Authors: Travis Lloyd & Kiran Singh

Abstract

In America, the educational system has been left fragmented and left to thrive or fail. With most of the educational funding stemming from taxpayer money, it is only logical that locales with less affluent constituents or a smaller population would inevitably have less income from taxes. This is directly reflected in that areas school districts, for better or worse. In this article we explored a School Improvement Grant and its applicants to better understand and predict the likelihood of an applicant being approved.

Business Background

Within the dataset there is significant demographic data that was found to have relevance in whether an educational institution has a chance of being approved. The SIG grant is such a powerful resource as it targets underprivileged educational institutions by adding funding to the school for improvements, not just structurally but in new textbooks, up-to-date technology like computers and so much more. Our aim was to bridge that gap and provide a great foundation to build upon by adding more grants and funding streams that schools can access and dedicate resources to application processes.

Problem Statement

With over 136,000 K-12 schools in the United States, one could logically deduce that there is a high variance in the quality of education based on location. With the rate that new technologies and textbooks are released, there can be a substantial gap in the information provided to children regarding what is most current and best quality.

Summary of the findings

The models evaluated included: K-Nearest Neighbors, logistic regression, decision trees, bagging classifier, adaboost, random forest, linear discriminant analysis, and neural network. Following the same ordering, the resultant F1 scores were as follows: 0.471, 0.496, 0.485, 0.644, 0.493, 0.473, and 0.498. While six of the seven models evaluated had similar performance, one was prominently superior to the others with a F1 score of 0.644, AdaBoost.

Business Questions

While we were able to find desirable results for this dataset, can this application be replicated with similar or better results for other grants? We also have to ask ourselves are there more grants with such rich datasets that we can perform an analysis on them.

Scope of analysis

The scope of this analysis is limited to primarily this one grant. While there are a lot of options in the ether of school funding, this dataset was rich and capable of giving great insights and a baseline to build a foundation on. We utilized training, testing, and validation splitting due to the dataset being unbalanced. Models used include: K-Nearest Neighbors, logistic regression, decision trees, bagging classifier, adaboost, random forest, linear discriminant analysis, and neural network. All models were trained and tested on balanced datasets, then validated using the validation holdout set that was unbalanced.

Limitations

As previously mentioned, this analysis only includes one grant and its relevant data. Expanding to other grants and their relative datasets could provide better or even less favorable results.

Solution details

This article has provided insight into the gap between financially disadvantaged districts and wealthy ones by introducing a method for educational institutions to locate and acquire supplemental finance through secondary grants and aid.

Concluding summary

Accurately predicting and reducing false-positive cases is the foremost priority of identifying private donors, which may help those schools that fall within low-income districts compete with wealthy schools to reduce disparities in educational outcomes. The performance of machine learning methods varies for each individual business case. The type of input data is a dominant factor that drives different ML methods. The number of features, number of transactions, and correlation between the features are essential factors in determining the model’s performance. Comparing all algorithm performances side to side, the adaboost model was prominently superior to the others with a F1 score of 0.644, which is a measure of a model's accuracy on a dataset.

Call to action (CTA)

Allocating and aggregating additional grants will further expand the capabilities and reach of this model. The eventual goal is to have a recommender system in place based on the school-provided demographic information. Reaching into smaller communities as well will also be a focus of further expansion of this project as the SIG Grant’s distribution is more to urban locations.