A Machine Learning Approach to Assess Education Policies in Brazil

Alexandre Simoes Gomes Junior – asimoes@stanford.edu

1. Introduction

With the recent popularization of Machine Learning and Statistical tools, many private companies already use these technologies in their management and operation processes. Even in developing countries, as Brazil, there are already many different companies, including national ones, taking advantage of these technologies in several ways. However, innovation always take longer to reach the government. Especially in countries with corrupt, inefficient and incompetent governments as is also the case of Brazil.

Recently, some organizations in the Brazilian government are trying to change this scenario. For example, the State Government of Sao Paulo adopted in 2012 a new approach to define the budget of the following years. The program is called Budget by Performance and the framework can be applied to multiple areas. The general idea is that, for any area in the public sector, it is possible to quantify the goal, based on the demands of the population. Given the goal, it is possible to derive what services and products needs to be offered by the government to the population and what are the resources needed for that, including human labor, infrastructure and capital.

In this context, the present project offers a new approach in the estimation of the budget for the Secretary of Education of the State of Sao Paulo. The idea, in accordance with the new framework adopted, is to start with a goal. In this case, it is a Quality Index called IDEB (Development Index of Basic Education), which is based on the scores of students in a national exam called Prova Brasil and on specific approval data from schools. Each school in the state has its own IDEB score, which is updated every two years, and its IDEB goal, defined for the next 4 years.

The spending data in the education sector represents the policies implemented by the government. This includes the segmentation of spending across several categories. The idea is to identify the minimum necessary budget to each category that will allow the school to achieve its goal.

Finally, since different schools might have different needs, a cluster algorithm having as input descriptive data on the schools is used to separate them in different groups with similar needs. The main assumption is that schools in the same cluster have similar problems and, therefore, need an equal distribution of spending across the categories in order to improve and achieve its goal. The final problem consists of finding the minimum total budget and its optimal categorical distribution for each cluster of schools.

2. Related Work

There are several studies evaluating the impact of increase in government spending for specific sectors in quality indexes, for example the cited works by Baldacci et al, Sutherland et al and Gupta et al. In the first example, different variations of the least mean squared error regression model are tested, as well as a covariance structure model based on latent variables. The models are used to find the relation between public spending and quality indexes. In all the studies there was statistical evidence of a correlation between increase in spending and increase in the indexes. The present project aims to explore this relation to create a tool for budget planning.

3. Dataset and Features

The target variable, as explained previously, is the IDEB score of each school. This data can be found on the page http://ideb.inep.gov.br/. It contains the evolution of the in Brazilian public schools, for the years of 2013, 2015 and 2017. This index combines the scores of students in a national mandatory exam with data provided by each school describing rate approvals, to assess the quality of basic education in public schools. In the year of 2017, the goals for the 2019 and 2021 were established.

The second data source is the Brazilian school census of 2013, http://portal.inep.gov.br/microdados, that has survey data on every public school in Brazil. This dataset contains information describing many aspects of the

infrastructure of the school, the qualification of teachers and the profile of students. Some examples of features include:

- Total number of students
- Number of professors by level and area of education
- Number of laboratories, computers and offices
- Number of students per race
- Number of classes per subject
- Total amount of time spent by students with extracurricular activities

In total, after the preparation of data, the dataset includes 353 features per school. Most of them are count variables, as the number of professors from each educational background, number of different equipment, etc. A big part of transforming the data included counting different categories in categorical variables. For example, there is one entry in the original data for each student and teacher in the school, in the final dataset there is only counts for the number of male/female students, mathematics/biology/chemistry teachers, etc.

The last data source is the website https://www.fazenda.sp.gov.br. It has data detailing all of the disbursement made by the state government of Sao Paulo since 2010. The database contains information on the targeted sector (education, health, transportation, etc.), the subarea (primary, secondary, higher education, etc.) as well as a more detailed classification of the purpose of the spending (scholarship for poor students, construction of new schools, purchase of food or transportation for students, etc.). Some of the expenditure categories include:

- Transportation for students
- Food for students and workers
- Constructions and maintenance of schools
- Salaries of school employees
- School supplies
- Contracts with third parties

After transforming the data, it ended up with 711 different categories. The final variables consist of the summed expenditure in each category for the years of 2014, 2015, 2016 and 2017. There are, however, two main limitations with this dataset. First, many categories are redundant, for example, there are approximately 5 different categories related to constructions in schools. Second, the source website only offers spending data per local administration center. There are 1382 different centers and 3152 schools. To get to spending per school the expenditure of each center was divided by the number of schools it attends.

4. Methodology

The final project includes three different models that process data in different phases. First, a regression model uses the descriptive data from the school census (353 features) to predict the Ideb of each school (3190) in 2013. The purpose of this model is to reduce the number of features from the census considered in the next phase. Only the most important variables detected by the algorithm in this phase continue in the dataset. The final set of variables have an accumulated feature importance of 0.99 in the model.

In this phase, 3 different algorithms are tested, SVM, Gradient Boosted Trees (GBT) and Ridge Regression. For the GBT, two different implementations are evaluated, scikit-learn and LightGBM. The evaluation metric chosen is the R^2, defined as:

$$R^2 = \frac{\sum_i (y_i - h_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The model chosen is the one that presents greatest R^2 in the test set (638 schools). After the selection of variables, 129 features continued to the next phase, the clustering algorithm. This second model uses the descriptive variables to separate schools in groups with similar needs. Since the final goal of the project is to define the budget and its optimal distribution for each school, there is the need to isolate the effect of other variables, not related to expenditure, that are correlated to the Ideb. This is the purpose of the second stage in the data processing framework.

The evaluation metric for the clusters is the mean Silhouette Coefficient. For one sample in the train set, the Silhouette is given by:

$$s = \frac{b - a}{\max(a, b)}$$

Where *a* is the mean intra-cluster Euclidean distance to the considered point and *b* is the Euclidean distance to the nearest point in other cluster. Two different approaches are tested, both of them use K-means as the main algorithm. In one of them, however, the original data is first transformed with Principal Component Analysis (PCA) in order to reduce the dimensionality of the dataset. The model used data from 9837 schools, this phase did not consider only schools administered by the state government but also those ran by the federal and city governments.

The final phase is a combination of multiple classifiers, one for each cluster. Each model predicts whether the school achieved its goal for the 2017 Ideb. The input variables are the expenditure data for each school, there are 3152 schools and 711 features. The assumption in this phase is that, after isolating the effects of descriptive variables in the Ideb, it is possible to find an expenditure distribution that will minimize the total sum of investments per school while allowing it to achieve its goal. This distribution will be equal for all schools in the same cluster.

The evaluation metric is the F1-score, that combines both precision and recall, in order to guarantee that the model do not present good performance only for the most common class. The F1-score is given by:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

In this phase only one algorithm was implemented, derived from the first part, the GBT implementation in scikit-learn. The final tool can be applied in the estimation of the budget for each school.

In the chosen approach, first, all the schools that achieved their goals and for which the model presented correct predictions are selected. The initial budget estimate for each category is the minimum value (greater than 0, if there is one) found for that category in this group of schools. If the model predicts success in goal achievement with this expenditure distribution, it is considered as the final budget.

If the model predicts fail in goal achievement, one of the categories is chosen to be increased. The probability of selecting a specific category is equal to the normalized feature importance of the variable that represents this category, according to the final model. The budget for the selected category than assumes the value of the second lowest expenditure for this category in the selected subset of schools. This process is repeated until the model predicts success in goal achievement. If a specific category achieves its maximum possible value its probability of being selected in the following iterations goes to 0.

5. Results and Discussion

5.1 Regression

The results for the each model tested is in table 1. The chosen model had a final test R^2 of 0.647.

Model	Train R^2	Test R^2
GBT – scikit-learn	0.745	0.647
SVM	0.069	0
Ridge Regression	0.674	0.575
GBT - LightGBM	0.812	0.550

5.2 Clustering

The 129 most important features in the previous model are then used in the clustering algorithm to separate schools in groups. The results of the two models tested are in table 2. The number of clusters varied from 4 to 20 and in the final model consisted of 10 clusters, in a tradeoff between increase in the silhouette and guaranteeing a reasonable number of schools in each cluster. Still, some clusters ended up with few schools, to the minimum of one. These consist of outliers and these clusters did not enter in the next phase.

Model	Silhouette
K-means	0.767
PCA + K-means	0.805

Although the Silhouette for the model with PCA preprocessing was higher, both algorithms presented a very similar result, with clusters almost identical. The output of both cases were tested in the classification phase and the clusters from the model with PCA presented a better weighted average F1-score for the classifiers. For this reason, this was the selected model.

5.3 Classification

Finally, for each cluster with at least 20 schools, one classification model was created to predict goal achievement. The results are shown in table 3. The final weighted (regarding number of schools) average F1-score of the classifiers was 0.692. The last row in table three shows the result of the model when there is no separation of schools according to clusters.

Cluster	Train F1	Test F1	# Schools
0	0.671	0.631	1791
2	0.903	0.774	141
3	0.864	0.830	399
5	0.824	0.625	38
6	0.827	0.751	759
All	0.694	0.675	3152

As expected, by separating schools into clusters the performance of the classifiers increase. This means that it is easier for the model to find patterns in expenditure data when schools with similar descriptive features are grouped together and isolated from other groups. This supports the initial hypothesis.

However, it is also possible to observe that some clusters, as number 0, presented a test F1-score lower than the model with all schools together. This might be an indication that this cluster is not homogeneous in terms of characteristics that might affect the Ideb. In addition, clusters as number 5, had problems with overfitting due to the small sample of schools it represents.

Cluster number 3 had excellent performance, which indicates that this cluster is homogeneous and that it is possible to find a common expenditure distribution for these schools that will allow them to achieve their goals. For this cluster the method of budget estimation described previously was implemented. The prediction for the initial budget estimation (minimum values for each category) was 1, therefore there was no need to iteratively search for the expenditure distribution.

Applying the minimum estimated budget for each school, there is a reduction of R$314,972,841.00 in the total spending of the government of Sao Paulo with the schools in cluster 3. In addition, according to the model, all schools would have achieved their goals using the estimated expenditure distribution, while with the current budget approximately 30% of schools in this cluster did not achieve their goals.

6. Conclusion and Future Work

The usefulness of the tool developed in this study depends heavily on the quality of the clustering algorithm. For the initial assumption about the clusters to hold, all the relevant factors associated with the schools that do not relate to their spending and that affect the Ideb, must be represented in the clustering variables. When this is the case, the separation of schools in groups will be able to isolate the effect of this variables and all variation observed in the Ideb will be explained solely by differences in the expenditure distribution.

In the present project, the assumption was valid for some clusters, mainly number 3. For this cluster, it was possible to create a good predictor of goal achievement only with expenditure features. When this condition is present, this tool can be very useful in minimizing the budget of the schools while guaranteeing they will achieve their goals.

However, other clusters, mainly number 0, are too heterogeneous to have the Ideb explained only with spending data. It means that, to create a good predictor for goal achievement more variables are needed. Therefore, for this clusters, it is not possible to explain goal achievement only as a function of spending distribution.

To solve the problem described above, the first step would be to incorporate new variables in the clustering phase. For example, sociodemographic variables of the region where the school is located are probably highly correlated with its Ideb also. Features as the average income of residents, average number of people in one house and distance from the center of the city are not present in the school census data used in the first two phases of the project.

The Brazilian Census have this type of sociodemographic data. However, in this dataset, locations are described as sectors and each sector has its own code. The problem when linking this dataset to the school Census is that the last one does not have information on the code of the sector where the school is. This needs to be solved in order to include data from the Brazilian Census in the clustering algorithm.

A second point of improvement is aggregating redundant categories of spending. This problem was detailed previously. Because of these redundancies, it might be difficult for the last model to identify the real impact of each subarea of investment on the Ideb.

Other limitation also explained previously is that the data provided by the government of Sao Paulo does not have detailed spending for each school. There certainly is, in the government database, this type of data, however it is not open to the public.

Finally, this project did not explicitly try to find a causal relation between the input features and the target variable, which is a necessary step in the design of public policies. A qualitative evaluation of the importance of the features in the first and third models, as well as the impact they have on the target variable needs to be conducted. This would be better performed with the assistance of specialists in the area.

This tool, however, is a good starting point for the government to explore quantitative tools in the design of public policies. In a real implementation, there would be an evaluation period when the tool would suggest the expenditure distribution and, after its implementation, the results would be reevaluated and incorporated in the model.

7. References

[1] - Bair, Eric. "*Semi-supervised clustering methods*." Wiley Interdisciplinary Reviews: Computational Statistics 5.5 (2013): 349-361.

[2] – State Government of Sao Paulo. “Orcamento por Resultados no Estado de Sao Paulo: experiências, desafios e perspectivas”. http://www.ppa.sp.gov.br/docs/OpR.pdf (accessed December 14, 2018)

[3] - Gupta, Sanjeev, Marijn Verhoeven, and Erwin R. Tiongson. "*The effectiveness of government spending on education and health care in developing and transition economies.*" European Journal of Political Economy 18.4 (2002): 717-737.

[4] - Sutherland, Douglas, et al. "*Performance indicators for public spending efficiency in primary and secondary education.*" (2007).

[5] - Baldacci, Emanuele, Maria Teresa Guin-Siu, and Luiz De Mello. "*More on the effectiveness of public spending on health care and education: a covariance structure model.*" Journal of International Development: The Journal of the Development Studies Association 15.6 (2003): 709-725.

Python libraries – scikit-learn, LightGBM