Where to aim? Factors that influence the performance of Brazilian secondary schools

Paulo J.L. Adeodato
Universidade Federal de Pernambuco
Centro de Informática
Recife - Brazil
pjla@cin.ufpe.br

Rogério L. C. Silva Filho
Universidade Federal de Pernambuco
Centro de Informática
Recife - Brazil
rlcsf@cin.ufpe.br

ABSTRACT

There have been discussions on where to invest the budget allocated to education. Most politicians want to invest in the schools' infrastructure, but is that the most efficient policy for spending? This paper presents analyses to help clarify that. It integrates the most recent data sources (2018) on the secondary students' assessment (ENEM), the School Census and the Teachers' Census and consolidates all microdata to the school level, making them features of the schools. These features are then grouped into three types: infrastructure, human education and socio-economic aspects. Then the features from each group are applied in logistic regression predictive models both isolate and collectively. In a 10-fold cross-validation comparison with the area under the ROC curve as metric. The experimental results show that infrastructure is significantly less influential than the other features. Further research needs to consider investment costs and time to produce effect on school performance.

Keywords

School quality assessment, Educational decision support system, Educational Data Mining, Domain-Driven Data Mining, Educational budget allocation

1. INTRODUCTION

International comparison of students' performance among countries by the Programme for International Student Assessment (PISA) has yielded strategic discussions in international education policies. The PISA, sponsored by Economic Co-operation and Development (OCDE), aims at assessing and providing a global perspective on secondary education (15-year-old pupils) across countries of the world [16].

Following the international efforts, the local governments have been concerned with standardized tests themselves, aiming at the assessment of students as much as at monitoring the quality of the educational system [1]. In Brazil, the National Institute for Educational Studies (Instituto Nacional de Estudos e Pesquisas Educacionais – INEP) produces the annual School Census which is a survey of the schools for secondary education in the country and the National Secondary School Exam (Exame Nacional do Ensino Médio – ENEM) that evaluates student performance at end of secondary education.

In 2009, ENEM became a mechanism for students' admission to higher education in public universities. That improved the quality of the information collected. Added to the technical knowledge of each student, ENEM also captures their socio-economic-cultural (SEC) information [2]. The integration of this information with the School Census data has become a relevant source of data for scientific studies and enables the Federal Government to define and validate public policies for Brazilian education [30]. However, secondary education is under jurisdiction of the constituent states of the federation, not the national government. Thus, despite the importance of the federal government role, there is considerable variation among the states in curriculum, teacher training, budget policies and other issues [10].

Many factors can influence the performance of the students. Studies have shown that school inputs, students' SEC background, parents' education are correlated with student achievement [13, 7, 12]. In Brazil, according to the last school census available for this research (2018), 42% of the secondary public schools still lack Basic Infrastructure Level. The definition of the quality of the levels was performed by Neto [26] being the Basic Level the second lowest of four levels which includes features like having a management room with computer and printer for administrative work only. This scenario makes Brazilian politicians focus most of their educational bills and budget allocations on improving school infrastructure [19].

Despite the importance of providing infrastructure to schools, it is common for politicians to invest in infrastructure such as computer labs, tablets, TVs etc. even for schools that have not reached the basic level yet. This paper does not discuss pedagogical issues related to infrastructure; just tries to help policymakers and education-related institutions on how to invest their budget to comply with both regulations and education quality goals in a long-run plan to secondary public education.

This paper presents experiments in a Domain-Driven Data Mining (DM) approach that assesses the quality of secondary

Paulo Jorge L. Adeodato and Rogerio Luiz C. S. Filho "Where to aim? Factors that influence the performance of Brazilian secondary schools" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020),* Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 545 - 549

schools in Brazil. The results show that the infrastructure predictors are less relevant than SEC and educational predictors at a 5% significance level. Experiments were carried out on the most recent data available (2018) with logistic regression models in a 10-fold cross-validation setting.

The paper is organized in 4 more sections. Section 2 presents the data sources and preprocessing. Section 3 describes the experimental project to elucidate the most influential group of features for predicting good students. Section 4 presents the results and discusses its impacts. And Section 5 presents the conclusions, difficulties found and suggestions for future work.

2. DATA SOURCES AND PREPROCESSING

This research has used two official public databases: Microdata from the National Secondary School Exam 2017 and 2018 containing the students SEC information and their grades on the test at the end of secondary education, and the School Census 2018 [2] detailing the conditions of the schools, from physical infrastructure to faculty information. The 2017 ENEM database was only used as an independent statistical sample to apply the process of granularity transformation described in Subsection 2.4. These databases refer to over 5 millions of students of 32,000 secondary schools across the country, but this paper will focus only on public (free) schools.

2.1 The Universe of Schools (Scope)

This research attempts to help policymakers optimize the budget allocation in order to improve Brazilian Secondary Schools. That does override the priority of the 42% of public secondary that have only an elementary infrastructure (just classrooms, electric energy, sanitation and piped water). A few schools (0.7%) were discarded from the database for being below that level.

ENEM is a democratic exam that any person can sit. That makes it necessary to apply some selection filters in student grain: a) Students who have no school assigned, are just training or are not in the last secondary school year (74%), b) students who do not follow a regular curriculum (2.5%) and c) foreign students (0.02%). The remaining 680,583 students were considered. To eliminate anomalies that could either divert from the goal or deteriorate the quality of the work, students who did not perform all the tests, including the essay, were also left out of the scope of this research.

Back to the school grain, for having critical mass, only schools with 10 or more students were selected, as established by INEP in the analyses. After this last filter, the total that remained in this research dropped to 14,579 secondary schools with 653,848 students which form the dataset used in this paper's experiments.

2.2 Problem Characterization and Goal Setting

In business, one of the most common decision strategies for selecting the eligible candidates for an action is ranking them according to a classification score and choosing those above a predefined threshold [17]. That is used in applications such as staff selection, fraud detection [6], and resources allocation in public policies, for instance. This score is computed by either weighing a set of variables based on human-defined parameters or by applying a function learned by a classification algorithm from a set of data with binary labels as desired response, according to specific optimization criteria.

In some domains of application, several problems are ill-defined simply because stakeholders do not reach consensus on either method [29]. That is particularly true for education where experts and faculty do not agree even on the characterization of a good school or a good student. To circumvent these issues, we have adopted the systematic approach proposed by [3] to characterize this as a binary decision problem. Thus, the problem can be solved by machine learning algorithms based on the supervised learning paradigm with a data dependent strategy where each example is labeled as "good" or "bad" for binary decision making. That involves solving two scientific issues which represent controversial points in the application domain: (1) which metrics should be used as a ranking score for evaluating the quality of the school and (2) which threshold should be adopted as a criterion to define what would be a "good" school in the binary decision.

The ENEM [1] has been conceived to assess the quality of the Brazilian secondary schools based on their students' evaluation on the test. Despite arguments among experts on education, they have agreed that the performance of the students at the last year would represent their performance in the secondary school and also agreed that the mean student score would be the most relevant indicator of each school, as already done in previous studies [3, 30].

2.3 Binary Goal Definition

Once having defined the quality metrics, the most controversial point is to set the threshold to characterize what would be a "good" or "bad" school in the dichotomic objective. Once again, to circumvent the controversy and lack of consensus in the field on the issue and bring a higher level of abstraction that enables future comparison across years, regardless of the degree of difficulty of the exams, this study used statistics concepts for setting the threshold as recommended by [28]. Quartiles of the distributions not only are robust against extreme values (outliers) [21], but also can be a straightforward data dependent dichotomizing criterion of interest for the application domain. The upper quartile has already been successfully used as threshold [28] on a continuous goal variable for creating a binary target-variable. This paper has adopted that approach for converting the problem into a binary classification where the upper quartile represents the "good" schools.

2.4 Granularity Transformation

The granularity of the attributes is a fundamental concept and its diversity brings great complexity to research of this nature. How can one associate to each school its family income attribute from the distribution of family income of their students? How can one associate to each school its faculty education attribute from the distribution of faculty education of their teachers? These transformations represent a difficulty for teams without professionals specialized in developing data mining projects. This difficulty is due to both the sheer volume of data to be handled and the need to use artificial intelligence to embed knowledge of experts in education in the transformation of the attributes for granularity change in a process coined Domain-Driven Data Mining (D3M)[23].

We considered and chose the Regression Granularity Transform (RGT) [4] as the most adequate approach for this research. It aims at maximizing the information gain towards the target class for categorical microdata present in Student and Teacher grains. Logistic Regression was the technique applied on the categorical attribute distribution having its histogram with the categories' relative frequencies as input. These transformations were learned from the previous year data (2017), to avoid having to discard data from the focus year of 2018. For numerical features, the average was the transformation adopted.

2.5 Preprocessing

Many factors affect the success of a data mining application. Data quality is among them [20]. Domain and data understanding allowed for the removal of irrelevant attributes (e.g. linked to elementary and fundamental schools and to other secondary school models that do not use the regular curriculum), attributes with a posteriori information and identification codes.

In the final data sample, just two binary features presented missing values, which were filled with "0", because they represented lack of that property. The categorical features that had the mode representing over 90% of the cases were removed.

For features with correlation higher than 0.8, only those with the highest semantic value for the domain were preserved. To reduce the influence of outliers and improve the quality of the Logistic Regression models, all numerical features were normalized using the α-winsorized values of the distribution ($\alpha/2 = 0.025$ at each tail) as their minimum and maximum.

3. EXPERIMENTAL PROJECT

The experiments were carried out using the Logistic Regression model in a 10-fold cross-validation setting. The features on school grain were partitioned into 3 different groups: 1) Infrastructure of schools, 2) SEC information of Students and 3) Level of Education of Parents and Teachers. The same held-out fold was used as test dataset for all groups and the models' performance on it was assessed by Area Under the Receiver Operating Characteristic (ROC) curve. ROC curve plots the true positive rate against the false positive rate, at all possible decision thresholds.

The goal is to experimentally compare the discriminant power of each group of predictors, focusing on groups 1 and 3, once that it is hard to produce any change in group 2 with educational policies. In one hand, it is widely known that features from group 3 are more influential than those of group 1 , but it is hard for education policymakers to intervene on that due to limitations on either the country's economy or the Cash Transfer policies [22]. On the other hand, investment in features from group 1 has been the main focus of government, either by the insufficient conditions in some schools or because these investments have their effect more easily assessed. There are some studies in the literature on public policies addressing teachers training and parents education [5, 8]. This paper aims at showing that the predictors of group 3 are more influential in predicting performance than those of group 1.

3.1 Performance analysis

Figure 1 shows the results for each test set in the 10-fold cross-validation process. In turns, one partition (fold) is separated for testing while the other 9 are used for training the model. The performance of the ROC curve at each fold, the average and the standard deviation across the 10 folds are the values reported. By comparing the results of groups 1 and 3, in the one-sided paired t-test, we accept the alternative hypothesis that the mean of the education group (3) is greater than that of the infrastructure group (1) at 0.05 significance level.

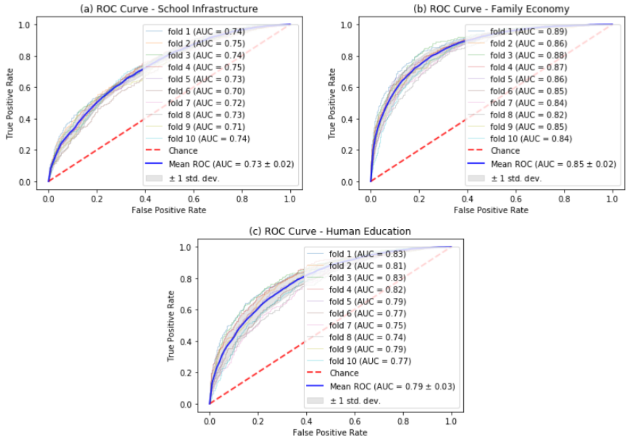

Figure 1: represents (a) the outputs for schools' infrastructure group, (b) the outputs for students' socioeconomic-cultural information group and (c) parents and teachers' level of education group.

The difference is highlighted even more when analyzing the number of variables in each group. Group 1 has 23 variables that represent the school structure while group 3 has only 3 variables, namely, the level of education of fathers, mothers and teachers. Analyzing group 3 in a logistic regression model on the whole dataset for assessing the features' influences according to their β coefficients, their predictive powers were in decreasing order, the father's education, the mother's education and the teachers' education. The qualification of teachers, in contrast to existing studies [18], does not have high explanatory power. This result is probably due to the fact that, in Brazil, the number teachers with M.Sc. and Ph.D. degrees in public secondary education is minimal (4.8% and 1.1%, respectively). Table 1 displays the beta coefficients of each variable and their p-value, well below the 0.05 significance level.

4. DISCUSSION

Several studies have improved the understanding of the determinants of school performance with the perspective of guiding educational policies. James Coleman, in 1960, had already identified the SEC factors of the students as the main determinant of their performance [13]. The correlation between parental education has been long established,

Table 1: Variable importance of the Logistic Regression model for group 3

Feature	β	p-value	Grain
Father's Education	3.87	0.00	Student Level
Mother's Education	2.94	0.00	Student Level
Teacher's Education	2.62	0.00	Teacher Level

as well [15, 11]. Other lines of research have also highlighted the relevance of aspects related to schools and teachers [18, 25, 14]. Much has been discussed in Brazil about the secondary education, as well as improving the budget allocation.

According to OECD, Brazil's public spending on education was close to the average of its member countries in the year of 2015 while the performance of Brazil in the last PISA exam was among the worst countries evaluated [27]. The quality of education still does not respond to the investments made. Therefore, it is crucial to improve the understanding of standardized national tests to help policymakers and education related institutions in developing educational public policies to produce an effective return on investment.

4.1 Parents Education as Proxy to SEC?

Separating out the independent effects of family education and SEC background is not a simple task. Some prior studies showed that those features are very correlated to family income, once parents who are more educated, earn higher salaries [24]. From another perspective, more education empowers parents and teachers to give the students better counseling and training. Some studies have tried to isolate the effect of each feature, aiming at determining causal relations between them in the educational outcomes [9] .

This Subsection attempts to dissociate these characteristics to find out if the parents' education influences the student performance in the ENEM Exam for families with the same constant income. We started by considering only students from schools with infrastructure at basic level or above. To block any effect of economics, the students were undistinguishable by their family income which was kept constant. The performance was measured by the fraction (percentage) of good students in the sample, for each level of education.

Figure 2 shows the fraction of good students against the parents' education for each income value. It is clear that higher parents' education is associated with higher fraction of good students, with the income constant. Despite being a categorical feature, the level of education is associated with time of schooling, therefore suited to line graph representation.

Wrapping up, the students' performance on ENEM increases with their parents' level of education no matter their family income.

5. CONCLUSION

This paper has presented a comparative study of the influence of groups of predictors in the quality assessment of secondary school in Brazil to help policy makers in educational budget allocation.

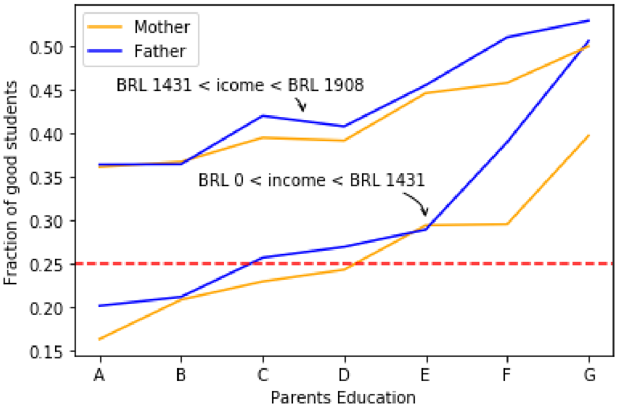

Figure 2: Fraction of good students as function of their parents' education level. Ranges from "A" (no schooling) to "G"(postgraduate) for different family incomes indicated in the curves.

The experimental procedure had logistic regression as predictive technique and the comparison was performed with single-tailed t-test on a 10-fold cross-validation setting on paired test sets. The predictors (features) were partitioned into 3 groups as planned: Infrastructure, SEC information and Education. The performance metric was the Area Under the ROC curve (AUC_ROC) widely applied for assessing binary classifiers in domains such as medicine, telecommunications, artificial intelligence etc.

The results show that both the groups of features of parents' and teachers' education and of socio-economic-cultural information are more influential than the group composed of infrastructure features with statistical significance of 0.05.

Some research found in the literature argue that there is a high correlation among the predictors and that there could be causality in SEC information influencing the Education predictors. We have shown that Education predictors have a positive effect on the students' performance no matter the family income. Nevertheless, much more analyses have to be made in that sense.

Furthermore, ensemble of predictors in general achieve higher improvement in performance with the increase of complementarity among their modules [31]. That suggests that the higher increase of AUC in the combination of SEC and education features versus the combination of SEC and infrastructure features might be related to the smaller correlation of education compared to infrastructure both in relation to SEC features. This needs to be further investigated. It is also important to extend the analyses presented here for 2018 to several years to verify if the results found hold across time. We are carrying out the research and the preliminary results show that the same behavior holds for the previous 9 years as well.

It is important that experts in education and policy makers collaborate in this research to help improve the Domain-Driven Data mining approach by embedding their expertise in the solution development.

6. REFERENCES

[1] Enem. *INEP*, 2020.

[2] Microdados. *INEP*, 2020.

[3] P. J. Adeodato. Data mining solution for assessing brazilian secondary school quality based on enem and census data. In *Proc. 13 CONTECSI*, pages 2658–2679, 2016.

[4] P. J. L. Adeodato, F. C. Pereira, and R. F. O. Neto. Optimal categorical attribute transformation for granularity change in relational databases for binary decision problems in educational data mining. *CoRR*, abs/1702.08745, 2017.

[5] S. Barretto. Politicas de formacao docente para a educacao basica no brasil embates contemporaneos. *Revista Brasileira de Educação*, 20:679 – 701, 2015.

[6] R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical science*, pages 235–249, 2002.

[7] J. Brophy and T. Good. Teacher behavior and student achievement in m. witrock (ed.), the third handbook of research on teaching (pp. 328–375), 1986.

[8] M. M. Burke. Improving parental involvement: Training special education advocates. *Journal of Disability Policy Studies*, 23(4):225–234, 2013.

[9] D. Card. The causal effect of education on earnings. In *Handbook of labor economics*, volume 3, pages 1801–1863. Elsevier, 1999.

[10] M. Carnoy, T. Khavenson, L. Costa, I. Fonseca, and L. Marotta. Is brazilian education improving? a comparative foray using pisa and saeb brazil test scores. *A Comparative Foray Using PISA and SAEB Brazil Test Scores (December 16, 2014). Higher School of Economics Research Paper No. WP BRP*, 22, 2014.

[11] A. Chevalier, C. Harmon, V. O'Sullivan, and I. Walker. The impact of parental income and education on the schooling of their children. *IZA Journal of Labor Economics*, 2(1):8, 2013.

[12] A. Chudgar, T. Luschei, and L. Fagioli. Constructing socio-economic status measures using the trends in international mathematics and science study data. *East Lansing: Michigan State University*, 2012.

[13] J. S. Coleman, E. Campbell, C. Hobson, J. McPartland, A. Mood, F. Weinfeld, et al. Equality of educational opportunity study. *Washington, DC: United States Department of Health, Education, and Welfare*, 1966.

[14] L. Darling-Hammond. Teacher quality and student achievement. *Education policy analysis archives*, 8:1, 2000.

[15] P. E. Davis-Kean. The influence of parent education and family income on child achievement: the indirect role of parental expectations and the home environment. *Journal of family psychology*, 19(2):294, 2005.

[16] P. Dolton, O. Marcenaro, R. d. Vries, and P.-W. She. Global teacher status index 2018. 2018.

[17] T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[18] D. D. Goldhaber and D. J. Brewer. Evaluating the effect of teacher degree level on educational performance. 1996.

[19] C. A. T. Gomes and M. R. T. Duarte. School infrastructure and socioeconomic status in brazil. *Sociology and Anthropology*, 5(7):522–532, 2017.

[20] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[21] R. A. Johnson, D. W. Wichern, et al. *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ, 2002.

[22] H. Jones. More education, better jobs? a critical review of ccts and brazil's bolsa familia programme for long-term poverty reduction. *Social Policy and Society*, 15(3):465–478, 2016.

[23] C. Longbing. Introduction to domain driven data mining. In *Data Mining for Business Applications*, pages 3–10. Springer, 2009.

[24] P. Lundborg, M. Nordin, and D. O. Rooth. The intergenerational transmission of human capital: the role of skills and health. *Journal of Population Economics*, 31(4):1035–1065, 2018.

[25] F. J. Murillo and M. Román. School infrastructure and resources do matter: analysis of the incidence of school resources on the performance of latin american students. *School effectiveness and school improvement*, 22(1):29–50, 2011.

[26] J. J. S. Neto, G. R. De Jesus, C. A. Karino, and D. F. De Andrade. Uma escala para medir a infraestrutura escolar. *Estudos em Avaliação Educacional*, 24(54):78–99, 2013.

[27] OECD. *Rethinking Quality Assurance for Higher Education in Brazil*. 2018.

[28] R. L. Silva Filho and P. J. Adeodato. Data mining solution for assessing the secondary school students of brazilian federal institutes. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 574–579. IEEE, 2019.

[29] B. Stauss, F. Nordin, and C. Kowalkowski. Solutions offerings: a critical review and reconceptualisation. *Journal of Service Management*, 2010.

[30] R. TRAVITZKI. *ENEM: limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar. 2013*. PhD thesis, Tese (Doutorado em Educação)–Universidade de São Paulo, São Paulo, 2013.

[31] D. H. Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.