05839 Interactive Data Science Final Report

Xinwen Liu (xinwenl), Xinyu Lin (xinyulin),
Yuxi Luo (yuxiluo), Shaobo Guan (shaobog)

December 2020

Project URL: https://share.streamlit.io/cmu-ids-2020/fp-vectorization/main

1 Abstract

With a Narrative track for this final project, our team will analyze the donation statistics from DonorsChoose.org, a funding platform for public school teachers to find resources in need, to make visualizations. Our essential question is: what is the current state of funding for public schools. We learn that public schools need more funding to provide quality education that will play an important role in students' learning outcomes. We would like to help users and potential donors understand the challenges that public schools face and what their needs are.

First, we will present a number of visualizations to provide useful information about the donation statistics over the past years, and help future donors to explore the options to make contributions for public education on their end. The visualizations will generally introduce two aspects of the data: Donation Cost Analysis for all donations, and Successful Rate Analysis for the project proposals hosted on the platform.

Second, based on our observation, a large number of projects are not fully-funded at the end. Therefore, we want to build a model to predict whether a project can be fully funded or not, to allow teachers to better understand what are some characteristics for projects to get fully-funded, and how they should refine their project proposals. In addition, we provide two interactive components to help users better understand our proposed model, in which they can explore the model's responses towards a given project description, and they can obtain useful feedback from the model to improve their own project proposals.

2 Introduction

What happens when the school lacks funding? Based on the data collected by the Virginia Department of Education, students have fewer choices in courses, less experienced instructors, lower test scores and college enrollment in high poverty schools compared to low poverty ones. Even worse, according to the Center for American Progress, many school districts are impacted by the Covid-19 pandemic and will expect a higher budget loss in the coming years. Thus, financial support becomes increasingly important at this moment. Even though the funding itself could not be the panacea, an increasing amount of funding will offer students more educational opportunities, better supplies, as well as mental services. Our goal for this project is: to use visualizations that we built to provide useful information about the donation statistics over the past years, and help future donors to explore the options to make contributions for public education on their end, and to help teachers refine their project proposals based on our model.

To relieve the stress and help more donors recognize the challenges public schools are currently facing, we use the datasets from DonorsChoose.org, a funding platform for public school teachers to find resources in need, to make visualizations. We would like to use our visualizations to present donors with useful information and attract more potential donors. We will present our visualizations in two aspects: Donation Cost Analysis for all donations, and Successful Rate Analysis for project proposals. Through these two aspects, we will discover some important findings about the donation statistics and provide users refined analysis on these specific visualizations.

To help teachers to better formulate their project proposals, we would like to utilize the dataset related to projects to understand the characteristics of projects that get fully-funded, and therefore develop a tool to interactively provide constructive feedback for teachers to better refine their project proposals. We will utilize Logistic Regression to analyze the factors that contribute to a project proposal's fully-funded status. In particular, we develop several features based on a project description to better understand what contributes most to a good project proposal which is fully-funded at the end.

In the Related Work section, we analyze related works and reports about donation statistics and the database. In the following sections of Methods, Results, and Discussion, we will divide our analysis into two parts: Visualizations and Model. In the Methods section, we describe the techniques and algorithms we used. In the Results section, we describe the overall results of our approaches. In the Discussion section, we analyze the effects of our approaches. At the end, in the Future Work section, we explore the areas of improvement and provide ideas to extend our current solutions.

3 Related Work

3.1 Kaggle

The dataset we are using for this project is from DonorsChoose.org through the Kaggle platform, which covers donation statistics from 2013 to 2018. The original motivation for DonorsChoose.org to release this dataset is to help them build targeted email campaigns recommending specific classroom requests to prior donors. Consequently, a number of data analysis is done to address the donor's related information for this particular dataset on Kaggle, to help them better decide what kind of donors they should target for their promotions. For example, DonorChoose: Complete EDA + Time Series Analysis provides detailed analysis on donors' information with respect to their location and the amounts of donation they made.

3.2 Resources

To offer our readers a deeper comprehension of the visualization, we reviewed some literature and connected the resources with our analysis. Some reports and conclusions were added in the background in order to highlight the difficulty public schools face when lacking funding (Duncombe, 2017). We also did some research on the funding section, where we explored the data revealing the top states with the most state funding for public students (US.News, 2020). In addition, we want to uncover the reasons behind some of the changes in our visualization. For instance, we discovered that the need for technologies surged from years to years, and we searched online for the reasons behind it and recognized the rapid development of educational technology in current days (Chen, 2019). We also met challenges when dealing with the dataset, like the meaning of free lunch percentage. We did not recognize its importance until we read the Income Eligibility Guidelines, which indicates that only the household income of a family reaches or is below the poverty level, can their children receive reduced or free lunch at school (2020). These evidence we collected support our analysis and build better understanding for our readers.

4 Methods

4.1 Visualization

The major technical challenge for the visualization part is handling the large volume of data. There are more than a million records in the project dataset, and each of them contains a paragraph of project description. Even after cleaning and compressing the raw dataset, the size of the dataset is still at 100MB scale, which a Streamlit server cannot handle decently with visualizations made by Altair. Therefore, the raw data need to be further preprocessed. Specifically, we firstly designed what kinds of visualizations are worthy of making. Then, we use Pandas to aggregate the original dataset to get the expected information for a visualization. Indeed, rather than sharing a common dataset, each visualization now has its own source data file, which is around 20KB, enabling the Streamlist server to show visualizations without huge loading time.

Another challenge for the visualization part is deciding to make what kind of visualizations. Despite the fact that the donation dataset has comprehensive information, using all information to make visualizations would overwhelm the audience and potentially the core ideas are hard to catch. Hence, we need a sensible way to choose truly important information to make visualizations, and an effective way to present the visualizations. On the one hand, we stick tightly with the purposes for our project mentioned in the proposal when choosing topics of visualizations. By doing so we could make sure that we would not make useless tables. On the other hand, we separate the visualizations in two major categories, and each visualization lies in a chapter with a concise title. By doing so, the audience could understand the relations among different visualizations, and form a general idea about the big picture of this section.

4.2 Model

In order to help teachers predict whether their projects can get fully funded within the designated time period, and to help them formulate better project proposals, we built a model that uses project proposal information to predict fully funded status. We first used spark and pandas to do feature engineering. We extracted 85 high frequency words in project descriptions, and computed the bag of words features. Other features such as length of the description, and project cost are also included. A full list of features can be found on our streamlit page. We split our dataset into training (80%) and test (20%) sets. Our dataset is skewed towards fully funded (78%), so we undersampled the fully funded class to make the two classes balanced in the training set. We trained a Logistic Regression model because it is easy to interpret and it works well with bag of words. F1-score was used to evaluate the model.

To present our model to the audience, we use bar charts to visualize the weight of each feature. We then present two sample project proposals, and allow users to highlight the words that incentivize them to donate. Once the users indicate whether they want to donate or not, our model prediction is displayed. We show how the model reaches the prediction by highlighting the parts that encourage/discourage the model to predict fully funded in green/red. The darker the color, the more it influences the model. Moreover, users can input their own project proposals on our streamlit web page to get the model prediction. How the model makes the prediction is also shown by the highlighted words in project description. If the model predicts that the project cannot be fully funded, we suggest ten new words to include in the proposal to help users increase the probability of getting the fund.

5 Results

5.1 Visualization

The major goal of the visualization section is to demonstrate the current state of donation funding of public schools and give new donors references in making donations. And we achieve this goal via constructing visualizations about donations in different dimensions. Specifically, the first section of our visualizations focus on revealing important dimensions of donation data (e.g., project mean cost, average donation, and etc.) with respect to chronological or geographical changes. Moreover, the second section of the visualizations demonstrates donation request successful rate over several important dimensions. These visualizations pose direct but insightful results about how different factors affect project successful rates, which is probably the most important dimension both donation requesters and platforms care about. Indeed, our results, while providing comprehensive information about donation related statistics, also reflect several interesting insights about our topic, which would be discussed further in the next section.

5.2 Model

Our Logistic Regression model results in a 0.77 F1-score on the test set. Because our model is highly interpretable, it is also worth analyzing the model for weights of our selected features. A complete overview of the weights information can be found on our website. For the features calculated from the project description, our model learns the number of “?” and “!”, and the number of specific numbers present in the project description can contribute positively towards the fully-funded status of the project proposal. This indicates that teachers may include more engaging descriptions in their project descriptions to describe

the current situation and seek for more help from potential donors. In addition, for the selected 85 words, our model learns that not all of them contribute positively towards the project fully-funded status. For example, our model indicates that "technology" contributes negatively towards the fully-funded status. For the features calculated from resource categories, our model learns that resources that are under the "Supplies" and "Technology" categories may have little chance to get fully-funded. However, from our visualization method, those two categories present similar fully-funded rates as other categories. Therefore, our model's learned weights and predictions may not reflect the overall statistics.

6 Discussion

6.1 Visualization

Through our visualizations, donors are able to learn about the donation information and take our analysis as references for making donation decisions. In our donation costs analysis part, donors can understand how the project costs are distributed geographically. To be more specific, they could check the mean and sum of project costs in different states from 2013 to 2018. While this information may not directly affect their donation decision, it is helpful for them to know which states have the most needs regarding educational resources. And one interesting insight we found is that the five top states (Vermont, New York, New Jersey, Pennsylvania and Wyoming) who received the most funding per pupil match with the darker states in our map, except for Pennsylvania.

Besides geographical information, donors can also learn about classrooms' needs based on the costs of various categories. By scrolling the year bar, they are able to tell the changes of needs through time and starting from 2017, they will be aware of the added categories of supplies including instructional technology, lab experiments, and educational manipulatives like kits and games. After going through the visualizations, donors can get a better understanding of the reasons behind the new categories from our written analysis.

In the last section of donation cost analysis, donors can have a brief overview of how much money each state donated through both average chart and sum chart, which are strong references for them to make their own donation decisions.

Our second section successful rate analysis is showing donors the results of the projects – whether projects are fully funded or not. The visualization in this section not only gives donors the information of the amount projects they donated may become successful, but also helps them understand what kind of the projects are more likely to become fully funded, so that they could continue supporting the same type of projects or pay extra attention to the ones that are less likely to get fully funded.

By reading the successful rate under different school metro types section, readers will learn how free lunch policies work and how it may impact the successful rate of schools with distinct metro types. Donors may tend to donate to schools who need the resources most or who start the most projects. From the successful rate based on states map, an interesting finding is that northern states have generally higher successful rates compared to the southern states. At the end, the successful rate based on proposed cost intervals diagram could help both donors and teachers get useful information in what cost intervals have comparatively higher success rates.

6.2 Model

The visualization of our model weights allows the audience to understand what words in the project description contribute to the fully funded rate positively. For example, the words "team", "play", "world" have a positive impact. By inputting their own project proposals and getting suggestions on proposal writing, the audience can improve their proposals, increase the likelihood of getting fully funded, and set a reasonable expectation of the project outcome.

7 Future Work

In this project, we focused on one donation site - DonorsChoose. In the future, we can compare donation statistics between different donation sites. We expect to give suggestions to the audience on which

donation site to choose for their projects to maximize the fund they get, by analyzing the sites' similarities and differences on marketing campaigns, donation reports, project areas, and target donors. The donation sites can also benefit from this analysis and make improvements accordingly. Moreover, our product currently allows the audience to highlight the words that influence their decisions. We can record their highlights and use this new data to improve our current features. In this way, our product is used as a data collection tool to collect donors' thoughts for free and we can further improve our model based on the new data.

8 References

Chen, G. (2019). Technology in public schools. Public School Review.

Child nutrition programs: Income eligibility guidelines (2020). Federal Register.

Duncombe, C. (2017). Unequal opportunities: Fewer resources, worse outcomes for students in schools with concentrated poverty. The Commonwealth Institute.

Health and physical education. (2020). PA Department of Education.

Leins, C. (2020). States with the most equitable school funding. U.S. News.

Partelow, L., Yin, J. & Sargrad, S. (2020). Why K-12 education needs more federal stimulus funding. Center for American Progress.