Finding the Clusters of the School Projects Funding Data Using K-means Clustering

We have the dataset that contains the school projects funding data in the year of 2012 and 2013, and the features include many key information such as school location, school categories (chartered school vs. non-chartered school), the poverty level of the area where the school is located, and total price of the projects, etc. Our task is finding the clusters within this dataset using unsupervised learning methods. We are interested in the clusters of the data on overall submitted projects, the projects that are not fully-funded within 60 days, and the top 5% of the predicted projects that are unlikely to be fully-funded within 60 days, which is predicted by the Logistic Regression Classifier model.

We set three clusters in each task, and here are the clusters that we discovered in the overall submitted projects. The first cluster (or cluster 0 in the Notebook) contains the projects with relatively **medium-to-high total price of support** that reach to **medium-to-large number of students**; also, most of the schools submitting these projects are in suburban and urban areas with high poverty level. The second cluster (or cluster 1 in the Notebook) has the projects with **low-to-medium total price of support** that reach to **low-to-medium number of students**, and most of the schools falling into this category are in suburban and urban areas with high poverty level. The third cluster (or cluster 3 in the Notebook) incorporates the projects with **medium-to-high total price of support** that reach to **low-to-medium number of students**; likewise, most of the schools are in suburban and urban areas with relatively high poverty level.

Regarding the data on projects not fully funded within 60 days, the first cluster has the projects with **medium-to-high total price** that reach to **low-to-medium number of students**, and the schools are in the areas with **the highest poverty level**. The second cluster has the projects with **medium-to-high total price** that reach to **medium number of students**, and the schools are in **high poverty level** areas. The third cluster has the projects with **medium-to-high total price** that reach **large number of students**, and the schools are in **high poverty level** areas.

For the data on the predicted top 5 percent, the first cluster has the projects that request **high**

total price and **reach to medium-to-large number of students**; and the schools are in **rural and suburban areas** with **moderate poverty level**. The second cluster has the projects that request **high total price**, and the schools are in **suburban and urban areas** with **high poverty level**. The third cluster has the projects that request **high total price**, and the schools are in **rural and suburban areas** with the **highest poverty level**.