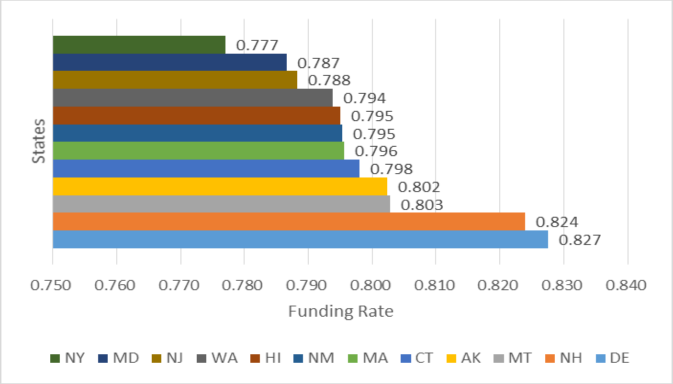
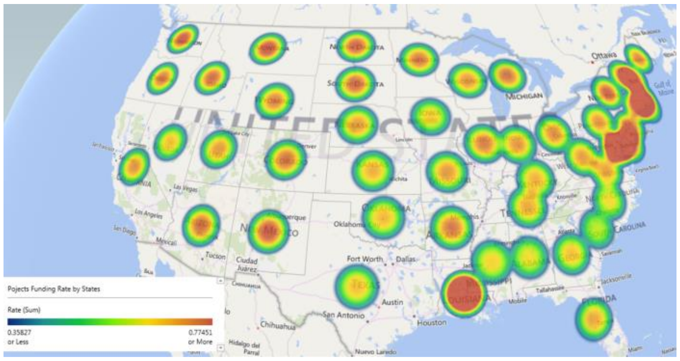


Figure 4 Project Proposal Distribution Over State

Figure 5(a) shows the project funding rates for the top 12 states. DE and NH have higher funding rates than other states. Figure 5(b) illustrates the funding rate over all the states. Comparing Figure 5(a) and Figure 5(b), we can see states with more project proposal do not have more fully funded projects. There is no clear trend how the funding rate will be influenced by location, but it might be influenced by locations. Therefore we use our prediction model to evaluate relationship between location and funding rates later.



(a)



(b)

Figure 5 Funding Rates Over States

We can see from fully funded project distribution of different resource type, as shown in Figure 6, that technologies and supplies are most needy resources for fully funded projects. Therefore we infer resource type are a factor that may influence the possibility that a project will get fully funded.

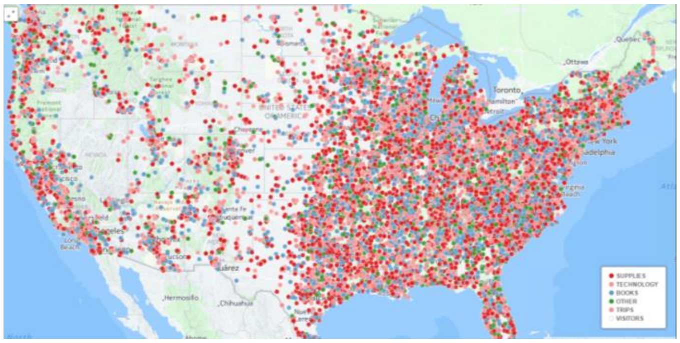
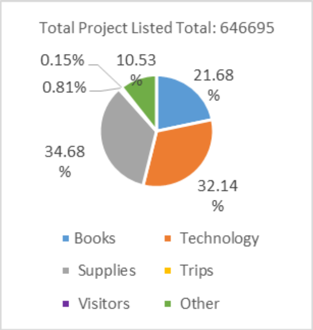
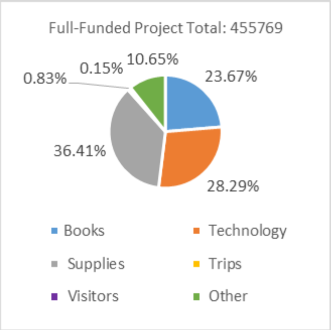


Figure 6 Fully Funded Projects Distribution of Different Resource Type

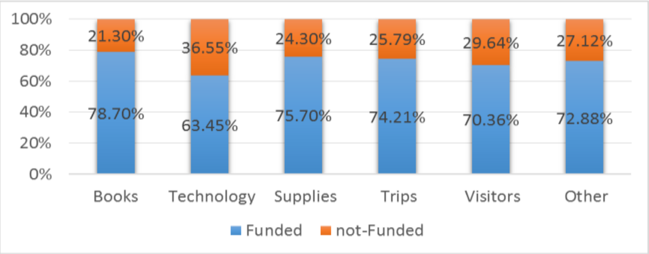
A detailed analysis is illustrated in Figure 7. More projects are requesting for technologies and supplies than books, but the proportion of technologies project in all fully funded projects drop down while proportion of supplies and books raise up. Therefore we infer people may be more likely to donate to books and supplies instead of technologies and the rest resources.



(a)



(b)



(c)
Figure 7 Projects Proposal / Fully Funded Projects / Funding Rate for Different Resource Type

Figure 8 shows that the funding rate for different poverty level areas. As expected, area with highest poverty receive more money than the rest. Difference among the rest of the projects is not obvious. The observation implies poverty rate might influence the odds to get fully funded, therefore we decide to put the poverty level into random forest as a feature.

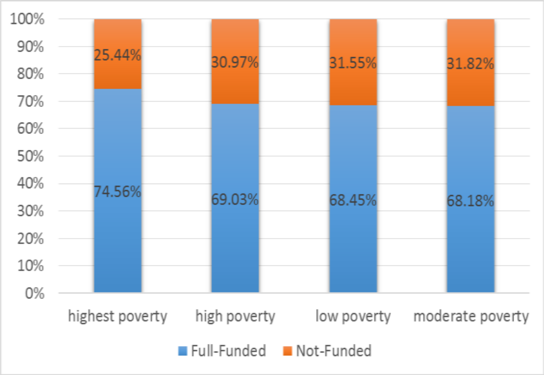
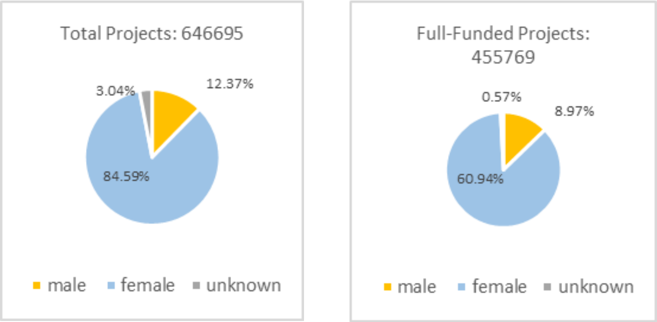


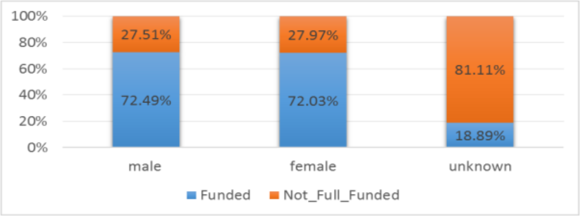
Figure 8 Funding Rate for Different Poverty Level

Figure 9 shows female teacher proposed more projects and has more projects being funded but their funding rate is slightly lower than male teachers. Therefore we infer teacher gender is also a factor that will influence a project's chance to get funding.



(a)

(b)



(c)
Figure 9 Projects Proposal / Fully Funded Projects / Funding Rate Over Teacher Genders

Figure 10 shows the total money requested and donated each year for all the projects. Clearly, DonorChoose.org asked more and received more money every year. The peak of the gap between donation and requested money appearing in year 2009 could be explained to financial crisis.

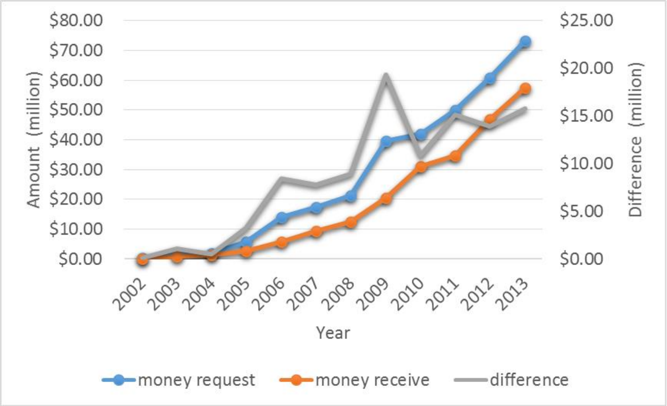


Figure 10 Requested Money vs. Donation Received

Word Classification. According the word classification process result, we've organized the most-frequented words by different essay pattern, totally have 8 kinds of patterns, such as , all the project essays, all the fully funded projects and 6 kinds of project areas includes sport and health, history, applied learning, literacy, math and science, and music and art. From Table 1 , the words appear in all essays and fully funded essays are almost the same, so generally, the word usages don't have much influences to the donors' donation. However, we can still see the top words are most related to the education, such as "student", "school", and "learn". On the other hand, if we select the essays by topics, the appeared terms will have more relatives to the topic, for example, we can find "activity", "physic" in sports and health topic, "history" in history topic, " learn" and "skill" in applied learning, "book", "read" , " language" and even an interesting word "love" appear in Literacy topic, the other two analysis are for Math and Science, and Music and Arts, no to say, both the two words in the topic are almost have high-frequency to appear in the essay on the top 10 words.

Table 1 The most powerful words

All Essays	Fully Funded Essays	Sports & Health	History	Applied Learning	Literacy	Math & Science	Music & Arts
student	student	student	student	student	student	student	student
school	school	school	school	school	read	learn	school
learn	learn	help	learn	learn	book	school	art
read	read	activity	help	help	school	math	music
Help	help	learn	world	classroom	learn	help	learn
need	book	physic	classroom	work	help	science	help
book	need	play	read	year	classroom	classroom	class
classroom	classroom	equip	history	children	love	class	work
all	from	class	class	class	class	work	year
from	class	educate	book	time	work	year	instrument
would	many	time	study	skill	grade	experience	create
class	work	children	teach	teach	year	grade	project
work	do	teach	project	project	write	hand	play
many	more	ball	year	book	time	materi	classroom
do	make	healthy	grade	bulli	skill	project	teach
more	year	year	work	grade	level	skill	experience
make	able	work	time	material	teach	time	love
year	teach	game	social	day	Language	technology	grade
able	grade	skill	resource	read	children	activity	opportunity

Random Forest Prediction Model. We first build our Random Forest using all the attributes that might influence the chance that a project will get fully funded. Then we eliminated some attributes we inferred as "important" from our statistical analysis. By comparing and evaluating the prediction accuracy, we concluded which attributes really matters to the funding status. We split our dataset into 90% training data including 565,929 records, and 10% testing data including 62881 records. We use Mahout Random Forest Library, setting the input split size to be -Dmapred.max.split.size=187423 so that we have 60 mappers.

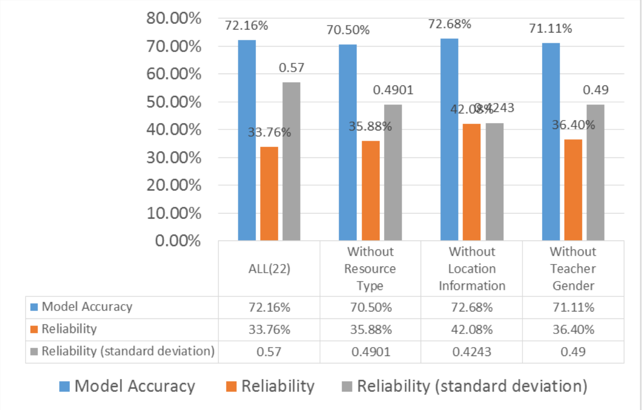


Figure 11 Prediction Accuracy Using Different Attributes As Features

Figure 11 illustrate the prediction results if using different Attributes as Features. Table 2 lists the attributes used as features in our dataset. If all the 20 attributes are utilized to train the forest, the model achieves 72.16% accuracy. The full attributes model was used to compare with other model missing some of the attributes. If we throw away resource type from training sample, prediction accuracy drop down by 2 percent which proved our hypnoses that resource type is a factor that will influence the chance to get fully funded. If we throw away school_state, school_state, school_longitude, school_latitude, school_zip, the prediction accuracy increase slightly by 0.5 percent. Therefore we conclude that location information does not impact a project's funding status. If teach_prefix (denotes teacher gender here) is eliminated, the prediction accuracy drop down by 1 percent. Therefore our hypothesis that teacher gender matters to the funding status is validate.

Table 2 Features used in Random Forest

school_longitude	school_nlns
school_latitude	school_kipp
school_state	teacher_prefix
school_city	teacher_teach_for _america
school_zip	teacher_ny_teaching_fellow
school_county	primary_focus_subject
school_charter	resource_type
school_magnet	poverty_level
school_year_round	total_ptice
resource_type	students_reached

VI. Future Work

We might want to do more analysis from the project attribution, such as whether the students' grade level, the requested money amount, or the beneficial student number will affect the donation behavior, and what percentage of these factors matters? Based on these further detailed analysis, we can build a more well-established recommendation system to those teachers who want to post the project request. By our system teacher can have a rough sense to know whether their project will get the fully funding or not, and even know what factors may cause their failure.

VII. Conclusion

From the analysis and our prediction result above, we can find out that 1) online donation behaviors are influenced by the resource requested, teacher gender. If the requested resource type is technologies, that this project might have less chances to get the full funding. Also project posted by male teacher might have slightly higher odds to get the project fully funded. 2) Project describing words might not have positive effect on increasing a project's chance to get more funding. But we can still see those words includes help, need, students, etc. are still have large relatives to the education and their own project focused area. 3) According to our geographic analysis, education resource are uneven distributed online. There are more project requested in East Coast rather than West Cost. But it doesn't mean that the most request states can get the most full-funding. In this aspect, we can figure out some states might have more request to their education resource but, in fact, get just little or not enough response.

Acknowledgment

Thanks for DonorChoose.org to provide such a great dataset as well an worthy topic to be research about. Also thanks the HPC teams for their support, and we are really appreciated the free space and powerful machines provided from NYU HPC. And finally we especially would like to thanks Professor Suzanne McIntosh who give valuable guidance throughout out project.

REFERENCES

[1] http://www.donorschoose.org/

[2] http://jcberk.com/donorschoose/#p2

[3] Should teachers have to beg online for funding for classroom supplies?

[4] http://www.kdd.org/kdd2014/

[5] http://data.donorschoose.org/male-teachers-get-donations-female-teachers/

[6] http://data.donorschoose.org/the-odds-are-in-your-favor-our-2-cents-on-how-to-get-more-funding/

[7] http://nces.ed.gov/

[8] http://www.datascience-labs.com/hive/hiveql-data-manipulation/

[9] Hadoop_ The Definitive Guide, 3rd Editions

[10] A Text Classification Based Method for Context Extraction from Online Reviews

[11] An algorithm for suffix stripping

[12] Development of a stemming algorithm. MIT Information

[13] Processing Group, Electronic Systems Laboratory, 1968.

[14] http://tartarus.org/martin/PorterStemmer/

[15] An automated domain specific stop word generation method for natural language text classification

[16] http://www.lextek.com/manuals/onix/stopwords1.html

[17] http://www.webconfs.com/stop-words.php

[18] http://www.ranks.nl/stopwords

[19] Mahout in Action