Capstone Project 2 - Milestone Report

Prepared By: Vishal Kumar

Prepared On: March 22, 2018

Application Screening at DonorsChoose.org

Milestone Report

Introduction

DonorsChoose.org is a United States–based 501(c)(3) nonprofit organization that allows individuals to donate directly to public school classroom projects. Founded in 2000 by former public school teacher Charles Best, DonorsChoose.org was among the first civic crowdfunding platforms of its kind. The organization has been given Charity Navigator's highest rating every year since 2005. In January 2018, they announced that 1 million projects had been funded. In 77% of public schools in the United States, at least one project has been requested on DonorsChoose.org. Schools from wealthy areas are more likely to make technology requests, while schools from less affluent areas are more likely to request basic supplies. It's been noted that repeat donors on DonorsChoose typically donate to projects they have no prior relationship with, and most often fund projects serving financially challenged students.

Charles envisioned a platform for individuals to connect directly with classrooms in need, providing materials requested by teachers. With the help of his students, he built the first version of the site in his classroom and invited colleagues to post material requests. Charles anonymously funded the first 10 project requests to demonstrate the effectiveness of the site.

In 2003, the Oprah Winfrey Show featured Charles Best in a segment on Innovative Teachers, which aired on June 20. Traffic generated from the show crashed the site, but viewers donated $250,000 to classroom projects. The sustained exposure from the Oprah Winfrey Show allowed DonorsChoose.org to expand from the New York City region to key metropolitan areas across the country.

In 2006, following the destruction of hurricanes Katrina and Rita, the site opened to public school teachers in Louisiana, Mississippi, Alabama, and Texas. In 2007, the site opened to every public school in the United States.

As of September 2016, donors have contributed over $400 million to fund more than 750,000 classroom projects posted on the site, reaching 19 million students in public schools across the United States

Problem Definition

DonorsChoose.org receives hundreds of thousands of project proposals each year for classroom projects in need of funding. Right now, a large number of volunteers is needed to manually screen each submission before it's approved to be posted on the DonorsChoose.org website.

Next year, DonorsChoose.org expects to receive close to 500,000 project proposals. As a result, there are three main problems they need to solve:

1. How to scale current manual processes and resources to screen 500,000 projects so that they can be posted as quickly and as efficiently as possible

2. How to increase the consistency of project vetting across different volunteers to improve the experience for teachers

3. How to focus volunteer time on the applications that need the most assistance

Potential Clients

This model can be used by any fund raiser companies as a filter to select certain qualifying projects to be posted on the internet. For example, online fund raiser companies would get hundreds of request everyday and they must analyze projects importance and need, before they can be posted on the website, this model will help them reduce the time spent on analyzing these project manually.

Causes, Crowdrise, DonateNow/Network for Good, FirstGiving, FundRazr are few companies which can make use of this model with little bit of update with the model.

I believe this model can also be used by public school teachers across the U.S. to prepare better proposal and submit it to the DonorsChoose.org.

About Data

The dataset contains information from teachers' project applications to DonorsChoose.org including teacher attributes, school attributes, and the project proposals including application essays.

https://www.kaggle.com/c/donorschoose-application-screening/data

File descriptions

- train.csv - the training set
- test.csv - the test set

- resources.csv - resources requested by each proposal; joins with test.csv and train.csv on id
- sample_submission.csv - a sample submission file in the correct format

Data Cleaning/Wrangling Techniques

Generally we do not get the data in the formate best suited for our problems. We always need to do some data cleaning and wrangling to bring the data to the desired format. Since this is an ongoing competition dataset, the dataset is pretty much clean. But to do some of the analysis we have to wrangle some of the data like time series and text based analysis.

Most of the dataset is clean but we did find some missing values. NaNs, and there were some NaNs which got generated from calculations. These values dealt with ***fillna()*** method accordingly to the situations.

- Teacher's prefix column has 4 NaN values, which were filled with "Teacher" prefix which was more suitable.
- During calculation of price range, there were price ranges in which there were no submission, for calculation of approval rate these values got NaN values, I filled it with 0 (zero).
- The 3rd instance of NaN values were due to process change while filing the application. The 4 essay parts became only 2. Hence NaN values in columns essay_3 and essay_4. We have combined all 4 part to just 2 parts according to the given definitions of all the 4 essays.

Data Wrangling

Here, will discuss more about the data wrangling steps I have taken to perform my analysis. Some of the techniques have been used through out the notebook.

- During analysis based on gender of the teachers, I had to combine doctors and teachers as unknown genders, hence categorizing the entire dataset into three category, Female, Male and Not Known.
- I have used ***split()*** function on project subject category and project subject sub category. By doing this we could have reduced the number of unique subject to 9 and unique subject sub category to 30, which was easy to handle and made more sense for the analysis.
- Time Series analysis, we derived 5 new columns using ***datetime*** data type of python. These new columns are date, month, year, day of the week, time. This helped me analyze which time during a year has the highest number of submission and what time it remains relatively dull. Then I divided the day into 4 categories as below and grouped the data accordingly.
  - 06:00 - 11:59 Morning
  - 12:00 - 15:59 Afternoon
  - 16:00 - 19:59 Evening
  - 20:00 - 05:59 Night(nights are long)

- To create a price range, I have used ***itertool*** package along with ***tee, islice, chain*** of python. I have also used a special pandas method ***pandas.cut()***. The ***cut*** function is useful for going from a continuous variable to a categorical variables.
- For text and word cloud analysis, I have used some basic string operations of python like ***.join() and split()*** method.

Apart from this, ***round() and lambda()*** methods have been used in almost all the section to bring the floating point data in to 2 decimal places, which is more readable in percentage calculation.

Initial Findings

The overall approval rate of a project, going to be posted on DonorsChoose.org is at **84.77%** which is very good. It means that any teacher's proposal at the time of submission has 85% chance that his proposal will be approved to be posted on the website.

Now let's see some insights of the datasets.

1) States wise:

California has the highest number of submission in the U.S. **Texas, New York**, and **Florida** are on 2nd, 3rd and 4th position respectively. California also tops the list of most number of approved projects, New York and Texas to follow. So when the number of submission is high the number of approved proposals are also high.

- All the states has more than **81%** of approvals of the total proposals submitted to DonorsChoose.org
- **Washington D.C, Texas** and **New Mexico** have slightly higher rejections compared to other states.
- a very small state in east **Delaware** has the highest approval rate followed by **Wyoming** and **Ohio**.

2) Teacher's gender wise:

Female category has the highest number of submission and hence approved proposals from Female teachers are very high. Proposal submitted by **female** teachers **88%** compared to male which is only **10%** of the total submission. **Male** and **Female** approval rate is almost same around **84%** and **85%** respectively. Even teachers with not know gender has a very good approval rate of almost **80%.** There are around **2%**of teachers who can't be identified as male or female.

3) Grade wise:

The proposal submissions for grades **Prep-2** and grades **3-5** is pretty high compared to grades **6-8** and grades **9-12**. However the approval rate for each of the grade category is between **83%-85%**, which is good.

4) Subject category and subcategory wise:

Subjects like **Care & Hunger** and **Warmth** have 0% approval rate for some of the sub-categorical subjects, thats because there is no submission in the sub-categorical subject. **Applied Learning, Literacy & Language, Maths & Science** subjects have better approval rates ranging from **60% to 80%.**

5) Time series:

The training dataset contains full one year data from end May 2016 to April 2017. **April, May and June** months are dull, the DonorsChoose.org does not get large number of proposals compared to other months.

Aug and Sep are very busy months in terms of proposal submissions. Thats because the the time when the school starts new terms.

Tuesdays, Wednesday, Thursday are most preferred days of the week to submit the proposal. In fact slightly more than **50%** proposals are submitted on these 3 days. **Saturdays and Sundays** are the least preferred day of the week to submit a new proposal.

Most teachers choose **night time** (after 8:00 Pm and before 06:00 Am) to submit a proposal. Morning time from 6am to 12 noon is the least chosen period of day to submit a proposal.

Number of approved proposals are in the same order as of the total number of submission.The approval rate is around **80%-85%** in all category.

6) Previously posted proposals wise:

The number of proposals by **new teachers**, that is teachers with no record of submitting proposal are more than **50K** which is almost like **one third** of the total number of proposals. So it doesn't look like the number of previously posted project has any affect on approval of the project applications. There are teachers very few, with very high number of previously posted proposals, with **100%** approval rates.

7) Resources and the Price wise:

The minimum total amount requested for a project is **$100** and the maximum is around **$15300. 75%** projects requested about **$690.** As the price **increases** the number of projects **decreases** significantly. There are just 1 or 2 projects which requested for high amount of more than **$10K.**

8) Text analysis and WordClouds :

Project Title:

Funny titles like **Wiggle While You Work!** is the title which has been used very significantly and the success rate is more than **90%** which is amazing. **Technology** and related words are most used words in the project title. **Science** and **math** are also seems very popular words. Words like **classroom, books, building, flexible seating** reflects the infrastructure need of the students for better **learning**.

Essay:

Essay 1, most of the frequently used words are related to **students** as this field required to write down about students and class. There is no much difference in the frequently used words in approved and rejected proposals, except word "**Classroom"** which has used frequently in proposals which has been rejected.

Essay 2 is more about problems which students face, things which will help them and how exactly will they help to improve the learning. Words like, **"small group, seating options, flexible, well, use"** are frequently used words. There seems to be many difference in frequency of words used in essay 2 for approved and rejected proposals.

Resource summary:

The most frequently used words are related to the object, like **chair**, devices like **ipad mini** and words that describes these objects. As the summary is about the resources requested we would expect something like this to improve the learning environment. The are sports accessories related words are also frequently used in describing the resources. The words are almost same in both, approved and rejected proposals, just frequency of these words are different.