“Predictive Analytics for Social good “

Predicting Project Success at DonorsChoose.org

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.

2. Problem

For DonorsChoose.org, It would be better to know if a project would succeed or not well in advance. Project success here is defined as – A fully funded project before the expiration date. By identifying the project success, they could improve funding outcomes, better the user experience, and help more students receive the materials they need to learn. DonorsChoose.org could utilize these predictions to identify projects that may not reach their funding goal and help them improve on the factors they are lacking. For e.g. Lowering the Total Expenses, writing a better essay to sell the project better.

So the goal of this project is to predict if a DonorsChoose.org project would be fully funded before its expiration date or not. Apart from predicting project success, there are other questions that are going to be answered through the exploratory data analysis.

3. Dataset

The data used for this project was retrieved from DonorsChoose.org Open data web site.
Link to data - *http://data.donorschoose.org/open-data/overview/*

- **Projects** - All classroom projects that have been posted to the site, including lots of school info such as its NCES ID (government-issued), latitude/longitude, and city/state/zip.

  IDs - Project ID, Teacher Account Id, School Id

  School Location

  School Types - 6 types - Charter,Magnet,School Year Round,School NIns,School Kipp,School charter Ready Promise

  Project Categories - Primary/Secondary focus subject, Resource type, Poverty level, Grade level

  Project Pricing and Impact - Total price of project, Number of students a project reaches

Project Donations - Total donation amount, Number of Donors

Project Status - Funding Status-Completed/Expired/Live, Date Created/completed/expired

- **Donations** - All donations, including donor city, state, and partial-zip (when available)

  IDs - Donation Id, Donor Account Id, Project Id

  Donor Info - Donor location, Is teacher account ?

  Donation Amount and Type - Type includes - via_giving_page: True if the donation was made through a Giving Page. for_honoree: Donation included an honoree.

  Payment Type - different types of payment includes - Account credit, Gift cards, Credit card, Check
- **Project resources** - All materials/resources requested for the classroom projects, including vendor name
- **Project written requests / essays** - Full text of the teacher-written requests accompanying all classroom projects

Limitations -

We do not have referral data i.e. For most of the cases, we do not know how a donor came to know about Donorchoose.org. However, we do have buyer data for donors who used a web-purchased gift card. Having referral data for all donors would be important for the company and help in ad campaign targeting. It could also help us in determining what donors are more likely to return for a donation or refer other donors.

4) Data Cleaning/Wrangling

Raw data was collected in a structured format in the form of csv files and loaded into pandas' data frames. I had to merge projects data and donations data 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 –

- I had to remove whitespaces from the columns while reading data from the csv files
  donations = pd.DataFrame.from_csv('opendata_donations.csv', index_col=None).ix[:,0:23]
  donations = donations.rename(columns=lambda x: x.strip()) # removing whitespaces from columns
- Some rows had a mismatch of row values vs the columns they belong to. So they were handled.
- Additional unnecessary columns created as a result of bad values in donor comments were removed.

Other Data wrangling –

- I had to exclude projects that have 0 as their total expenses as it is just bad data.
  projects = projects[(projects.total_price_including_optional_support > 0)]

- Extracted year and month from the date the project was posted to use them as features
  projects['date_posted'] = pd.to_datetime(projects['date_posted'])
  projects['year'] = projects['date_posted'].dt.year
  projects['month'] = projects['date_posted'].dt.month
- I had to merge Projects and Donations data to get better insights
  # Join donations and projects data
  projects_donations = projects.merge(donations, on='_projectid',how='inner')
  projects_donations.head(5)

Data Preprocessing - One Hot Encoding

Decision tree was one of the classifiers used for my problem. And since Decision tree classifier in scikit-learn doesn't deal with categorical data, I had to use One Hot Encoder provided by scikit-learns preprocessing module, 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

One Hot Encoding

 # Categorical features- Using OnHotEncode for categorical features
enc = preprocessing.OneHotEncoder()
a = enc.fit_transform(proj[['school_metro','primary_focus_subject','school_state','poverty_level', 'grade_level', 'resource_type','year','month']]).toarray()

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

Data Assumptions –

For solving my problem and answering the questions, I am only considering "**completed**" and "**expired**" projects, since "**live**" projects are not of much value as we cannot use them for training in the supervised learning model built as they are ongoing projects.

5) Methodology and Analysis

After gaining all the domain knowledge from the DonorsChoose.org website and doing a lot of exploratory data analysis had to be done to determine the key features needed to build a predictive model. The features chosen are –

1) *primary_focus_subject*
2) *school_state*
3) *poverty_level*
4) *grade_level*
5) *resource_type*
6) *school_metro*
7) *posted_year*
8) *posted_month*
9) *school-types –*
   'school_charter','school_magnet','school_year_round','school_nlns','school_kipp','school_charter_ready_promise'
10) *eligible_double_your_impact_match*
11) *eligible_almost_home_match*
12) *teacher_teach_for_america*
13) *teacher_ny_teaching_fellow*
14) *total_price_including_optional_support*

Model

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

I have used **GridSearchCV** to make an exhaustive search over the specified parameter values for the estimator. The scoring function used for the Random Forest Classifier was **"f2score"**. And I have used 5-fold cross-validation while searching for the best estimator.

6) Results

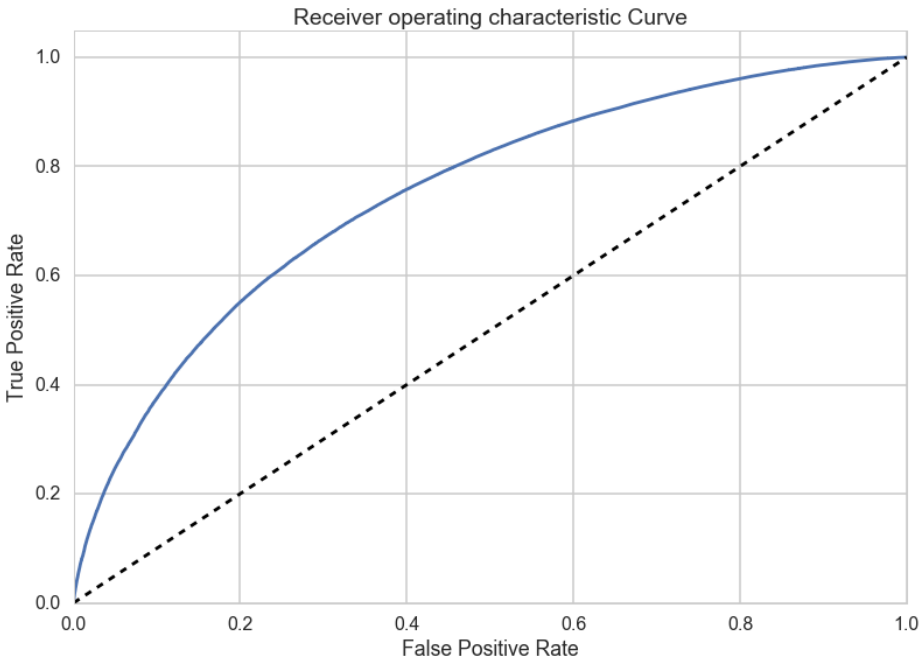
The best estimator was found to have the following parameters –
20 number of trees in the forest,
5 number of features considered when looking for best split,
10 minimum number of samples in newly created leaves,
2 minimum number of samples required to split an internal node.

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. Confusion Matrix

	Predicted(p)	Predicted(n)
Actual(p)	212642	24600
Actual(n)	64262	36929

5. ROC curve

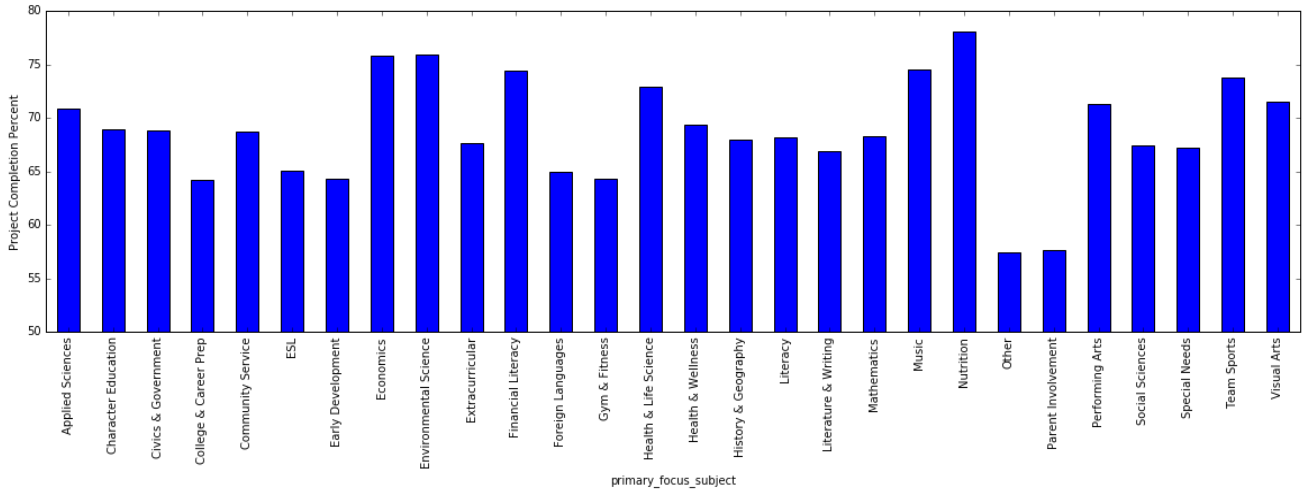


6. Area under Curve(AUC) – **0.755**

7) Other Exploratory Data Analysis

A lot of other exploratory data analysis has been done to give important insights into various questions and problems. Some of the interesting analysis is given below.

1. Plot 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

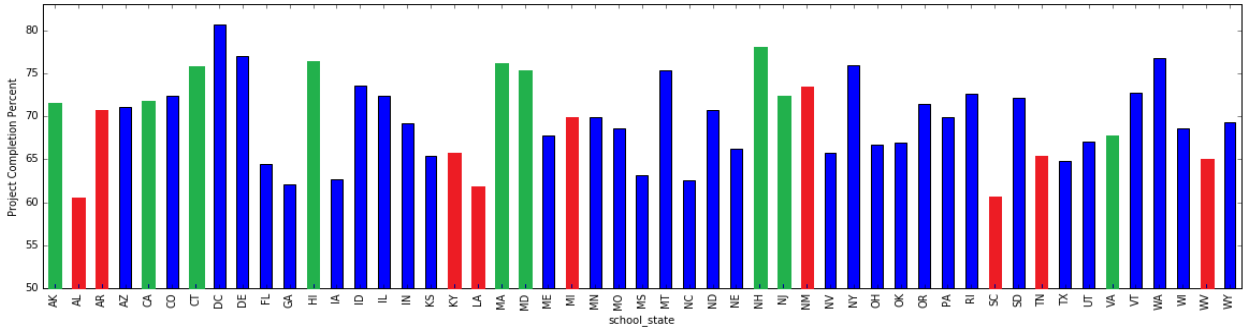
States	Percentage Share
IA	25.656841
IL	25.329843
NY	8.592648
CA	6.960143
WI	3.600954

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

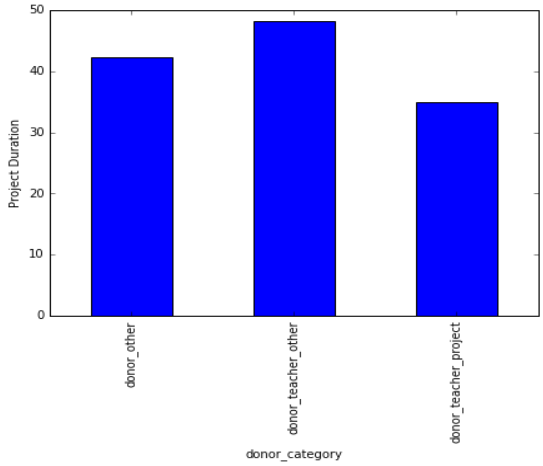


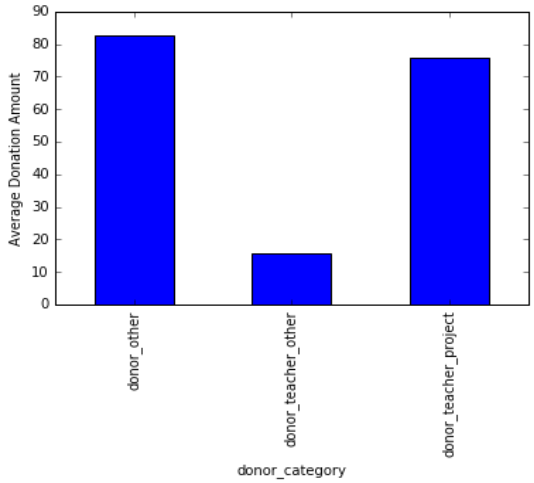
According to *U.S. Census Bureau's 2014 American Community Survey (ACS)*, the poorest and richest states based on the average median household income, are as follows-

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)

Looking at the data 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. Plot 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, we observe the following –

- 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

8) 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.

9) Recommendations to Client

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

2) Based on the data shown in the Exploratory 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 Exploratory Data Analysis section, if DonorsChoose.org decide to build a recommendation system that recommend projects to donor's based on their interests, 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.