A Machine Learning Approach to Assess Education Policies in Brazil

Alexandre Simoes Gomes Junior – asimoes@stanford.edu

Overview

- Goals:
  1. Estimate the budget for government spendings in public education necessary to achieve pre estipulated goals (quantitative approach)
  2. Identify subareas that should be prioritized by the government when deciding where to allocate the amount defined by the budget (quantitative + qualitative approach)
- Models and Results:
  1. Regression model to predict current quality index of schools (only descriptive features): Gradient Boosted Trees (GBT). R^2(test) = 0.647
  2. Clustering model to identify groups of school with similar profiles: K-means. Silhouette = 0.805
  3. Classification model to predict goal achievement in schools (only spending data): GBT. F1-score(test mean) = 0.692
- Main assumption: it is possible to cluster schools according to their descriptive features . For each cluster there is an optimal distribution of spendings that will allow the school to achieve its goal

Data and Features

1. Government spending data: detailed anual information of all government spendings in public education in the state of Sao Paulo. Source: State Government of Sao Paulo. Example of features:
   - Spendings with transportation for students
   - Spendings with food for students and workers
   - Spendings with constructions and maintenance of schools
   - Spendings with salaries of school employees
2. School census: descriptive data on each school in the state of Sao Paulo. Survey conducted in 2013. Source: Inep (National Institute of Educational Studies)
   - Number of students
   - Number of professors separated by level of education
   - Number of laboratories, computers and offices
3. Education quality index: Ideb (Development Index of Basic Education) score for each school in the state of São Paulo. Source: Inep
   - School performance according to Ideb in 2013, 2015 and 2017
   - School goal for Ideb in 2013, 2015 and 2017

Regression

- Input variables: descriptive data from school census (290 features)
- Target feature: Ideb score of 2013
- Purpose: detect the descriptive features most correlated with the quality index
- Evaluation metric: R^2

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

Where h is the output of the model

- Results:

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

Clustering

- Input variables: most importante variables from school census (129 features). Accumulated feature importance in the regression model of 0.99
- Purpose: separate schools into groups according to characteristics relevant to the Ideb
- Evaluation metric: mean Silhouette Coefficient of samples. For one sample 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.

- Results:

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

- The final model used 10 clusters

References

- Bair E., Semi-supervised clustering methods, 2013
- CS229 Machine Learning lecture notes, 2018

Classification

- Input variables: spending data (711 categories) for each school
- Target feature: 1 if school achieved goal in 2017, 0 otherwise
- Purpose: this tool can be used to estimate the necessary budget for each school given the Ideb goal. It is also useful to evaluate how the spending should be distributed among diferent categories. One model was constructed for each cluster in order to isolate the effect of the school descriptive variables in the Ideb.
- Model: GBT – scikit-learn
- Evaluation metric: F1-score

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

- Results:

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

Discussion

- As expected, the inclusion of information on the cluster of the school improved the performance of the classifiers, except for clusters 0 and 5
- The most relevant spending features for the classifiers changed for diferent clusters. Example:
  - Only for cluster 2, financial assistance for students appeared among the 10 most important features (third). For schools in this cluster, increasing this type of assistance for students and their families might have a big impact on their education

Future Work

- Inclusion of sociodemographic data of school location
- Evaluation of the causal relation between spending in different categories and the Ideb result
- Grouping of categories in spending data to avoid redundancy
- Exploration of better approaches to identify spendings per school (spending data is associated with regional administration centers, not schools)