Report: funding school projects on DonorsChoose.org
Using Clustering to understand the projects

The online platform DonorsChoose.org helps fund school projects for students in need. The platform provides a space where donors can find school projects and contribute to them. Although all projects seek to fulfill some kind of educational need, not all projects meet their funding goal. The client for this project seeks to help those projects which are at the highest risk of not getting fully funded to intervene on those projects.

For this setting, a Machine Learning solution was developed to predict the 5% projects at the highest risk of not getting fully funded within 60 days

The objective of this report is to understand the projects that were submitted to the platform, as well as the projects our Machine Learning predicted as being at the highest risk of not getting fully funded.

Data

The whole data set spans Jan 1, 2012 to Dec 31, 2013. It has 124,976 entries. Each entry is a project. The data includes information at three levels: 1) school context (location, charter, focus subjects and areas, poverty level), 2) project (resource type, grade, total amount of money, number of students reached, date posted, date fully funded, etc.) and 3) teacher (prefix). 12,329 observations had to be dropped since they did not have a representative value for the variable "not fully funded within 60 days", i.e. the last date for the dataset did not have information about the projects for least 60 days since they were posted.

The prediction data set spans July 1 to October 30, 2013. It has 1,561 entries. The best model was trained on data from January 1, 2012 to May 1, 2013 and predicted projects posted between July 1 to October 30, 2013, and the prediction is made at the time of posting. The 1,561 entries represent the projects posted between January 1, 2012 to May 1, 2013 among the 5% with highest risk of not getting funded.

Results

Using KMeans Clustering Method, we calculated 4 clusters[1] for each of our samples.

<u>Cluster sizes</u>

For our whole data set, Cluster 0 had 98,592 projects, Cluster 3 had 2,740 projects, Cluster 1 had 188 projects, and Cluster 2 had 9 projects. We appreciate significant concentration in Cluster 0.

For our whole data set, Cluster 0 had 1,445 projects, Cluster 3 had 109 projects, Cluster 1 had 6 projects, and Cluster 2 had 1project. We also appreciate significant concentration in Cluster 0.

[1] K (3 clusters) was selected based on visualization of the clustering results. *Run_clustering.ipyb* contains detailed reports for 2, 3 and 4 clusters and the user can choose any other values for k.

Proportion of projects with some characteristics

Figure 1 shows that the whole data set, the clusters of projects were fairly similar in terms of being part of a school charter and being part of a school magnet, and some similarity for being eligible for a corporate program to match the money raised (except for Cluster 2), the gender of the teacher of the project (except for Cluster 2), and the actual data on being fully funded within 60 days (except for Cluster 2). In other words, in Cluster 2, the projects are generally not less eligible for matching program, less female teachers applied to them, and less fully funded.

In contrast, the same figure shows data for the projects predicted. A very interesting finding is Cluster 3 (109 projects). These projects were less part of a school charter, more generally part of a school magnet, and considerably less proposed by female teachers.

Poverty

Figure 2 shows that the full dataset of has 4 very similar clusters in terms of poverty of the school. The projects for Cluster 2 come from slightly less-disadvantaged contexts than those of the other Clusters. For the projects with the highest risk of not getting funded, we note that Clusters have a more uniform distribution of projects coming from low, moderate, high and highest poverty levels.

Students reached and total price per project

Figure 3 shows that Clusters for the whole data set have no pattern between the students reached and the price of the project. However, the Clusters of predicted projects have a clear differentiation in terms of price. Projects for Cluster 0 requested less money than the projects for Cluster 3. The six projects for Cluster 1 were most expensive than those in Clusters 0 and 3. Finally, Cluster 2 has the most expensive project, which might be the reason this deserted Cluster was formed.

Annex

Figure 1: Binary columns for all data set (right) and prediction data set (left)

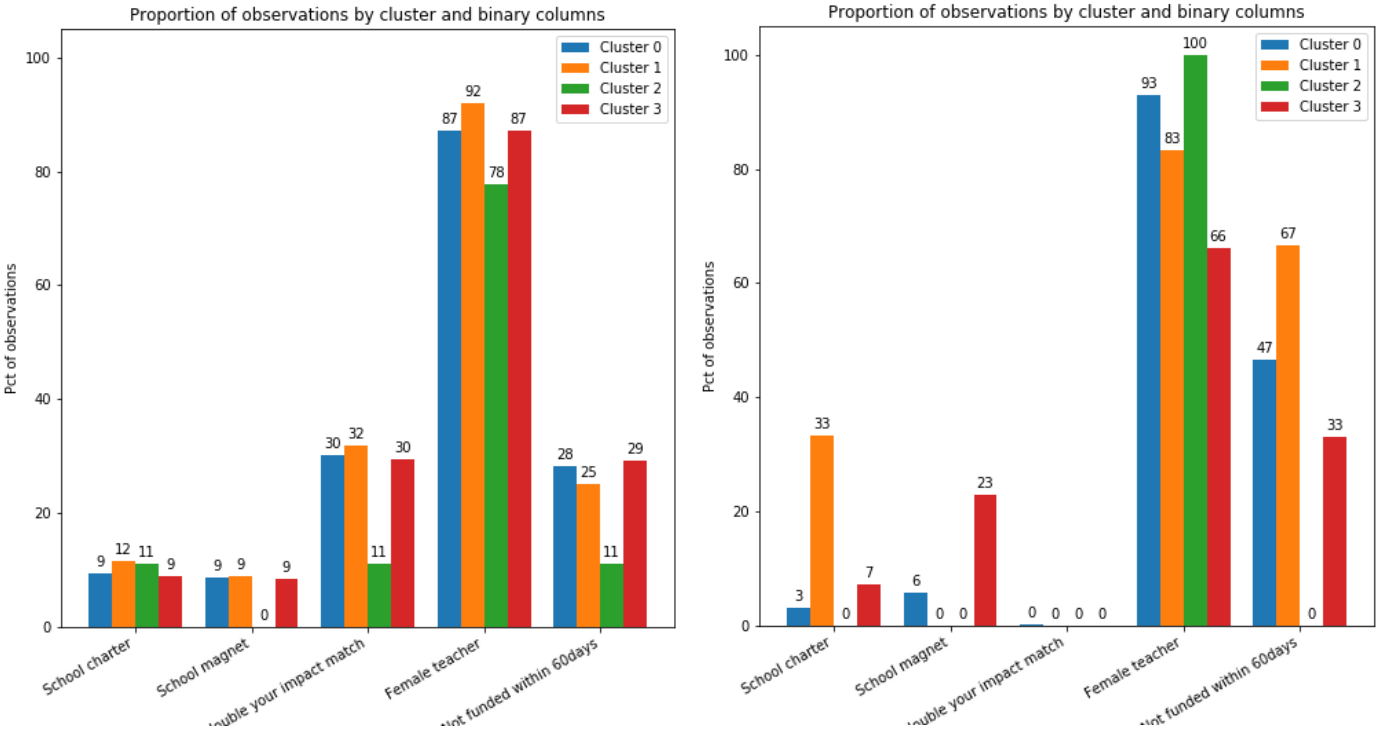

Figure 2: Poverty level by cluster for all data set (right) and prediction data set (left)

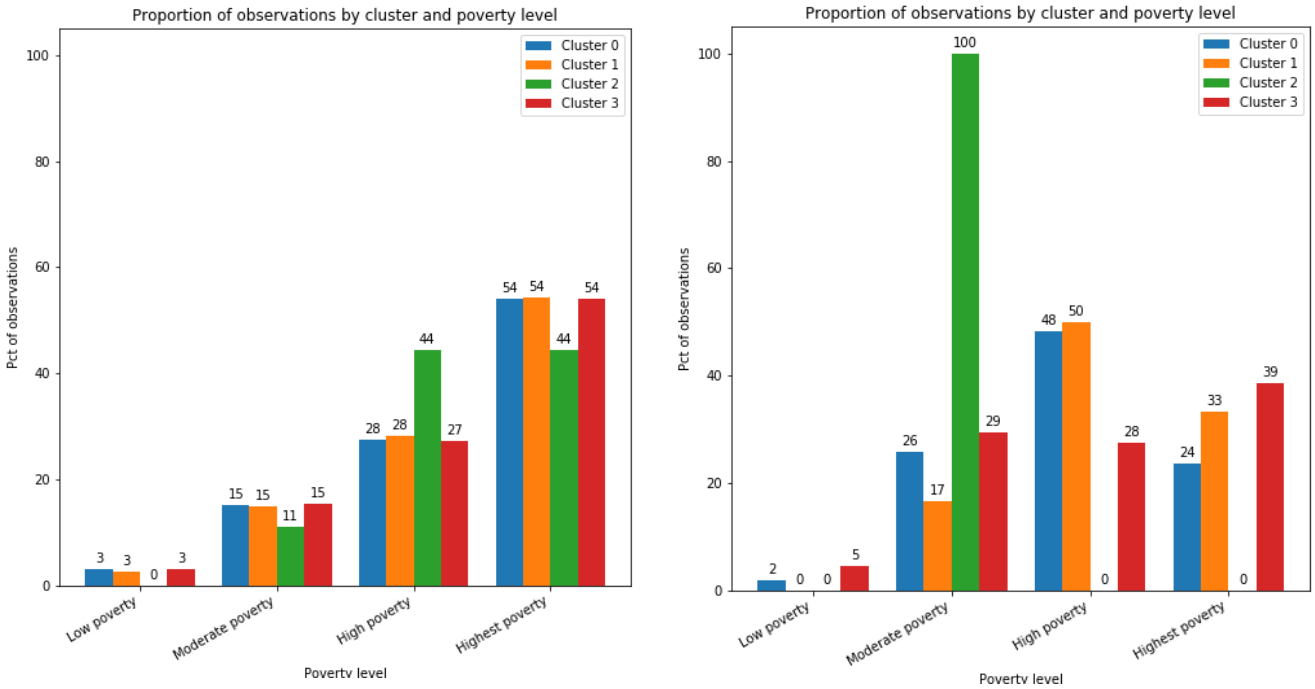

Figure 3: Log price of project versus log students reached for all data set (right) and prediction data set (left)