Matthew Sparacio

Data Mining Final Project

12/8/22

# Introduction

This analysis has two major goals: the first is to identify exciting projects, and the second is to gain a better understanding of donors who donate to DonorsChoose. By being able to identify exciting projects early, the company hopes to get better outcomes out of campaigns, focus more on projects that are likely exciting, and help more students. To achieve this goal, we have built a predictive model. To better understand groups of donors, we will do a cluster analysis in which donors will be categorized into groups with similar characteristics. These groups will allow us to tailer more to different customer experiences and increase donations as a result.

# Data Quality

The data for identifying exciting projects is based on over 300,000 past projects from DonorsChoose with labels of if they were exciting. Additional information is given on the teacher and school requesting the funding, the subject area for the funding, and the cost of the project. Some of this data had many missing values, so any variable with over 20% missing values was removed. We filled in the other missing values with the median of the data.

The data for the grouping of donors is based on over 1 million records of individual donors. For each donor, there is information on the amount of the donation made, the method of donation, and whether the donation was made through an individual campaign page. Information about where the donor was from had to be removed because it was largely incomplete. There were also 8 rows of data that were incorrectly filled in which were also removed. Lastly, to address the outliers of the amount that some people donated, we added another to describe donors, making categories of 100 to 1000 dollars donated, and more than 1000 dollars donated. This kept the information that some people donated much more than average but does not drastically influence the groupings.

# Identifying Exciting Projects

To identify exciting projects, the final model we used was a Random Forest Model. This predictive model is very effective at predicting if a project will be exciting. A random forest works by combining the predictions of many worse, smaller decision trees into a bigger and better model.

The random forest we made has an accuracy of 0.92, but more importantly, because exciting projects only occur around 10% of the time, it is more important to consider the precision and recall of the model. The precision on the test set of the data was 0.65 and the recall was 0.71. This is a decent model performance. Although it could be higher, the model still performs quite well. Below is a matrix where these calculations come from and shows how many predictions the model got right vs wrong.

Chart, treemap chart

Description automatically generated Using this model on new projects, we can quite effectively predict if a project will be exciting. The two most important factors that the model identified as important for if a project will be exciting are the number of donors that were teacher referred, followed by the number of donors that were non teacher referred. This means that the number of donors donating to a project has a big impact on if the project will be exciting. The two next most important are the total price and the number of students reached. These factors had the greatest impact on the model and are important to focus on for the future.

# Understanding Donors

In order to better understand the donors, we created a set of clusters that categorize donors into groups. We created 4 clusters each with 100,000 to 300,000 donors in them. Below is a description of each of the clusters’ main attributes that differentiate them. They are only grouped based on their donations, there is no information about demographics.

Cluster 1: This cluster contained everyone that donated above $100. This group had a diverse range of payment methods, and they had the highest proportion of including optional support in their donations.

Cluster 2: This cluster contained half of the donors who donated $10 to $100 dollars. This group had a fairly wide range of payment methods, most often paying with a credit card. But this group never made their donation in cash.

Cluster 3: This cluster contained the other half of the donors who donated $10 to $100 dollars. This cluster almost only made their donation in cash, with a small percentage also using PayPal. This group also often included a corporate sponsored gift card and was almost never promo matched.

Cluster 4: This cluster contained the people who donated under $10. This group most often paid with cash and importantly, this group had the most teacher accounts in it.

# Conclusion

Using these two analytical methods, we now can predict if a new project will likely be exciting, and we now have a set of clusters that characterize groups of donors. In order to keep these analyses relevant, data should continue to be collected and recorded about projects and donors. Over time, the models will become outdated as trends for donations change. As such, these models should be periodically retrained with newer, more relevant data to keep up with this. Additionally, more metrics could potentially be useful for these models. As data capabilities increase, more detailed information can be collected which could also help to improve these models over time.