Donorschoose.org

(Help DonorsChoose.org connect donors with projects they care about)

Springboard Data Science Intensive Program: Capstone Project #1

Reena Turak

11/13/2018

Table of Contents

Motivation/Background

The impact and power of Data Science are very important factors which captivated me towards it. I was very sure that in my first capstone, I want to do the project which can be used to empower human. So the business use case should be on similar lines and I was looking for some interesting problem to be solved.

My friend told me about this DonorChoose.org and that he donated to it. He shared about some fun projects teacher posted there and how the projects get funded. I found it interested and started looking for their website and projects. While exploring, I came across Kaggle competition and within few seconds I knew this is my capstone project.

Introduction

Founded in 2000 by a Bronx history teacher, DonorsChoose.org has raised $685 million for America's classrooms. Teachers at three-quarters of all the public schools in the U.S. have come to DonorsChoose.org to request what their students need, making DonorsChoose.org the leading platform for supporting public education.

To date, 3 million people and partners have funded 1.1 million DonorsChoose.org projects. But teachers still spend more than a billion dollars of their own money on classroom materials. To get students what they need to learn, the team at DonorsChoose.org needs to be able to connect donors with the projects that most inspire them.

Problem Statement

DonorsChoose.org has funded over 1.1 million classroom requests through the support of 3 million donors, the majority of whom were making their first-ever donation to a public school. If DonorsChoose.org can motivate even a fraction of those donors to make another donation, that could have a huge impact on the number of classroom requests fulfilled.

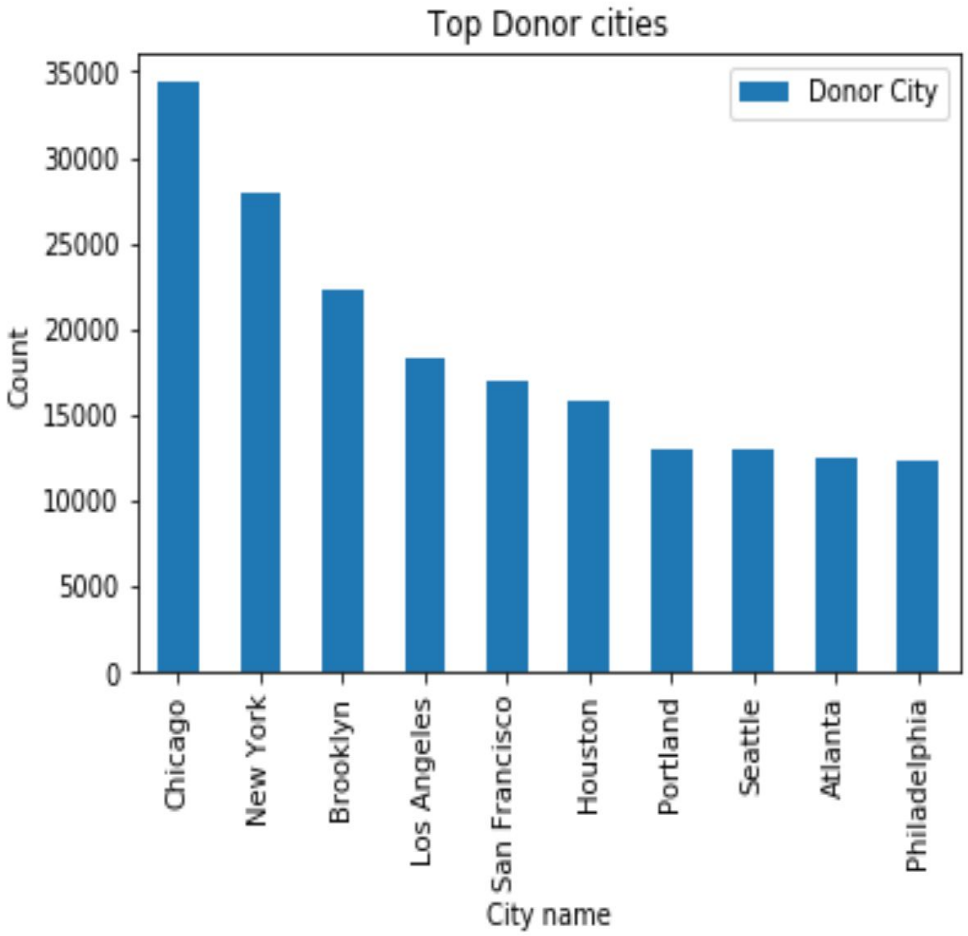
Data Acquisition:

Data source is Kaggle where Donorschoose.org provided dataset for Kaggle competition " Data Science for Good: DonorsChoose.org" -Help DonorsChoose.org connect donors with projects they care about.

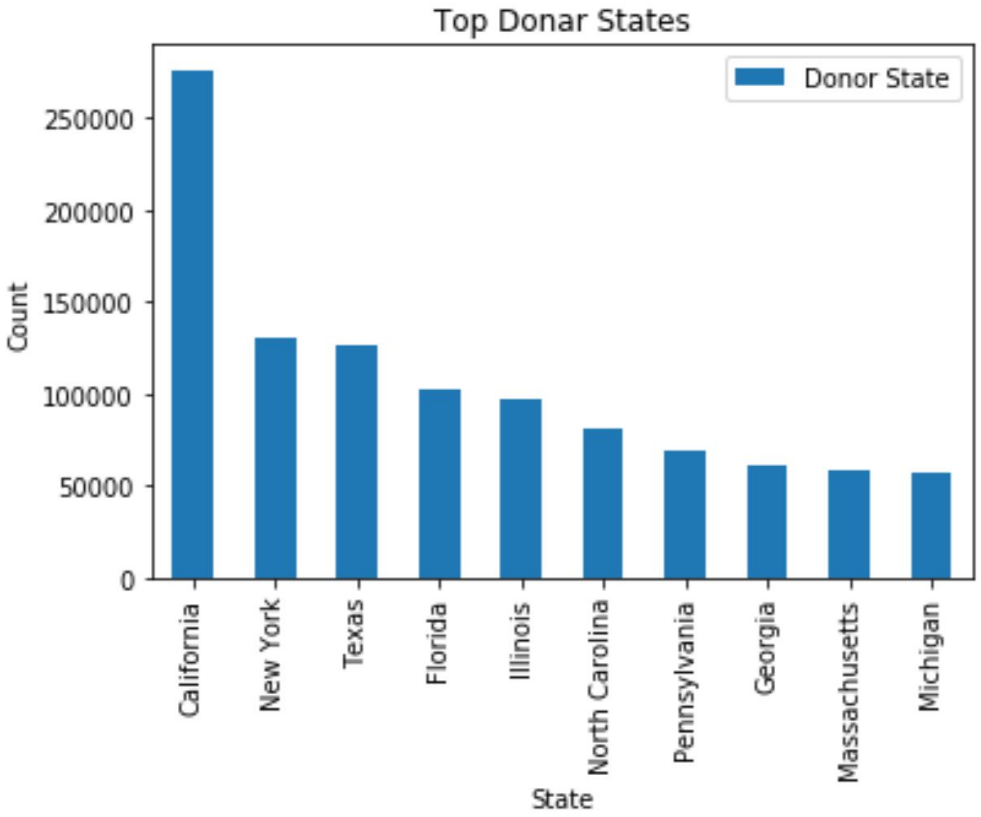
Link: https://www.kaggle.com/donorschoose/io

Exploratory Data Analysis:

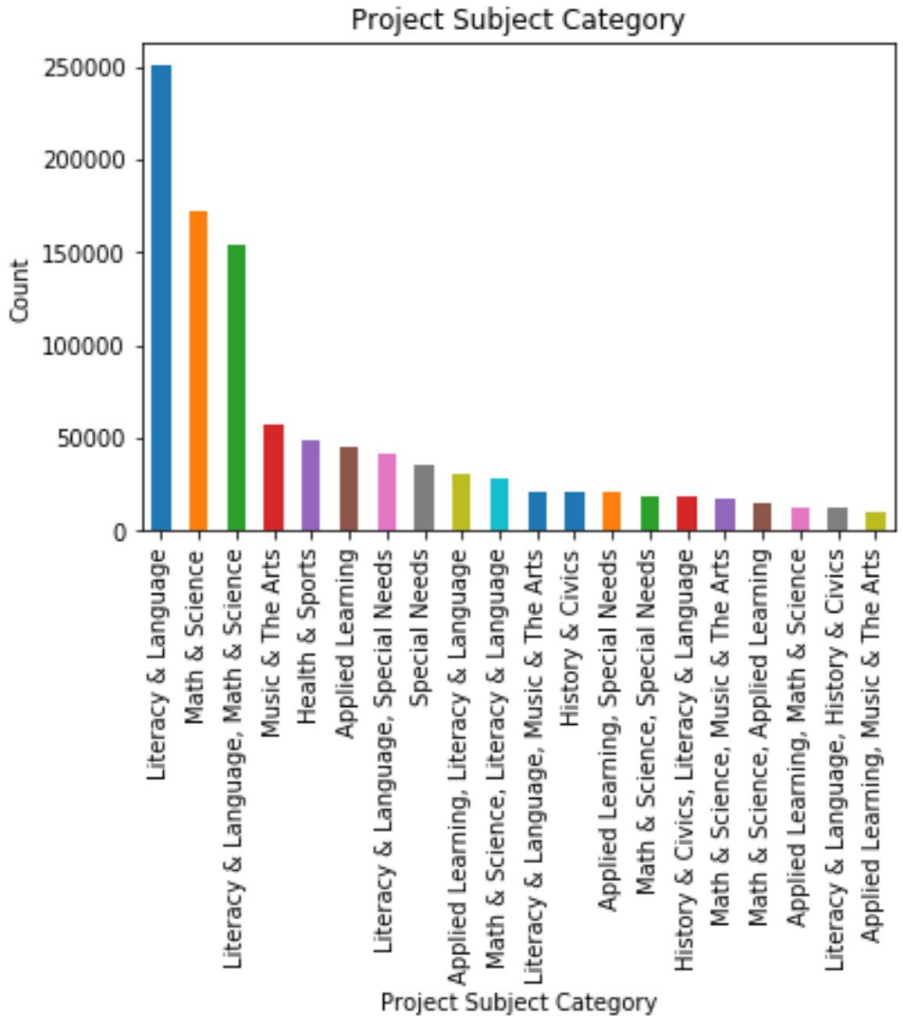
1. Total donation amount raised by Donorchoose.org is $284408243.28
2. Minimum Donation amount is USD 0.01 , Mean donation amount is USD 60.67 and Maximum donation amount is USD 60000
3. Top Donor cities include Chicago, New York, Brooklyn, Los Angeles and San Francisco



4. Top donor state is California, New York, Texas Florida and Illinois

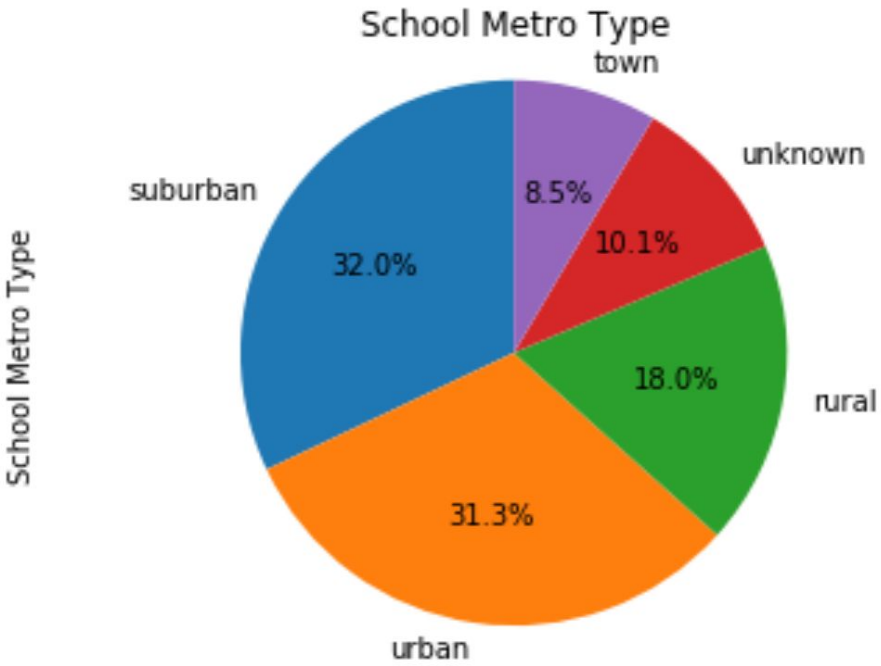


5. 9.5 % Donors are teacher and 90.5% are non teacher.
6. Yearly Donations trend show clearly that No of Donors are increasing
7. donations are mostly received on working days like tuesday and wednesdays.
8. Top Project sub-categories are : Literacy & Language , Math & Science , Music & Arts.



9. Top Project subject Sub-categories are : Literacy , Mathematics & Writing.
10. Most of the projects are led by Teacher and very small amount of projects are Professional Development and Student Led.

11. Metro categories



- Suburban - having 31.5 % schools
- Urban - having 31.2 % schools
- Rural - having 17.8 % schools
- Town - having 8.38 % schools
- Unknown - 11.1 % schools

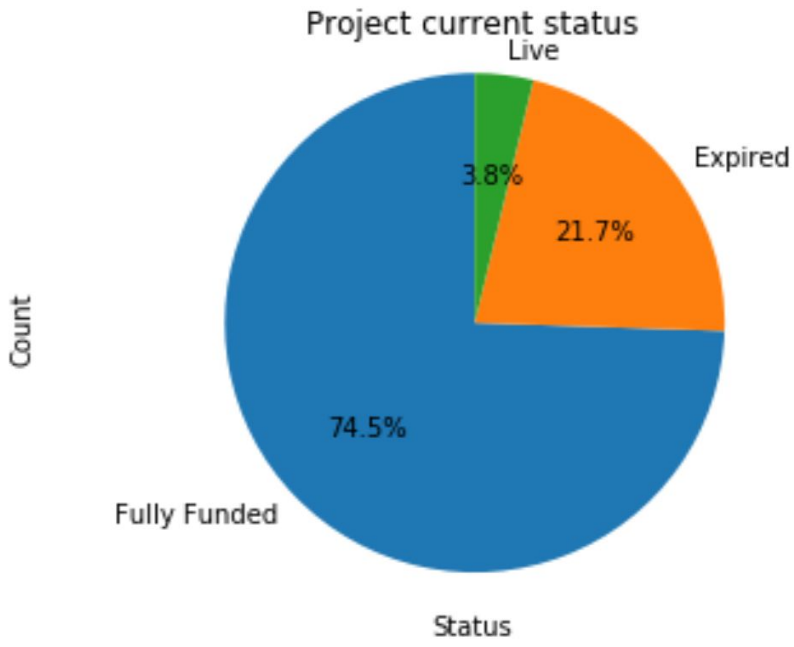
12. 86.4% Teacher who posted the projects are females

13. Top weekdays when teachers posted their first project : Sunday - Approx. 73 K

    Saturday - Approx. 66 K Monday - Approx. 61 K

14. Project Status:

- Fully Funded - 74.5%
- Expired - 21.5%
- Live - 3.8%



Model Selection:

Selected Model:

Bipartite graph recommendation system

Why this Model:

As we know the main problem for DonorChoose is to find matching projects for the Donors so that they can run the email campaign specifically and more efficiently. Basically we need to recommend projects to the Donors which are of more interest to them based on their historical data.
One of the way to approach the problem is user (Donor) similarity.

Below are few very commonly used recommendation systems:

- Content based
- Collaborative
- Hybrid

Though I thought to experiment with the Bipartite graph based on below papers

- https://iopscience.iop.org/article/10.1088/1742-6596/887/1/012056/pdf
- https://ieeexplore.ieee.org/document/7814492
- https://download.atlantis-press.com/article/4652.pdf
- https://link.springer.com/chapter/10.1007/978-3-319-68385-0_14

How it works:

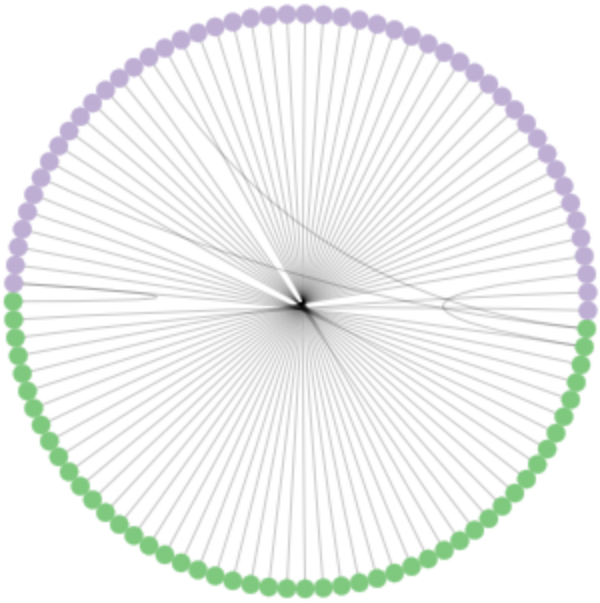
Step 1: Build the bipartite graph

Bipartite have two sets of nodes and nodes from one set can ONLY connect with node/s from other set. Hence all edges share a vertex from both set A and B, and there are no

edges formed between two vertices in the set A. So this is perfect for the use case where we have Donors and projects.

There is one Donor ID and other Project ID graph which have edges connecting from Donors to Projects only.

Below is Bipartite graph:



Step 2: Find most similar Donors

- Most similar donor for the given donor is computed based on similarity matrix.
- Similarity score is calculated based on this formula:
    - Total shared nodes / Total Project nodes

Step 3: Recommend projects

- With given similar Donor , intersection between the two donors is computed.
- For ex Donor #1 (to whom we need to recommend project) and Donor #2 (most similar Donor

Sample of the recommendation system:

For Donor A below are the actual and predicted projects:

Actual Projects	Recommended Projects
Research Tech	Block Out Distractions!
iPad Fundraiser	Robotics Reality
Goodies Galore!	Goodies Galore!
Positioning Ourselves for Success!	Research Tech
None	iPad Fundraiser
None	Learning Tablets for Title 1 Students
None	Writing Is The Paining Of The Voice- Voltaire
None	Social Skills Library
None	Tech For Ms. M's Room

Evaluation metrics used for recommendation system:

- Precision@k = (# of recommended items @k that are relevant) / (# of recommended items @k)
- Recall@k = (# of recommended items @k that are relevant) / (total # of relevant items)

- F1 score is interpret as a weighted average of the precision and recall.

    - **Precision @10 is 0.33**
    - **Recall @10 is 0.75**
    - **F1 score @10 is 0.46**

Technical and Data Limitation:

1. Computing 4.6 million records was not possible. So only sample records were considered.
2. Leveraging GPU and on demand cloud power would have led to better accuracy and speed to train and test the model with different variations.
3. Data was randomly split in training and test set. So it's difficult to find the same Donor who have donation transaction in training, in test set.

Future work:

1. Building weighted edges in the Bipartite graph based on the donation ratio and number of donations to the same project.
2. Leveraging the project category while recommending the projects worth efforts to try.
3. There are many projects whose project title are not exactly matching though the meaning or purpose of the project is same. So using NLP to match and find similar projects like this would be helpful to increase accuracy of recommend projects.