Predictive Analytics for Social Good:

Project Success at DonorsChoose

1. Introduction

DonorsChoose.org is a United States–based 501(c)(3) nonprofit organization that allows individuals to donate directly to public school classroom projects. DonorsChoose.org enables teachers to request materials and resources for their classrooms and makes these project requests available to individual donors through its website. Donors can give $1 or more to projects that interest them, which are searchable by school name, teacher name, location, school subject, material, and keywords. DonorsChoose.org then purchases necessary supplies and ships them directly to the schools. Every project contains a line-item budget and a description of the project. An example could be a project that is needing computer mice and flash drives for the students to use in their Computer Lab or a project that asks for expenses for a field trip to a nearby museum.

2. Problem

For DonorsChoose, it can be useful to know the likelihood of a project succeeding. By identifying the project success, they could improve funding outcomes, improve the user experience, and help more students receive the materials they need to learn. Project success is defined as a project that is fully funded before the expiration date. DonorsChoose could utilize these predictions to identify projects that may not reach their funding goal and help them improve on factors that are lacking. For example, lowering the total expenses, and writing a better essay to sell the project.

The goal of this project is to predict whether or not a DonorsChoose project would be fully funded before its expiration date. Apart from predicting project success, there are other questions that are going to be answered through the exploratory data analysis namely - the factors contributing towards a donor donating to a specific project and factors contributing towards a project not being fully funded.

3. Dataset

The data used for this project was retrieved from DonorsChoose.org open data web site[i].

Projects - The projects dataset includes all classroom projects that have been posted to the site, including school information such as its GPS coordinates and location information such as city, state and

zip code. There are also variables denoting the school's type (charter, magnet, year-round, etc.), the project's category (focus subject, resource type etc.), the project pricing and impact, the total donation amount and the status of the project (completed, expired, etc.)

Donations - The donations dataset includes all the information related to the donors that contributed towards projects such as the project identifier and location information (city, state and zip code). It also includes other information such as donation amount, the source of donation (referral, donation included an honoree etc.) and payment information (account credit, credit card and gift check)

Project resources - The projects resources dataset contains information on all the materials and resources requested for classroom projects, including vendor information such as name and location.

Project written requests / essays - The project essay dataset contains full text of the teacher-written requests accompanying all classroom projects.

For solving this problem, only "completed" and "expired" projects were considered, since "live" projects are not of much value as they cannot be used for training in the supervised learning model built as they are ongoing projects.

There were some limitations of this data. For example, we do not know how a donor came to know about DonorsChoose. Having referral data for all donors would be important for the company and help in ad campaign targeting. It could also help us determine what donors are more likely to return for a donation or refer other donors. We do, however have buyer data for donors who used a web-purchased gift card.

4) Data Preprocessing

Raw data was collected in a structured format in the form of csv files and loaded into pandas data frames. Projects data and donations data were merged in a single data frame to get meaningful information such as - correlation between donor's state and project's state, plot between Primary focus subject and Average donation amount etc.

Initial Data Cleaning:

- Whitespaces from the columns were removed while reading data from the csv files
- Rows that have a mismatch of row values vs the columns were handled.

- Additional unnecessary columns created as a result of bad values in donor comments were removed.

Other Data preprocessing:

- Projects that have 0 as their total expenses were removed as zeros suggest data collection problems since a project cannot cost nothing.
- Year and month were extracted from the date project was posted to use them as features
- Projects and Donations data were merged to get better insights

Data Preprocessing - One Hot Encoding

A decision tree was one of the classifiers used for my problem. And since Decision tree classifier in scikit-learn does not deal with categorical data, One Hot Encoder provided by scikit-learn's preprocessing module was used to convert the categorical features into multiple binary features.

Original Dataframe-

	school_metro	primary_focus_subject	school_state	poverty_level	grade_level	resource_type	year	month
878748	suburban	Health & Wellness	CA	highest poverty	Grades 6-8	Other	2015	8
878763	urban	Literacy	TX	highest poverty	Grades PreK-2	Books	2015	8
878770	urban	Mathematics	TX	highest poverty	Grades 6-8	Books	2015	8
878792	urban	Visual Arts	CO	highest poverty	Grades 3-5	Supplies	2015	8
878804	urban	Literacy	NC	highest poverty	Grades 3-5	Supplies	2015	8

Encoded Dataframe-

	school_metro	primary_focus_subject	school_state	poverty_level	grade_level	resource_type	year	month
878748	2	15	4	1	2	2	13	7
878763	3	17	44	1	4	1	13	7
878770	3	19	44	1	2	1	13	7
878792	3	28	5	1	1	3	13	7
878804	3	17	28	1	1	3	13	7

5) Methodology, Analysis and Results

After exploratory data analysis, the features chosen were - primary focus subject of a project, school information such as poverty level, location of the school(state), whether the school resides in a metro or not, type of school namely – charter school, magnet school, year around school, new leaders school, and kipp (knowledge in power program) school.

Other features include grade level of the students (Grades PreK-2, Grades 3-5, Grades 6-8 and Grades 9-12), the type of resource (Books, Technology, Supplies, Trips, Visitors and Other), total price needed for the project, the year and month the project was posted.

Teacher attributes that were chosen as features were - Does the teacher belong to "Teach for America" organization or has the teacher undergone the "NYC Teaching Fellows" program?

Features related to project donations include:

1) Eligibility for **Double Your Impact** match - A Double Your Impact offer is a way for a corporation or a foundation to match individual donations for projects that meet certain criteria.
2) Eligibility for **Almost Home** match - As an Almost Home partner, a corporation or foundation makes the following offer: We'll fund qualifying projects down to under $100 to go (typically $95-$98), as long as other donors pitch in and fund the rest.

Model

For this problem and the given data, Decision Tree was chosen as the base model mainly because of the dependencies that exist among different features. And for better results, an ensemble approach of Random Forests was used.

A 5 fold cross validated grid search was used to make an exhaustive search over the specified parameter values for the estimator. The scoring function used for the Random Forest Classifier was **F1 score** which can be interpreted as a weighted average of the precision and recall.

$$F1 = 2 * (Precision * Recall) / (Precision + Recall)$$

Precision - The precision is the ratio *tp / (tp + fp)* where *tp* is the number of true positives and *fp* the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The precision here determines, how accurately the classifier was able to predict the successful projects.

Recall - The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples (projects that were funded in this case)

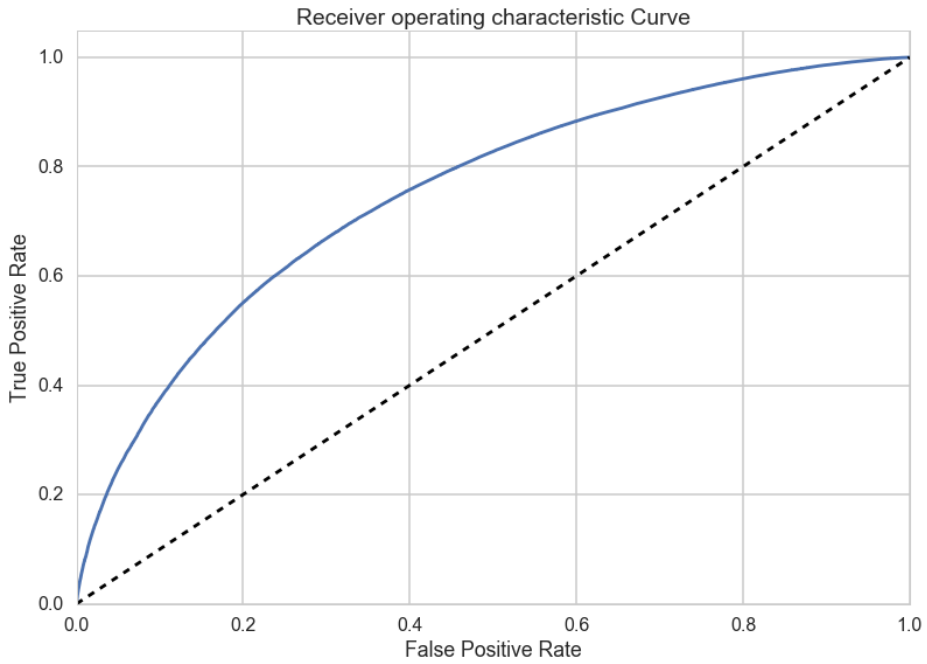
F-Score – Weighted average of the precision and recall

Results

The best estimator was found to have the following parameters –
20 number of trees in the forest,
5 number of features considered when looking for best split,
10 minimum number of samples in newly created leaves,
2 minimum number of samples required to split an internal node.

The results of the best estimator when trained over 60% of data and tested on 40% of data using 5-fold cross validation are:

1. Precision – **0.77**
2. Recall – **0.90**
3. F-Score – **0.83**
4. ROC curve



5. Area under Curve(AUC) – **0.755**

ROC-Curve – ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings[ii]. A random classifier will perform along the dotted line, the further above we are from, the better because it shows that the number of true positives are more than the number of false positives.

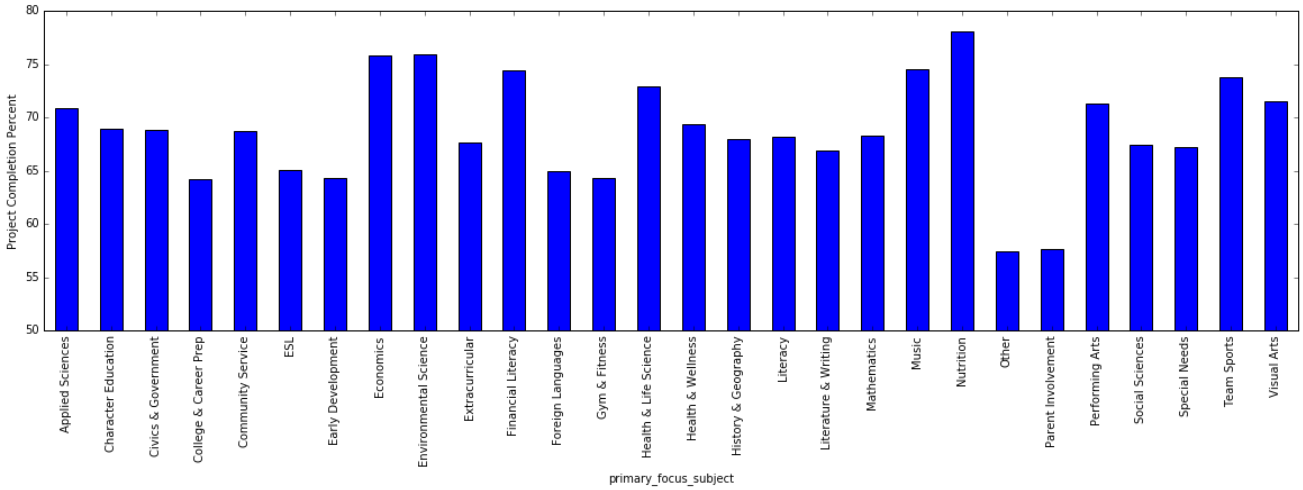
The classifier achieved a decent precision 0.77, AUC 0.755 and very good recall of 0.90 which means, 90% of the actual number of funded projects were predicted correctly, which is a very relevant to the

problem being solved as it is important for the company to identify successful projects so that they can better prepare in terms of resource allocation and ordering from vendors.

6) Other Analyses

Other analyses have been done to give important insights into various questions and problems. Some of the analyses are given below.

1. Analysis between the **subject category** and **project completion percent** (where project completion percent = number of projects completed under a category/ total number of projects under a category



Based on the above figure, we observe the following:

- 'Other' scored the lowest-
  'Other' comes under 'Applied Learning' category. So for a teacher, it is better to have a project assigned to some category rather than having it in 'Other'
- 'Parent Involvement' projects also scored very low compared to other categories.

So looking at the data above, DonorsChoose.org could have a small tip while registering a project saying – Chances of getting a project noticed are high if you have an assigned category.

2. Correlation between project's state and donor's state

States	NY	DC	CA	SC	NV	IA
NY	**57.3121**	0.423177	6.1593	0.0900104	0.0861968	0.161488
DC	11.7059	**27.2631**	4.60637	0.128401	0.0393744	0.275746
CA	9.40677	0.34173	**58.4186**	0.0485834	0.106149	0.0574352
SC	10.8145	0.222739	5.34664	**42.3813**	0.0739814	0.115496
NV	9.52517	0.311637	26.9805	0.021724	**36.9311**	0.0664827
IA	8.59265	1.21878	6.96014	0.0369313	0.071499	**25.6568**

Table - State wise Projects vs Donations

A row here is a state in which a project belongs and the columns are the states where the donations came from. For example, for a project that belongs to New York(NY) state, on an average 6.1593% of the total donations come from California(CA) state. Looking at the diagonal line in this table, it clearly shows that a donor is more likely to donate to a project that belongs to a school in their own state. Apart from that, another interesting finding is that a donor is more likely to donate to a project that belongs to a neighboring state. For example, look at the states below –

States	Percentage Share
DC	27.263134
NY	11.705908
VA	9.365090
MD	7.012761
CA	4.606375

Top 5 donations for a project in District of Columbia

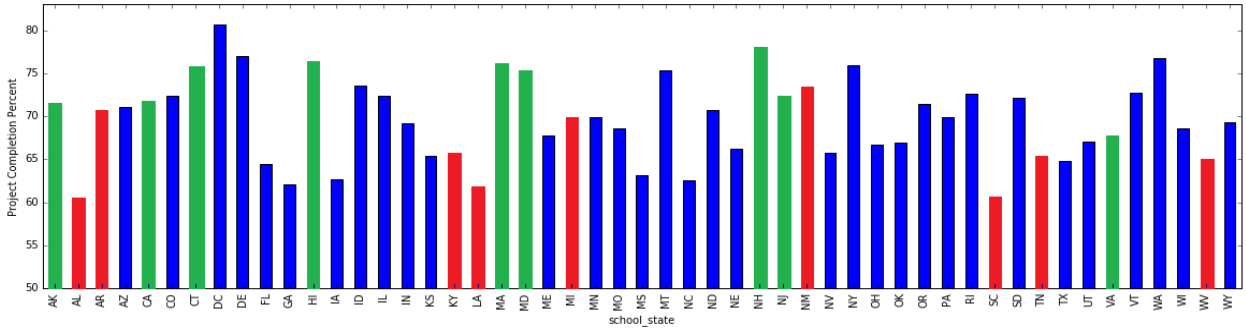
States	Percentage Share
IA	25.656841
IL	25.329843
NY	8.592648
CA	6.960143
WI	3.600954

Top 5 donations for a project in Iowa

For a project in DC, a significant number of donors come from its neighboring states *Virginia* and *Maryland*. Similarly, for a project in Iowa, a significant number of donors come from its neighboring states - *Illinois* and *Wisconsin*.

So looking at the data above, if DonorsChoose.org decide to build a recommendation system that recommend projects to donors based on their interests, the state in which the donor resides could play an important feature as Donors generally prefer a project in their own state or a neighboring state.

3. Correlation between Average Median Household income of state and Project success rate

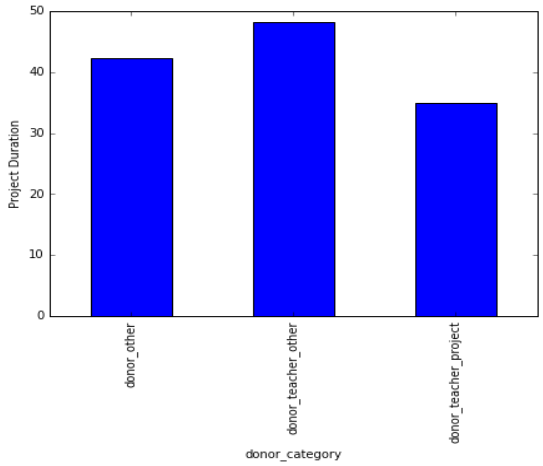


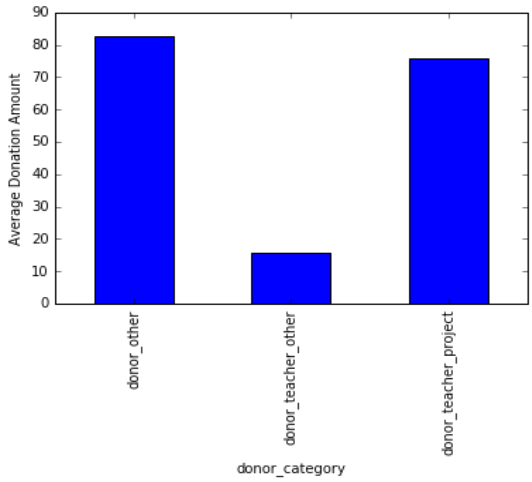
According to *U.S. Census Bureau's 2014 American Community Survey (ACS)*,[iii] the poorest and richest states based on the average median household income, are as follows-

Poor States	Rich States
Mississippi(MI)	Maryland(MD)
West Virginia(WV)	New Jersey(NJ)
Arkansas(AR)	Alaska(AK)
Alabama(AL)	Connecticut(CT)
Kentucky(KY)	Hawaii(HI)
Tennessee(TN)	Massachusetts(MA)
Louisiana(LA)	New Hampshire(NH)
New Mexico(NM)	Virginia(VA)
South Carolina(SC)	California(CA)

The color red in the figure above indicates poor states and the color green indicates rich states in terms of the average household income. Looking at the table and the figure above, we can see a definite correlation between the average median household income of a state and the success rate of a project (the success rate here is the percentage of projects fully funded). This could mean that, the state of the school in which the project is based of has a significant impact on the donation.

4. Analysis between donor category and project duration / donation amount





donor_teacher_project – Teachers donating to their own projects
donor_teacher_other – Teachers donation to other projects
donor_other – Donors other than teachers

Based on the above two charts, the following can be observed:

- A person who is not a teacher is likely to be more generous than a teacher.
- There are a lot of teachers who donate a significant amount to their own projects.
- Teachers contribute very less to projects that they don't own.
- There is a possibility of a teacher donating a significant amount to his/her project to meet the funding requirement and hence the duration of the project is much less compared to other projects

7) Future Enhancements

For the future, a lot of text processing can be done by considering project essays data. This data has a teacher written essay that describes the project. Also the responses given by teacher for each donation could be processed and the sentiment could be found out too. These could be important features in determining a project would be successfully funded or not.

A Recommendation engine could be built which can recommend a donor on what projects to choose based on his interest – subject category, state he is in, poverty level etc. This engine would definitely

help the projects reach their funding goal soon and hence benefit the teacher and the school students and ultimately benefit DonorsChoose.org as the process would be more efficient and this could pull in more projects and donors in future.

8) Recommendations to Client

1) Based on this analysis, the client DonorsChoose.org would know in advance what projects are likely to be fully funded. Because of which the client could prepare in advance by ordering resources, improve funding outcomes, improve the user experience, and help more students receive the materials they need to learn.

2) Based on the data shown in the Preliminary Data Analysis section, we know that projects that have 'Other' as a subject category assigned are more likely to not meet their funding goal. The client could have a small tip while registering a project saying – Chances of getting a project notices are high if you have an assigned category.

3) Based on the correlation we saw in the Preliminary Data Analysis section, the state in which the donor resides could play an important role as Donors generally prefer a project in their own state or a neighboring state. So this information could be useful while displaying advertisements or sending promotional materials to target donors.

---

[i] https://data.donorschoose.org/open-data/overview/
[ii] https://en.wikipedia.org/wiki/Receiver_operating_characteristic
[iii] http://finance.yahoo.com/news/america-richest-poorest-states-040318647.html