Modeling

Juan David Nieto García, Giovani de Almeida Valdrighi

1 Introduction

In this context of school evaluation, our primary goal is to pinpoint schools facing challenges in SAEB scores that could receive extra support and funding to improve their academic standing. To achieve this objective, we embark on a binary classification of schools, discerning those in need of assistance. Given that our dataset comprises only SAEB scores, which manifest as continuous values ranging from 0 to 500, a crucial initial step involved discretizing these scores into two distinct classes.

Once the binary classification framework is established, it is important to consider the different types of errors. It is imperative to note that our primary aim is to provide support rather than impose punitive measures. Consequently, while extending assistance to schools that may not necessarily require it presents some issues, the substantial issue lies in not identifying schools genuinely in need of support. An important consideration in our study was to perform a selection of features, as the original data present a large number of features that could be difficult to analyze. In the next sections, we describe our steps of modeling.

2 Related Works

Among some reported works demonstrating the use of plug-in methods and risk applications in education or student evaluation, the following are mentioned:

Thomas Asril (2020), focuses on enhancing higher education by utilizing the k-Nearest Neighbor (k-NN) algorithm to predict student study periods based on final grades [1]. The work was carried out at BINUS University with data from 1989 computer science students (2016–2019), the study achieved high accuracy rates (93.2% for on-time completion, 91.5% for total study years, and 75.63% for total study semesters). The research concluded the k-NN algorithm is effective in forecasting student study durations, ultimately contributing to increased graduation rates.

On the other hand, Tanner (2011), examined the early prediction of student performance in an online touch-typing course, employing k-NN and its variants [2]. The study based on a database of 15000 students, demonstrated the effectiveness of k-NN in accurately predicting performance from the initial lessons. The findings suggested that early skills tests served as strong indicators for final scores in skill-based courses.

Also is reported the work of Asogbon (2016), who conducted a study to addressed the crucial role of students educational data in higher learning institutions and emphasized the significance of effectively evaluating and predicting the performance of incoming students [3]. The author constructed a Multi-class Support Vector Machine (MSVM) using an educational dataset from the University of Lagos (Nigeria), to assess the performance of the MSVM predictor. The experimental findings demonstrated that the MSVM, particularly with K-fold cross-validation (k=7), accurately predicted student performances across all categories. This approach signifies a valuable step towards achieving qualitative education standards by ensuring appropriate student placements in faculty programs.

Works such as the above demonstrate the vital role of plug-in and risk-minimization methods methods in education. The effectiveness of the k-Nearest Neighbor and Support Vector Machines methods, in the early prediction of performance and student placement was demonstrated, emphasizing the importance of machine learning advanced methods to improve educational results and inviting the exploration of other methods of the same family.

3 Methodology

3.1 Data Preparation

To use our dataset in the models, a few extra steps that were not done previously in the data preparation stage were necessary.

First, a column with the same value for all the rows was removed. Also, the ID of the city was removed because it has too many unique values that wouldn't be possible to use. The original dataset presented answers in a coded format; for example, the information on school location was number 1 for urban or number 2 for rural. To facilitate the interpretation of features, this coding was reverted with some columns: the region of the school (south, north, ...), the state of the school, if the school is in the capital, if the school is in the urban area.

Next, there were many columns with hierarchical relations, i.e., if question Q was not answered, it was unnecessary to answer question R. However, most of the questions of type R (the child in the hierarchy), have the format "How many times does event A occur?" (or similar), and not answering the question had the same interpretation of an answer equal to 0. For that reason, columns similar to the R described, in which the value was absent due to the hierarchy, were replaced by 0.

After this transformation, the columns were divided into binary features, categoric features, and numeric features. The binary features presented numeric values but only 1 or 0, which should be interpreted as a positive or negative answer. The categoric features presented text values with different quantities of unique values, and it was necessary

to perform one-hot encoding in these columns. Then, the numeric columns presented values in different ranges, and to have more interpretable learned weights, it was the columns were standardized (mean 0, variance 1). A last step was performed to remove the variables from the data that presented a correlation bigger than 0.6 with any of the other variables. This is important to reduce the dimension of the dataset and improve the learning of some algorithms.

3.2 Discretization of regression

To transform the problem into a classification problem, it was necessary to select a threshold τ in the SAEB score to define positive and negative classes. Schools with SAEB scores lower than τ will be of positive class, i.e., support is necessary, and schools with scores bigger than τ will be of negative class. One initial intuition to select the threshold is to consider 50% or 60% of the max score, i.e., 250 or 300. To make this decision with more information, we look at the distribution of scores. In Fig. 1 we display the distribution of each score, separated by the socioeconomic index, which is a feature present in the data that has 7 different values of socioeconomic condition, the bigger, the better. It is possible to see a big correlation between the socioeconomic index and the scoring; the higher the socioeconomic index is, the higher the mode (peak of distribution) is. Most of the distribution mass is located inside the interval $[200, 300]$. Inspired by this Figure, we opt to use the threshold has 250.

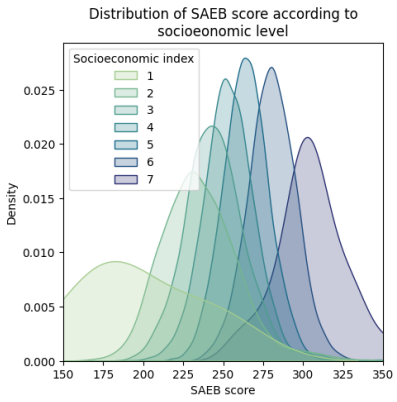

Fig. 1: Distribution of SAEB scores of schools grouped by socioeconomic index.

3.3 Models

We considered a diverse set of models in our classification problem. The models can be divided in two-groups, plug-in methods that try to estimate the function $P(Y = 1|X)$ and risk-minimization methods, which have an iterative approach to minimize a loss function (risk). In the first group, it was considered the models k-Nearest Neighbors and Logistic Regression, and in the second group it was considered Support Vector Machine (SVM) and Extreme Gradient Boosting (LGBM).

k-Nearest Neighbors: The method classifies or predicts values based on feature similarity to the k closest examples in a training data set. When used for classification, the distance between the new point and the nearest neighbors is calculated and the most common label is taken, while for regression the values of the k neighbors are averaged. k is an adjustable hyperparameter that affects the accuracy of the model, a low value of k tends to increase the variance of the model, meaning that it overfits the training data, while a high value tends to increase the bias, meaning that complex relationships in the data are underestimated data [4].

SVM (linear): It is used for data classification, its main objective is to find a hyperplane that maximizes the separation between two classes in a multidimensional space, which is known as the maximum margin. The support vectors, which are the points closest to the hyperplane, play a crucial role in defining this margin. Linear SVMs are effective in binary classification and can be extended to multi-class problems. They can handle data that is linearly separable, and by using kernel functions, they also deal with nonlinearly separable data by mapping it to a higher dimensionality space [5].

Logistic Regression: Mostly used in binary classification problems, it predicts whether an item belongs to one of two categories. It uses a logistic function to calculate the probability that a data point is in a specific class, and then makes a decision based on a threshold. The model is trained to adjust the optimal coefficients that best fit the data, minimizing the cost function. This method is widely used for its simplicity and ability to provide classification probabilities [6].

Gradient Boosting: Used in regression and classification problems, it operates by building a predictive model in stages, where each stage improves the deficiencies of the previous model. At each iteration, a weak model (usually decision trees) is fitted to predict the residual errors of the previous model. The weak models are then weighted and combined to form a strong model that is highly accurate. The process is repeated iteratively until optimal performance is achieved. This method is effective at handling noisy data and complex relationships, but may require proper hyperparameter tuning to avoid overfitting [7].

All methods have hyper-parameters which are important to consider in the training step. To do that, the data was divided into training and tests, and the training subset was used to perform a 5-fold validation. The folds were used to identify the best hyper-parameters in the models. For each of the models, we considered the following hyper-parameters:

- **k-Nearest Neighbors**: the number of neighbors k to consider in the prediction (7 combinations).
- **SVM**: the regularization weight C that penalizes model complexity (5 combinations).
- **Logistic Regression**: the regularization weight C that penalizes model complexity (7 combinations).
- **Gradient Boosting**: the max depth of decision trees, the number of decision trees, the fraction of samples used by each tree, and the weight of l_2 penalty λ (144 combinations).

3.4 Metrics

As previously mentioned, we want to identify schools in need of support, and in that case, it is worse to not correctly identify those schools than to classify a school that does not need support as it needs. For that reason, one important

metric to consider is the recall: recall $= \frac{TP}{TP+FN}$. If we classified every school that needs support as 1, then, the recall will be equal to 1. We could achieve recall equal to 1 by classifying every sample as positive; for that reason, it is necessary to look at other metrics, and in that case, we will look at the balanced accuracy, that is the fraction of correction predictions by class, weighted inversely by the prevalence of the class. These metrics were utilized in finding the best model: the hyper-parameters of each model that were selected were the ones that presented the highest mean recall in the 5-folds.

3.5 Feature Selection

One obstacle in our analysis is the number of features. We have a few features collected about each school on the day of the test, but we have hundreds of features answered by the school principal. It's an obstacle because it causes large fitting times for the models, and when interpreting the trained models, it can be overwhelming to understand how it works with large quantities of features. For that reason, it is important to perform a feature selection. It can be done with the help of experts in this area and with statistical approaches. We performed it with two methods: using Mutual Information score and Causality selection based on the Markov Blanket.

3.5.1 Models with high Mutual information

Mutual information score measures the degree of statistical dependence between two variables by quantifying the amount of information one variable provides about the other [8]. It is based on information theory, assessing how much knowing the value of one variable reduces uncertainty about the other. A higher mutual information score indicates a stronger relationship, in feature selection, it helps identify and retain features that contribute valuable information to the prediction task, facilitating the creation of more effective and efficient models.

For this work, we conducted a detailed analysis on the dataset, including ordinal encoding of categorical features. A mutual information-based scoring function was developed to select the top 25 relevant features. Subsets of numerical, binary, and categorical features were then created, restricted to the selected ones. Subsequently, a data processing pipeline was built, incorporating standard scaling and one-hot encoding for the chosen features. The four models trained in the initial part were trained again, this time using the reduced feature set through a grid search for hyperparameter optimization. The results of each model were stored in pickle files for subsequent analysis and comparison. This approach enables the assessment of classification model performance on more relevant feature subsets, emphasizing the impact of feature selection on model accuracy.

3.5.2 Models with Causality selection based on the Markov Blanket

The method identifies a minimal set of variables, known as the Markov Blanket, which, when conditioned upon, renders a target variable independent of the remaining

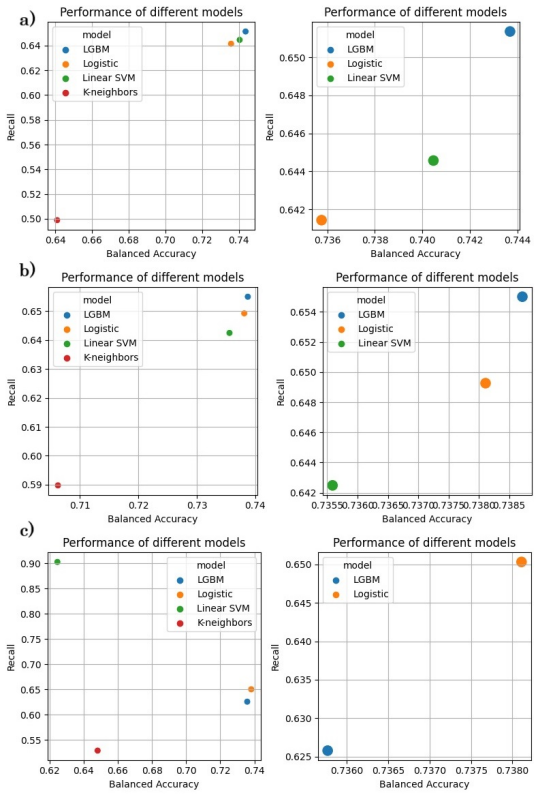

Fig. 2: Results of models on the test set for all features on a), features selected with mutual information on b), and features selected with causal independence on c).

variables [9]. This blanket effectively shields the target variable from the rest of the system, revealing direct causal relationships. In the context of feature selection, utilizing the Markov Blanket helps identify the most influential features directly affecting the target variable, reducing dimensionality and enhancing model interpretability by focusing on the variables with a direct causal impact on the outcome.

In particular, we made use of PPFS from Hassan *et al.* [10]. This is a recent approach to performing a large number of statistical tests in a feasible computing time. In comparison with other approaches to causal feature selection, this method presents better computing time. It is also easily utilized with the Python package[1]. The objective is to identify the Markov Blanked of the target variable, which is the set of variables that upon conditioning, the target variable is independent of every other variable. The algorithm performs this by training a selected model and comparing the model error with the original feature values and shuffled feature values. If the feature presents no important relation to the target, it will not impact the error. This would not consider causality without incorporating the idea of the Markov Blanked, which is performed by changing the feature set. The feature set X will contain the feature k that will be shuffled and all other features upon k is conditioned.

Despite the fast algorithm proposed, it still has high computing time. For that reason, we performed experiments using 15 simulations (the default value is 30) and using only 1000 random samples from our data. The good results

1. https://github.com/atif-hassan/PyImpetus

obtained by the final model show that this decision did not present much impact on the final result. Another consideration is that the method works with categorical variables, yet, the Python implementation only has support for numeric features inside the interval of [0, 1]. For that reason, it was necessary to perform the one-hot encoding and standard scaling of all features. The method selected some categorical features, for example, "UF = Ceará", which is not an original feature, but a feature created by the one hot encoding. In these scenarios, we opt to train the models using the original feature "UF", and perform one hot encoding in it, using all the generated columns, not only "UF = Ceará".

3.6 Model with Bayesian Risk and Calibration

In this last phase of the work, the LR model (with causal features) was employed to predict the performance in a Bayesian risk framework. Trained classifiers were loaded, and the model's predicted probabilities were obtained for both the training (`X_train_causal`) and testing datasets (`X_test_causal`). To optimize the classification threshold and minimize Bayesian risk, an iterative process adjusted the threshold across a range of values. The process involved evaluating the Bayesian risk at each threshold and selecting the threshold that yielded the minimum risk. Subsequently, the new threshold was applied to the testing data, generating a confusion matrix and assessing performance metrics, including balanced accuracy and recall.

For the calibration, we performed the model calibration on the LR classifier. The process involved calculating predicted probabilities for the positive class on the training set, followed by the computation of the Brier score to assess the accuracy of these predicted probabilities. Subsequently, a calibration curve was generated to visually represent the alignment between predicted probabilities and the actual occurrence of positive instances. This curve serves as a diagnostic tool, allowing us to evaluate the reliability of the model's predicted probabilities.

4 Results

In Fig. 2 a), we plotted the balanced accuracy against the recall calculated on the test set for the four models considered. These results were obtained with the best hyperparameters on the validation set. It's possible to identify that discarding the result from k-NN, the models obtained similar results, with around 0.64 of recall and 0.74 of balanced accuracy. The small improvements between the models place the LGBM as the best one, followed by the linear SVM. These results are positive; the recall bigger than 0.7 can be interpreted that based on the school attributes, 70% of the schools that presented lower educational results can be identified. In Figure 2 b), the results obtained for the four models using mutual information are presented. It is possible to show that by reducing the number of features, LGBM continues to be the best model. In Figure 2 c), the results obtained for the four models using causality selection based on the Markov Blanket, in this situation, Logistic Regression showed better results.

In reference to the results for the prediction of the LR model in a Bayesian risk framework, the obtained results reveal that the optimized threshold, set at 0.29, effectively minimized Bayesian risk to 0.59. Figure 3 illustrates the comparison between the confusion matrix of the original model and the model using the threshold that minimizes the Bayes risk. A balance between true positives and true negatives, with an 83% recall, indicating the model's ability to identify positive instances. The new balanced accuracy of 0.74 signifies robust overall performance. This approach not only enhances the interpretability of model predictions but also suggests the implementation of strategies, particularly catering to schools with underperformance. Finally, Figure 4 presents the calibration curve of our model, illustrating how the predicted probabilities align with the actual frequency of positive events.

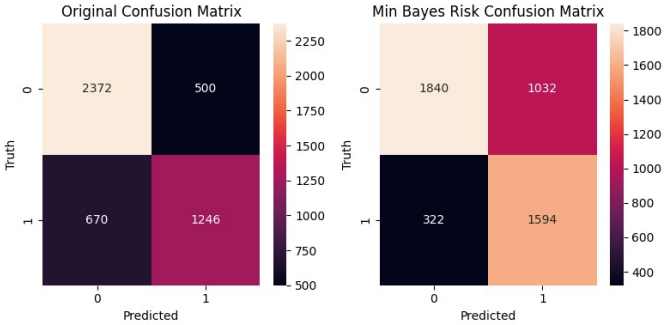

Fig. 3: Comparison between confusion matrix of original model and model using the threshold that minimizes the Bayes risk.

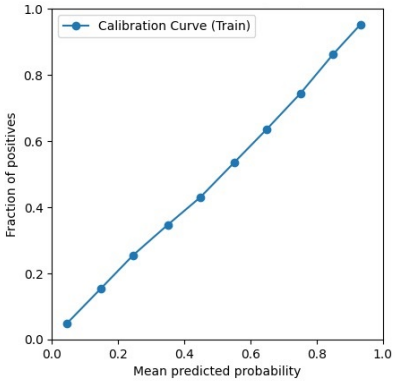

Fig. 4: Calibration Curve.

5 Discussion

This study aimed to develop predictive models for identifying schools in need of support, selecting hyperparameters, and using feature selection techniques. The 5-fold validation process was crucial for optimal hyperparameter selection, emphasizing recall. Dealing with a large number of features was addressed through effective feature selection. Results highlighted the consistent outperformance of the LGBM model, emphasizing its suitability. Scatterplots illustrated the balanced accuracy-recall trade-off, offering a comprehensive view. The Bayesian risk framework optimized the LR model, achieving a balance between true positives and true negatives. Findings support practical application in educational decision-making for targeted interventions in underperforming schools.

REFERENCES

[1] S. M. I. Thomas Asril, "Prediction of students study period using k-nearest neighbor algorithm," *International Journal of Emerging Trends in Engineering Research*, 2020.

[2] H. T. Tuomas Tanner, "Predicting and preventing student failure – using the k-nearest neighbour method to predict student performance in an online course environment," *International Journal of Learning Technology*, 2011.

[3] M. G. Asogbon, O. W. Samuel, M. O. Omisore, and B. A. Ojokoh, "A multi-class support vector machine approach for students academic performance prediction," *International Journal of Multidisciplinary and Current Research*, 2016.

[4] D. T. Larose and C. D. Larose, *k-Nearest Neighbor Algorithm*, 2014, pp. 149–164.

[5] S. Ray, "A quick review of machine learning algorithms," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 35–39.

[6] L. Connelly, "Logistic regression," *Medsurg Nursing*, vol. 29, no. 5, pp. 353–354, 2020.

[7] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[8] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, pp. 175–186, 2014.

[9] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation." *Journal of Machine Learning Research*, vol. 11, no. 1, 2010.

[10] A. Hassan, J. H. Paik, S. Khare, and S. A. Hassan, "Ppfs: Predictive permutation feature selection," *arXiv preprint arXiv:2110.10713*, 2021.

APPENDIX
MODEL CARD FOR SCHOOL CLASSIFICATION LOGISTIC MODEL

Classifier of schools in need of special support (schools that present low results in SAEB) based on feature values such as locality, administration type, etc.

Model Details

Model Description

Ensemble Classifier of schools in need of special support, i.e., schools that presented results lower than 250 on SAEB. The model used is Logistic Regression and uses a subset of features obtained from the day of application of the test, the school's characteristics, and answers from the school director.

- **Developed by:** Juan David Nieto and Giovani Valdrighi
- **Model type:** Logistic Classifier

Model Sources

- **Repository:** https://github.com/giovanivaldrighi/school_eval_ethical
- **Model files:** https://drive.google.com/drive/folders/1-8MBKLY2732kdCHe1S0H_xZHLao0VlaZ?usp=sharing

Uses

Direct Use

Identify schools that are underperforming in lessons of math and Portuguese and need support and special projections to surpass this condition.

Out-of-Scope Use

This model is not suited to identify schools that should be punished due to underperforming. Some of the features are related to the socioeconomic characteristics of the school, and the label obtained from SAEB scores is also related to the socioeconomic characteristics of students and teachers. The classification is not able to reflect the effectiveness of the school in teaching.

Bias, Risks, and Limitations

The model presented good performance, however, the data is based on the answers of school directors in a particular year. It can have learned some spurious correlations. The target variable is also a proxy of the intended label. The system could be gamed by changing features and values that are easily changed and do not provide a change in school performance.

Evaluation

Metrics

Train: ROC: 0.85 Balanced accuracy: 0.76 Recall score: 0.66 Precision score: 0.76

Test: ROC: 0.82 Balanced accuracy: 0.74 Recall score: 0.63 Precision score: 0.73

Model Examination

Feature selection

Features were selected using causality formulation, i.e., using the Markov Blanket independence test. Yet, a total of 50 features need to be validated with experts, as it was obtained from observational data, which can be not representative of the real causal system.

Feature importance

However, we note a potential bias associated with the socioeconomic level of schools, substantiated by the significant importance of the socioeconomic dimension in modeling (Fig. 5 and Fig. 6). This underscores the need to carefully consider this aspect in future iterations of the model.

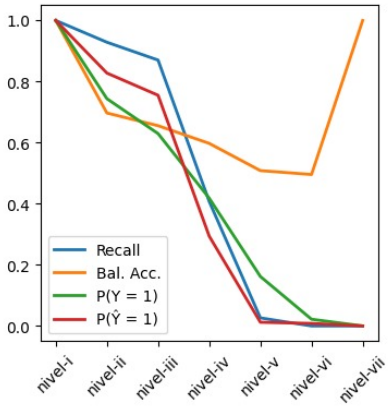

Fig. 5: Performance Metrics Across Economic Levels.

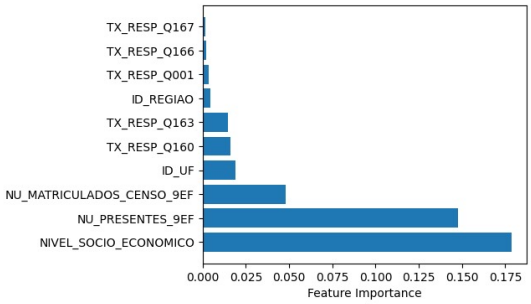

Fig. 6: Feature selection for some features.