CS6460 Educational Technology Project Proposal

ykang84@gatech.edu

Introduction

DonorsChoose.org, founded in 2000 by a history teacher, is a website that allows crowdfunding to make a difference in public education. It has raised $685 million for America's classrooms so far [1]. Teachers at 79% of public schools have posted projects that need help on DonorsChoose. Stephen Colbert, the well-known talk show host, has endorsed the platform, become a board member, and funded all the classroom projects in South Carolina in 2015 [2].

This research-oriented work aims to investigate a problem proposed by DonorsChoose: how to efficiently connect donors to the projects that motivated them. In order to efficiently feed projects to potential donors, machine learning (ML) techniques can be utilized to explore and enhance the performance of this platform. Improving public awareness of a platform like this will also substantially contribute to public education.

Literature review

ML approaches have been emerging in various fields recently and education-related data is not an exception. One of the most common application of ML techniques to education data is to predict students' performance. For example, Cortez et al. studied student achievement in secondary education of two Portuguese schools and built classification models to predict grades [3]. The results found that high predictive accuracy can be achieved provided that previous school period grades are known.

However, there was limited literature targeted at public education donations in particular. Instead, existing studies focused on education-related fundraising were reviewed.

Liang used various ML approaches, including Gaussian Naive Bayes, random forest, and support vector machine algorithms, to study fundraising success in higher education [4]. These algorithms were able to distinguish promising donors from non-promising donors, at an overall accuracy of 97%.

Wastyn did a qualitative analysis based on his 14 years of expertise in fundraising [5]. Her study concluded that where donors and non-donors differed was in the ways in which they socially constructed their college experiences to create their own realities. However, such research can be biased and subjective depending on the researcher. Hence, its value may be limited to a specific circumstance.

Wesley and Christopher used statistical logit regression analysis to predict the individuals who would give higher (e.g., $100,000) or lower ($1,000) donations based on the data from the alumni database as well as the geo-demographic information [6]. The models were developed from the alumni database at Northwestern University for both major gifts and annual fund prospects.

However, there is a difference between the existing studies and the current study. Previous studies were mainly focused on identifying promising donors, while the current study aims to explore relationships between multiple groups.

Data and methodology

History data on previous donations have been made publicly available, which is 1.27 GB with 6 files. The data includes information on schools, teachers, resources, donors, projects and donations. Unique IDs were created for each teacher, school, donor, donation and project, respectively. Multiple tables can be joined using these IDs. A lot of information will have to be extracted from the text, and possibly categorized by labels before use. The descriptions of the tables are as follows:

- Donations: project ID, donation ID, donor ID, Donation Amount, date
- Donors: donor ID, donor city, donor state, donor is teacher (yes or no), zip code
- Projects: project ID, school ID, teacher ID, project type, project title, project essay, etc.
- Resources: project ID, item name, quantity, unit price, vendor name
- Schools: school ID, school type, state, city, zip code, metro type, etc.
- Teachers: teacher ID, prefix and first post date.

It is a bit unexpected to me that the data only included projects that received at least one donations. No response variable is present as well. Thus it will not allow investigating the reasons some projects failed to get donations.

One of the most feasible approaches is to build profiles of the donors, the teachers and possibly the projects. Unsupervised learning methods, e.g. clustering, can be utilized to group similar profiles into clusters, hoping to achieve maximum heterogeneity between clusters. This will be similar to problems like auto-grouping students' profiles in C. Adam Harper's project in previous semesters.

Prior to unsupervised learning methods, the data has to be preprocessed. Applications like one-hot encoding and possible natural language processing techniques are expected. All analyses would be implemented using python packages. Nonetheless, this study attempts to reveal valuable information from the available data.

Implementation plan

Deliverable	Schedule	Weekly Task
Status check 1	6/24/2018	Data preprocessing
Milestone 1	7/1/2018	Text mining
Status check 3	7/8/2018	Exploratory analysis
Milestone 2	7/15/2018	Data Visualization and modeling
Status check 5	7/22/2018	Model estimation and wrap-up
Final submission	7/29/2018	Final presentation

Milestone 1

Data cleaning and preprocessing is crucial, as it is the foundation for subsequent analysis. For this study, the data preprocessing includes extracting information from textual data. Particularly, most important information regarding projects will need text mining efforts. So for the first milestone, the data should be cleaned and ready to be analyzed on. Information should be extracted and transformed to new columns if suitable.

Milestone 2

If the data were cleaned in the previous weeks, the next step would be exploratory data analysis and visualization. As shown in the data descriptions, multiple tables included spatial information. The geo-spatial info could be handy in revealing interesting findings, e.g. areas that receives the most or least donations. In this phase, it would also be clear what kind of in-depth analysis is available. Considering the data itself is mainly spatial and textual, Latent Dirichlet Allocation and geospatial mapping would be explored first.

To summarize, after data cleaning, perform explanatory data analysis on each one of the files to get meaningful insights firstly. Then build profiles of donors, projects and etc. based on the input data. Visualize data and implement in-depth analysis. What will be available for in-depth analysis is to be found out. For instance, it might be feasible to estimate a count model for donation frequencies.

References

1. DonorsChoose. (2018, June). Retrieved from https://www.donorschoose.org/
2. CBS. (2016, Mar 10). *Stephen Colbert on $14M DonorsChoose.org funding by public figures*. Retrieved from https://www.youtube.com/watch?v=I4GMX0MJNYw
3. Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)* pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
4. Liang, Y. (2017). *A Machine Learning Approach to Fundraising Success in Higher Education* (Doctoral dissertation, Department of Computer Science, University of Victoria).
5. Wastyn, M. L. (2009). Why alumni don't give: A qualitative study of what motivates non-donors to higher education. *International Journal of Educational Advancement*, 9(2), 96-108.
6. Lindahl, W. E., & Winship, C. (1992). Predictive models for annual fundraising and major gift fundraising. *Nonprofit Management and Leadership*, 3(1), 43-64.