International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-9, Issue-3; Mar, 2022
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.93.33

A Machine Learning Approach on the Problem of Corruption

Luciano M. C. Doria[1], Felipe F. Doria[2], Paulo Figueiredo[3], Adilson Sampaio[3], Renelson Ribeiro Sampaio[1]

[1]Departamento de Modelagem Matemática, SENAI CIMATEC Faculdade de Tecnologia, Salvador, Bahia, Brazil
[2]Instituto de Física, Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil
[3]Escola de Administração, Universidade Federal da Bahia, Salvador, Bahia, Brazil
Email: lucianomcdoria@gmail.com

Received: 06 Jan 2022,

Received in revised form: 01 Mar 2022,

Accepted: 11 Mar 2022,

Available online: 30 Mar 2022

***Keywords*— Corruption, Machine Learning, São Paulo State Court of Auditors, XGBoosting.**

***Abstract*— This work presents a step-by-step building of a model that effectively classifies a given municipality as corrupt or not. The output is the likelihood of the city being corrupt, which can be a valuable tool in preventing future corruption cases. This model was constructed to utilize the already existing API from the São Paulo State Court of Auditors and was built to deploy monthly reports. The XG Boosting model was the most robust among the many models trained and presented the best AUC score and accuracy.*

I. INTRODUCTION

Corruption is a phenomenon observed in human society over time [1-3]. Due to being so old, its definition tends to vary depending on the time and culture [1]. There is no international consensus on the meaning of corruption [4]. However, it can be measured not directly but through indices such as the Corruption Perceptions Index (CPI) [4]. As it is generally a survey of opinion and experience, all these indices need to be used with caution. The CPI would be measuring the perception of corruption and not corruption itself [4],[5].

Machine Learning models for classification have been extensively used to predict several different problems accurately, like Churn [6-8], People Analytics [9], Credit Risk [10-12], Corruption [13],[14], etc. Mainly, the corruption models employed an index of perception to mark the corruption.

An exciting style of machine learning models to classify corruption in Brazilian municipalities has been introduced [15],[16]. The approach to marking the answer variable is dependent on the annual public audit reports from the Brazilian Government Accountability Office (CGU). As this is a yearly variable, the actions taken may have a slower pace.

This raises the question, “Is it possible to create an effective model with good performance without using indices with more regular outputs?”. In this work, the target variable was classified through a federal police operation, which made it possible to discretize the marking of municipalities every month. Still thinking about celerity, all variables used were monthly so that the outcome would be monthly instead of annual.

The previous results observed that the private sector, financial development, and human capital features are the most critical predictors. In contrast, the public sector and political features are only minor contributions [15]. Nevertheless, as our work focused on the data retrieved exclusively from the public sector and attained good results in classifying a corrupt municipality, it is likely to assume that some correlations were not previously taken into consideration between the public sector features and the formerly thought to be the most critical aspects in measuring corruption. Thus, the most vital features are the expenses and revenues of a given municipality, specifically the expenses with basic sanitation.

This work will present a model capable of providing the likelihood of a given city containing traces of fraud in its management. The next session visits some machine learning classification models and explores which metrics would be suitable to measure the model's performance. Then in session III, the steps necessary to prepare the data for the modeling are shown. Session IV shows the results of the trained models and discusses their implications, in particular, the XG Boosting model and the AUC score for the Out of Time dataset and the importance of the most important features. Finally, it talks about the conclusion and next steps of the research.

II. METRICS AND ML MODELS

This section briefly explains the models utilized in this work and the metrics employed to measure its performance.

2.1 Models

2.1.1 Logistic Regression

A Logistic Regression calculates a weighted sum of input variables and outputs the logistic of the result. If the estimated probability is superior to 50%, the model predicts one class. Otherwise, the model predicts another class [17].

2.1.2 Decision Tree

A Decision Tree is a Machine Learning algorithm capable of performing classification and regression. Each node of the tree is tagged with a variable. This variable is split to maximize the Gini impurity or Entropy. Then the leaf obtains the value of 0 or 1 (one of the two classes) depending on the value of the most significant class [17],[18].

2.1.3 Random Forest

A Random Forest algorithm is nothing more than an ensemble of Decision Trees. It means that there are many decision trees voting for each prediction, and the class with a majority of votes ends up as the predicted class [17].

2.1.4 Gradient Boosting

Gradient Boosting works by consecutively adding up predictors to an ensemble. It tries to fit the new predictor to the residual errors made before by its previous iteration [17].

2.1.5 XG Boosting

XG Boosting stands for Extreme Gradient Boosting, and it is nothing more than an optimized implementation of Gradient Boosting [17].

2.2 Performance Metrics

There will be a prediction and the true value for a given dataset in a binary classification model. If the true value is positive and the forecast value is correct, the result is a True Positive (TP). If it is wrongly classified, the result is a False Negative (FN). If the true value is negative and the predicted value is negative, the result is a True Negative (TN); if it is incorrectly classified, it is a False Positive (FP).

2.2.1 Accuracy

The accuracy measures the overall hit and miss of a classification problem. It is the ratio of all the hits by all the values [19]. It is represented by the equation below,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

2.2.2 Precision

Precision is the fraction of hits of a given class by all that was classified as that class [19],

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

2.2.3 Recall

The recall is the ratio between the hits of a class by all the actual values of that class [19],

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

2.2.4 AUC score

Sometimes all the other metrics fail to show how well the model fares separating the classes. For this situation, the Receiver Operating Characteristic (ROC) curve can be plotted with the FP rate at the x-axis and the TP rate at the y-axis. The area under the curve (AUC) can then be calculated, and this is the AUC score [19],[20]. Values of 50% represent a case where the rate of TP and FP is the same, which is equivalent to a toss of a coin, so the values of AUC go from 50% to 100%.

2.2.5 Gini

To get a more intuitive score, the AUC score can be normalized getting a score from 0% to 100% [21] by the following equation,

$$Gini = 2 \times AUC - 1 \qquad (4)$$

III. METHODOLOGY

This paper's main idea and final objective are to create a model capable of analyzing monthly data from a city to find the probability of fraud or corruption.

At first, it was necessary to define what was considered as the target variable from the model. Given its nature, it Is challenging to measure corruption directly, and utilizing any perception index can be tricky, possibly creating unexpected bias [4].

For objectivity, the target variable was marked as 1 for a given municipality if the mayor was found with traces of fraud since the start of his term and 0 otherwise. The data to pinpoint the target variable came from the "Prato Feito" operation, a federal police operation in Brazil.

Triggered on May 9th of 2018, the “Prato Feito" operation was investigating the embezzlement of public resources intended initially for education (such as school lunches, school supplies, and uniforms) in 30 municipalities [22].

The data comprised all expenses and revenues of the municipalities in question and was obtained via API from the São Paulo State Court of Auditors (TCE-SP) [23]. It was arranged based on the city, the year, and the month it was recorded, meaning there will be a line in the database for each month of each year. As the features were based on the observation of the 30 cities, some don’t spend in specific areas; for them, it was filled with the value 0. The raw dataset comprised 1979 rows with 695 rows marked as a city with traces of fraud.

The next step was to split the dataset. First, the data of four cities were separated in the Out of Sample dataset (OOS); it was also added the data of 3 more cities that did not belong in the operation; the reason for this is that this data will not be used for training, then the year 2019 was separated in the Out of Time dataset (OOT), this data as well will not be used for training, and finally the remainder of the dataset was split in training comprised of 70 % and test consisting of 30%. To Ensure reproducibility, a random state was arbitrarily chosen as 42.

Subsequently, the model was trained with five different classification machine learning algorithms: Random Forest, Gradient boosting, XG Boosting, Decision Tree, and Logistic Regression.

This model was built to utilize the API data efficiently; in the end, based on the monthly value of the features, the output will be the probability of a municipality having some corruption.

The next session discusses the results and their meaning.

IV. RESULTS AND DISCUSSIONS

After splitting the data in the training dataset, test dataset, out of sample dataset, and out of time dataset, multiple algorithms were utilized for the model training.

As it is a classification problem (is the city classified as having fraud or not?) and supervised learning are the class of models being utilized, it is possible to measure the efficacy with metrics checking the hits (True Positive/True Negative) and misses (False Negative/ False Positive) of the output in comparison to the Test, OOT and OOS answer.

Table.1: Accuracy, precision, and recall of the trained models. XG Boosting has better discrimination and stability.

		Test	OOS	OOT
Random Forest	Accuracy	81%	76%	78%
	Precision	79%	67%	79%
	Recall	67%	64%	57%
Logistic Regression	Accuracy	74%	73%	68%
	Precision	69%	68%	64%
	Recall	54%	45%	39%
Gradient Boosting	Accuracy	79%	75%	78%
	Precision	76%	66%	83%
	Recall	64%	60%	53%
XG Boosting	Accuracy	82%	75%	81%
	Precision	79%	65%	85%
	Recall	70%	64%	62%
Decision Tree	Accuracy	73%	71%	70%
	Precision	66%	59%	63%
	Recall	57%	58%	52%

At TABLE 1, accuracy is being measured by the ratio between the hits of the model and the True Value of the variable; precision is being measured by the ratio between True Negatives and the sum of True Negatives and False Negatives; recall is being measured by the ratio of the True Negatives and the sum of True Negatives and False Positives. The model performance at the Test dataset shows the model’s behavior in an environment very similar to the training. As specified before, this is the first time the model sees the Test dataset. However, the model could not be robust enough to score datasets with never-seen municipalities or score datasets with different timeframes. The results for the OOS dataset can be applied to check the model's capabilities to get the correct outcome with never-seen cities. For the different timeframes, the results for the OOT dataset can be employed.

The XG Boosting model shows a better performance for the Test dataset with an accuracy of 82%, a precision of 79%, and a recall of 70%. The Random Forest model with an accuracy of 76%, precision of 67%, and a recall of 64% displays marginally better performance for the OOS dataset than the XG Boosting. For the OOT dataset, XG Boosting once more presents a superior performance with an accuracy of 81%, precision of 85%, and a recall of 62%. It is crucial to understand that the data of the OOT dataset is more like how the data is obtained periodically by the API; therefore, having a good performance at this dataset is essential for the robustness of the model.

As the dataset per se is unbalanced, it is good to analyze the prediction results deeper. The area under the ROC curve score (AUC score) is the metric for exploring this kind of data and avoiding coincidence [19]. An AUC score of 50% means that the model classifies True Positive and True Negative values at the same rate. When the value is higher than 50%, the True Positive rate is greater than the False Negative. If trained correctly with care not to be biased, the more an AUC score gets closer to 100%, the better at separating the model's classes [21][24].

Table.2: AUC score and Gini coefficient of each model for the OOT dataset.

		OOT
XG Boosting	AUC	90%
	GINI	80%
Random Forest	AUC	89%
	GINI	77%
Gradient Boosting	AUC	82%
	GINI	65%
Decision Tree	AUC	67%
	GINI	35%
Logistic Regression	AUC	68%
	GINI	36%

The Gini coefficient can be derived from the AUC score normalizing the metric and quantified from 0% to 100% [21].

At TABLE 2, each model's AUC score and Gini Coefficient are placed for comparison. And one more time, the XG Boosting presents the best performance with an excellent AUC score of 90% and a Gini coefficient of 80%.

Fig.1 shows the ROC curve for the OOT dataset of the multiple models trained. The XG Boosting model obtained a better AUC score based on the size of the area under the curve. The next better performance is from the Random Forest model, followed by the Gradient boosting model, the Decision Tree and Logistic Regression models performed similarly bad.

Considering the AUC score, the Gini coefficient, the accuracy, the precision, and the recall, it is reasonable to infer that this XG Boosting model is suitable for identifying fraud traces in a municipality.

In Fig. 2, the feature importance is drawn, considering the variable frequency utilized to split the data across all trees composing the model [19].

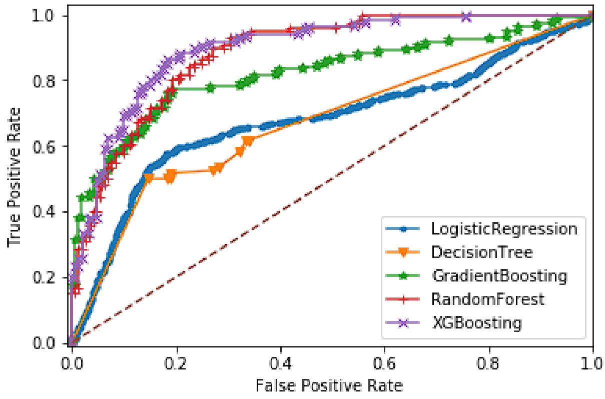

Fig. 1: OOT ROC curve of the models. The XG Boosting is represented by the purple curve with "x" marks and obtained an AUC score of 90%. The Random Forest is represented by the red line with "+" patterns and received an AUC score of 89%. The green curve represents the Gradient Boosting with "^" marks and an AUC of 82%. The Decision Tree is represented by an orange line with "v" marks and an AUC score of 67%. The Logistic Regression is the blue curve with "." marks and an AUC score of 68%.

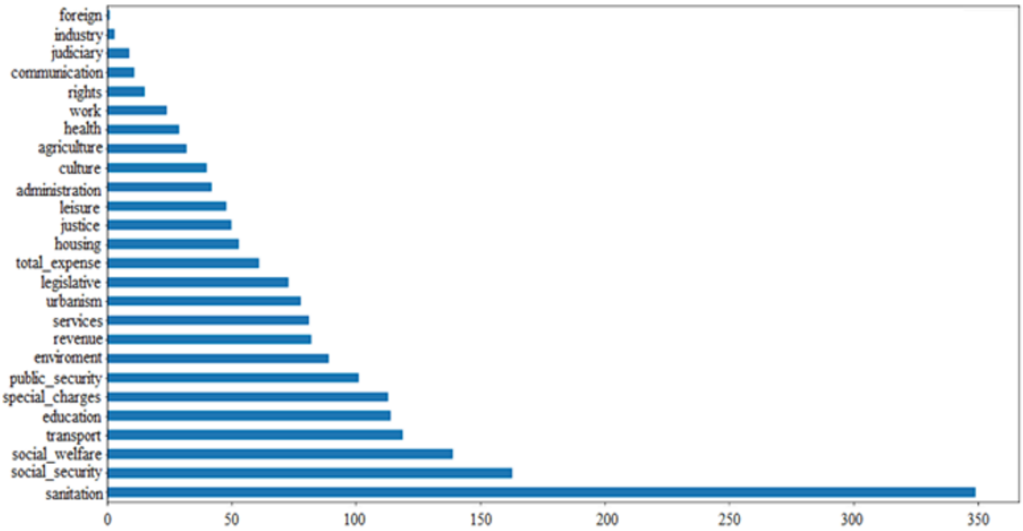

Fig. 2: XG Boosting Feature Importance. In the y axis are the feature names and in the x-axis is the frequency that features appeared in all the trees of the model.

By far, the most essential feature is the expenses with sanitation, followed by costs with social security, fees with social welfare, expenses with transport, and expenses with education. Those are charges closely related to the support of the city's population. It is feasible to infer that neglecting these features corresponds to a greater chance of the town showing traces of corruption.

The following session is about the conclusion of this research and the next steps for future works.

V. CONCLUSION

This paper focused on the problem of identifying a city that contains traces of fraud or corruption. Five machine learning classification models were trained. The models arranged by performance are XG Boosting, Random Forest, Gradient Boosting, Logistic Regression, and Decision Tree. Five metrics were utilized to measure the performance: Gini coefficient, AUC score, accuracy, precision, and recall. The best model had an AUC score of 90%, which is generally considered excellent according to the literature [25].

Moreover, according to the feature importance chart in Fig.2, it is possible to infer and extrapolate where corruption is possibly occurring. One evidence of this is that one of the top five most important features is the expense in education, which was the starting point of this article as the data utilized to mark the corruption was obtained from the cities in the "Prato Feito" operation. This result implies the importance of public sector variables differently as proposed in previous works [15].

The final model is particularly robust with datasets in different timeframes, which is how the data is obtained at the TCE-SP API, allowing the model to be a guideline for future references. The model's output is the probability that a city has traces of corruption in a month. It doesn't mean that that municipality is corrupt, but it would be a good course of action to investigate if the likelihood is high.

Indeed, the ability to return monthly reports is beneficial for the scope of a model created to help the police and auditors identify possible investigations subjects. However, monitoring the model for any change in the explicative variables and performance metrics would be necessary. This is because a policy change could reflect directly on the model's outcome. For example, how a resource should be spent or limitations in the budget could directly affect the expected delta of the previous years of a model. Consequently, to remedy that, the model should be retrained from time to time.

From this point, this work can evolve in two different ways. The first is to utilize more machine learning models like Neural Networks and quantum computation, using more refining techniques. The second is to find more triggered operations from the police and expand the dataset to the whole country.

ACKNOWLEDGEMENTS

Partial acknowledgment to SENAI CIMATEC, UFBA, and UFRGS.

REFERENCES

[1] Gomes, J. V. L. (2010). A corrupção em perspectivas teóricas. Teoria e Cultura, 5(1-2). Retrieved from: https://periodicos.ufjf.br/index.php/TeoriaeCultura/article/view/12234

[2] Ferreira Filho, M. G. (2001). Corrupção e democracia. Revista De Direito Administrativo, 226, 213–218. Retrieved from https://doi.org/10.12660/rda.v226.2001.47241

[3] Gico Júnior, I. T. (2011). Corrupção e Judiciário: a (in)eficácia do sistema judicial no combate à corrupção. Revista Direito Gv, São Paulo, 7(1), 75-98. Retrieved from SSRN: https://ssrn.com/abstract=3720565

[4] Rohwer, A. (2009). Measuring Corruption: A Comparison between the Transparency International's Corruption Perceptions Index and the World Bank's Worldwide Governance Indicators. ifo DICE Report, ifo Institute - Leibniz Institute for Economic Research at the University of Munich, vol. 7(03), pages 42-52. Retrieved from https://www.econstor.eu/handle/10419/166975

[5] Abramo, C. W. (2005). Percepções pantanosas: A dificuldade de medir a corrupção. Novos estudos CEBRAP, 73, 33–37. Retrieved from https://doi.org/10.1590/S0101-33002005000300003

[6] Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G. & Chatzisavvas, K. Ch. (2015). A comparison of machine learning techniques for customer churn prediction, Simulation Modelling Practice and Theory, 55, 1-9. Retrieved from https://doi.org/10.1016/j.simpat.2015.03.003

[7] Qureshi, S., Rehman, A., Qamar, A., Kamal, A. & [Rehman], A. (2013). Telecommunication Subscribers' Churn Prediction Model Using Machine Learning. 8th International Conference on Digital Information Management, ICDIM 2013. Retrieved from https://doi.org/10.1109/ICDIM.2013.6693977

[8] Lalwani, P., Mishra, M., Chadha, J. & Sethi, P.. (2022). Customer churn prediction system: a machine learning approach. Computing. 104. 1-24. Retrieved from https://doi.org/10.1007/s00607-021-00908-y

[9] Yahia, N. B., Jihen, H. & Colomo-Palacios, R. (2021). From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction. IEEE Access. 9, 60447-60458. Retrieved from https://doi.org/10.1109/ACCESS.2021.3074559

[10] Khandani, A. E., Kim, A. J. & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34, 2767-2787. Retrieved from https://doi.org/10.1016/j.jbankfin.2010.06.001

[11] Bhatore, S., Mohan, L. & Reddy, R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4. Retrieved from https://doi.org/10.1007/s42786-020-00020-3

[12] Ma, X. & Lv, S. (2019). Financial credit risk prediction in internet finance driven by machine learning. Neural Computing and Applications, 31. https://doi.org/10.1007/s00521-018-3963-6

[13] Lima, M. & Delen, D. (2019). Predicting and explaining corruption across countries: A machine learning approach. Government Information Quarterly, 37, 101407. Retrieved from https://doi.org/10.1016/j.giq.2019.101407

[14] Cordeiro, D. F., Cassiano, K. K. & da Silva, N. R. (2020). Identification of circulating information on corruption in Brazil using data mining and machine learning. Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 108. Retrieved from https://files.cercomp.ufg.br/weby/up/1197/o/LatinX_Cordeiro_Cassiano_DaSiva.pdf

[15] Colonnelli, E., Gallego, J. & Prem, M. (2020). What Predicts Corruption?. Documentos de Trabajo 017144, Universidad del Rosario. Retrieved from https://ideas.repec.org/p/col/000092/017144.html

[16] Ash, E., Galletta, S. & Giommoni, T. (2021). A Machine Learning Approach to Analyze and Support Anti-Corruption Policy. SSRN Electronic Journal. Retrieved from http://doi.org/10.2139/ssrn.3830220

[17] GeÌron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras and TensorFlow: concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly.

[18] Rivest, R. (2001). Learning Decision Lists. Machine Learning. 2. Retrieved from http://doi.org/10.1007/BF00058680

[19] Wang, Y. & Ni, X. (2019). A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. International Journal of Database Management Systems, 11, 1-17. Retrieved from https://doi.org/10.5121/ijdms.2019.11101

[20] Ling, C., Huang, J. & Zhang, H. (2003). AUC: a Statistically Consistent and more Discriminating Measure than Accuracy. Proc. 18th Int'l Joint Conf. Artificial Intelligence (IJCAI), 519-524. Retrieved from https://dl.acm.org/doi/10.5555/1630659.1630736

[21] Kaymak, U., & Ben-David, A. & Potharst, R. (2010). AUK: a simple alternative to the AUC. Erasmus Research Institute of Management (ERIM), 25. Retrieved from https://doi.org/10.1016/j.engappai.2012.02.012

[22] Associação de Docentes da Universidade de São Paulo. (2022, Mar. 18). Nota do balanço da Operação Prato Feito. https://www.adusp.org.br/files/docs/oppf.pdf

[23] Tribunal de Contas do Estado de São Paulo. (2022, Mar. 18). APIs. https://transparencia.tce.sp.gov.br/apis

[24] Dorfman, R. (1979). A Formula for the Gini Coefficient. The Review of Economics and Statistics, 61, 146-49. Retrieved from https://doi.org/1010.2307/1924845.

[25] Johnston, K., Barrett, K., Ding, Y. & Wagner, D. (2009). Clinical and Imaging Data at 5 Days as a Surrogate for 90-Day Outcome in Ischemic Stroke. Stroke; a journal of cerebral circulation, 40, 1332-3. Retrieved from https://doi.org/1010.1161/STROKEAHA.108.528976