Imbalanced Learning in Assessing the Risk of Corruption in Public Administration

Marcelo Oliveira Vasconcelos[1(✉)], Ricardo Matos Chaim[2], and Luís Cavique[3]

[1] Tribunal de Contas do Distrito Federal, Brasília, Brasil
mov@tc.df.gov.br
[2] Universidade de Brasília, Brasília, Brasil
ricardc@unb.br
[3] LASIGE, Universidade Aberta, Lisboa, Portugal
luis.cavique@uab.pt

Abstract. This research aims to identify the corruption of the civil servants in the Federal District, Brazilian Public Administration. For this purpose, a predictive model was created integrating data from eight different systems and applying logistic regression to real datasets that, by their nature, present a low percentage of examples of interest in identifying patterns for machine learning, a situation defined as a class imbalance. In this study, the imbalance of classes was considered extreme at a ratio of 1:707 or, in percentage terms, 0.14% of the interest class to the population. Two possible approaches were used, balancing with resampling techniques using synthetic minority oversampling technique SMOTE and applying algorithms with specific parameterization to obtain the desired standards of the minority class without generating bias from the dominant class. The best modeling result was obtained by applying it to the second approach, generating an area value on the ROC curve of around 0.69. Based on sixty-eight features, the respective coefficients that correspond to the risk factors for corruption were found. A subset of twenty features is discussed in order to find practical utility after the discovery process.

Keywords: Data enrichment · Imbalanced learning · Corruption · Public administration

1 Introduction

Corruption in public administration is a problem that could be addressed through machine learning to identify risk factors for mitigation by the supervisory body. This research explores this scenario in a Brazilian case study.

Corruption is a common problem in developing countries (Olken 2007), leading to an increase in the cost of public services, undermining economic growth (Mauro 1995), and impairing private business conduction.

Corruption is the abuse of the power entrusted to private gain (Transparency International, n.d.), and the cost of corruption is high. Part of that cost is the fee added to the contract value by charging the public budget. Another part of public administration

G. Marreiros et al. (Eds.): EPIA 2021, LNAI 12981, pp. 510–523, 2021.
https://doi.org/10.1007/978-3-030-86230-5_40

is poor public resources management that generates low-quality public service provision. There is also the devaluation of assets and losses of national and international investments (Padula & Albuquerque 2018).

This research sought to study and apply data mining techniques to create a predictive model for assessing the risk of the corruptibility of public servants in the Federal District, considering studies carried out on corruption and consultation with specialists on the subject.

This document contributes to the corruption data enrichment, the research of two classes of algorithms in imbalanced learning, and the impact of knowledge discovery in corruption literature.

The procedure used in this work can be summarized in three steps: (i) data enrichment and data cleansing, (ii) imbalance learning models, (iii) discussion of the findings. The proposed procedure can be presented in the following data pipeline: data pre-processing → learning models → findings discussion.

The remaining of the paper is organized as follows. Section 2 describes the data enrichment and data cleansing processes. Section 3 presents two imbalanced learning approaches. Section 4 provides computational results. The discussion of the results is reported in Sect. 5. Finally, in Sect. 6, the conclusions are drawn.

2 Data Enrichment and Data Cleansing

In the data mining process, data enrichment and data cleansing are essential stages to set a data frame. This section is handling the activities of these stages.

Data enrichment enhances collected data with relevant context obtained from additional sources (Knapp & Langill 2014). Data cleansing is the process of attempting to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data (Han et al. 2012).

Data enrichment and data cleansing have activities occurring in parallel and are explained in this section.

A compilation of eight different databases from the Brazilian Federal Government and Federal District was used to compose the data frame. These data represent the information from 303,036 civil servants, militaries, and pensioners of the Federal District.

Federal District is a legal entity of internal public law, which is part of the political-administrative structure of Brazil, of a *sui generis* nature because it is neither a state nor a municipality. A unique entity accumulates the legislative powers reserved to the states and the municipalities, which gives it a hybrid nature of state and municipality.

The concept of corruption adopted in this research was described in Brazilian Law No. 8,429/92, which defines corruption as an act of improbity that, under the influence or not of the position, causes illicit enrichment, causes or not mandatory, will be used to the purse or violate Public Administration principles (Brasil 1992).

The attribute corresponding to this definition is represented by "C.CorruptionTG", which will be described with the independent attributes in Table 1.

The data obtained from these databases were outlined by their attributes classified by domains according to the following areas of knowledge: corruption (C), functional work (W), political (P), and Business (B), as defined below.

Table 1. Extract of attribute's description

Attribute name	Type	Brief description
Corruption domain (C)		
C.CorruptionTG	Boolean	Cases of dismission by corruption, this is the target
C.CEIS	Boolean	Cases of individuals or legal entities with restrictions on the right to participate in tenders or to contract with the Public Administration by sanctions
C.TCDFrestriction	Boolean	Cases of person who are not qualified to exercise a position in a commission or a trust function within the Public Administration of the Federal District for a period of up to eight years due to severe irregularities found by the TCDF
Functional work domain (W)		
W.Salary	Numeric	Salary of the civil servant or military that included the salary received by any of the databases (SIGRH and SIAPE) or the sum of salaries in the case of civil servants who accumulate public positions as permitted by the Federal Constitution
W.QtySIGRHOff	Int	Quantity of positions that the civil servant or military held until Nov/2020 into the SIGRH determined only with the SIGRH database
W.QtySIAPEOff	Int	Quantity of positions the civil servant or military held in Public Security until Nov/2020 at SIAPE (Public Security, SIAPE)
W.QtySIGRHfunc	Int	Quantity of functions that the civil servant occupied until Nov/2020 in the SIGRH (Servers, except Public Security, SIGRH)
W.QtySIAPEfunc	Int	Quantity of functions that the civil servant or military occupied until Nov/2020 in SIAPE (SIAPE Public Security)
Political domain (P)		
P.CandEducation	Categorical	Candidate's level of education can be defined as non-disclosable, reads and writers, incomplete or complete elementary school, incomplete or complete high school, and incomplete or complete higher education
P.CandMaritalSt	Categorical	The civil status situation of the candidate civil servant: single, married, non-disclosable, widowed, legally separated, or divorced
Business Domain (B)		
B.OwnershipPerc	Numeric	Percentage of share capital that the civil servant or military presents at Nov/2020
B.QtFirmAct	Int	Number of secondary activities registered by the company in which the civil servant or military is a partner

(*continued*)

Table 1. (*continued*)

Attribute name	Type	Brief description
B.CodFirmAct	Categorical	The main activity of the firm/company in which the civil servant or military is a partner
B.CodFirmSize	Categorical	Size of the company that can be Individual Micro entrepreneur (MEI), Microenterprise (ME), Small Business (EPP), medium or large depending on the gross annual turnover of the head office and its branches, or that is, the global gross revenue defined in the tax legislation
B.DaysOwnership	Numeric	This attribute informs the number of days that the server is a partner in the company until Nov/2020
B.CodFirmTaxOpt	Categorical	This attribute informs if the company opted for the simplified taxation system - Simples Nacional - which aims to help micro and small companies concerning the payment of taxes

For this research, a dataset was created after an ETL process (extract, transform and load) collected from these different data sources, as described below:

- Expulsion Registrations maintained by Comptroller General of the Federal District (Portal da Transparência DF);
- SIGRH - Integrated Resource Management System maintained by Federal District Government;
- SIAPE – Integrated Human Resources Administration System maintained by Federal Government;
- Persons that by sanction are not allowed for the exercise commission position or a trust function within the scope of the Public Administration of the Federal District maintained by TCDF;
- Private Non-Profit Entities Prevented from contracting with the Public Administration (CEPIM) maintained by Office of the Comptroller General (Controladoria-Geral da União—CGU);
- Registration of Unfaithful and Suspended Companies (CEIS) maintained by Office of the Comptroller General (Controladoria-Geral da União—CGU);
- Electoral Data maintained by Superior Electoral Court (TSE); and
- Personal and Legal Data maintained by Secretariat of the Federal Revenue of Brazil (SRF/ME).

In the present work, the dependent variable (C.CorruptionTG) is binary. In preparing the data, to obtain processed and prepared data to demonstrate the understanding of the business. Database integration work took place for the integrated Resource Management System (SIGRH) and the Integrated Human Resource Management System (SIAPE). There are two different databases for payment of the Federal District Government civil servant/military that separate Public Security servers (SIAPE) from other civil servants (Education, Health, and other areas).

Categorical attributes were transformed into binary, i.e., variables that describe categories or classifications; for binary attributes, variables with a value of 0 or 1 express the existence or absence of the binary attribute. This procedure is also known as an application for dummy variables.

Another perspective of attribute construction used was the transformation of categorical attributes into counting attributes. This procedure was performed because the attribute when expressing quantity has meaning in the context of business understanding. In contrast, the categorical value does not express benefit in the context of the investigation.

For example, a categorical attribute means that the civil servant or military man/woman occupied in Public Administration has no meaning for this investigation. However, several positions he/she had occupied could inform that this one does not have a stable condition and could represent an anomaly.

Initially, the data set comprised 28 attributes (numeric and categorical) list in Table 1, which after necessary transformations of the categorical resulted in 11 numeric attributes and 1,116 binaries attributes described in Table 2.

Moreover, to avoid bias of the numerical attributes in the algorithm, these numerical attributes were normalized as the last transformation. Finally, each value was subtracted from the lowest value of the attribute and divided by the amplitude (highest value subtracted from the lowest value of the attribute), resulting in values between zero and one.

Along with all these steps, missing values and outliers were treated properly; some attributes were built to generate relevant information for the business from the original data. After cleansing data and building attributes, analysis of variance and correlation was performed.

Regarding the assessment of correlation between variables, four attributes with Pearson's correlation above 0.9 and seven attributes with a correlation between 0.8 and 0.9 were identified and excluded.

After calculating the variance of the attributes, one of them presented a null value, 294 attributes showed a variance less than 0.00001, and 421 attributes registered a variance between 0.00001 and 0.0001.

After excluding these attributes in this condition, the dimensionality reduced from 1,127 attributes to 397 attributes.

Another measure that reduces the dimensionality was excluding predictors that cannot be concluded if there is a statistically significant association with the response variable (target), i.e., when the predictor has a p-value greater or equal to the significance level, 0.05. Excluding these attributes (predictors), the model generated was left with sixty-eight attributes, and all of the excluding attributes in the cleansing process in resume in Table 2.

Table 2. Data cleansing

	Types of attributes			
	Numeric	Categorical	Binary	Total
Original attributes	11	13	4	28
1 - Transformation of categorical attributes into binaries	11	0	1116	1127
2 - Exclusions of attributes with:				
2.1 - Correlation between > 0.8	8	0	1105	1112
2.2 - Variance < 0.0001	8	0	389	397
2.3 – p-value > 0.05	8	0	60	68

3 Imbalanced Learning

In this section, the research's theoretical for imbalanced learning is presented. In the context of actual data related to corruption or fraud, the number of examples of the interest data predominately represents a small percentage of the dataset. This characteristic is considered a class imbalance. Therefore, the class of interest is reduced concerning the dominant class.

Most machine learning algorithms assume that all misclassification errors made by a model are equal. However, it is often not the case for imbalanced classification problems. For example, missing a positive or minority class case is worse than incorrectly classifying an example from the negative or majority class. There are many real-world examples, such as detecting spam email, diagnosing a medical condition, or identifying fraud (Brownlee 2020).

Zhu et al. (Zhu et al. 2018) suggest solving imbalanced datasets by two possible solutions: data-level solutions and algorithm-level solutions. Table 3 presents a variant of the taxonomy presented in Vimalraj & Rajendran (Vimalraj & Rajendran 2018), showing the algorithms that handle imbalanced data for both methods.

Table 3. Methods and algorithms to handle imbalanced data

Data-level	**Algorithmic-level**
Over-sampling (smote)	One class learning
Under-sampling	Cost-sensitive learning
Feature selection	Logistic regression

The data-level solutions are resampling data as a pre-processing step to reduce the negative effect caused by class imbalance.

Two methods are considered usual to minimize class imbalance in the pre-processing data phase: under-sampling and over-sampling.

The first deals with the random exclusion of observations from the majority class, while the second deals with the multiple creations of copies of observations from the minority class. However, both methods have disadvantages. For example, undersampling can discard potentially useful data instances, while oversampling can increase the probability of overfitting, which corresponds to the occurrence of a statistical model very well-adjusted to the set of data previously observed, but proves inefficient to predict new results.

A specific technique was created to minimize the effects of the previous techniques – synthetic minority oversampling technique (SMOTE), covered in detail in the next item.

The algorithm-level solutions aim to develop new algorithms or modify existing ones to deal with imbalanced datasets. For example, in Brownlee (Brownlee 2020), the author presents two approaches for modifying algorithms in Logistic Regression to apply for imbalanced classes: Weighted Logistic Regression Algorithm and Heuristic implementation for Logistic Regression.

3.1 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE technique was presented by Chawla (Chawla et al. 2002). SMOTE combines the oversampling method of the minority class (abnormal), under-sampling of the majority class (standard), and the creation of synthetic examples of the minority class. As a result, this new dataset can better perform the classifier (in the ROC space) than merely sub-sampling of the majority class.

This technique is widely used. An indicator of this fact is that the SMOTE article (Chawla et al. 2002) was cited more than 6,300 according to Web of Science. Furthermore, after 15 years of the publication of this paper, more than 85 extensions of SMOTE have been proposed by specialized literature. (Alberto Fernandez et al. 2018).

3.2 Logistic Regression

Logistic Regression is a method developed under the leadership of the statistician Ronald Fisher. It involves estimating parameter β of a probability distribution of random variable X with a certain number of independent observations.

Logistic Regression is usual in situations where the dependent variable is of a binary or dichotomous nature, while the independent variables can be categorical or not. The Logistic Regression seeks to estimate the probability of a given event concerning a set of variables that explain the phenomenon.

In Logistic Regression, the probability of an event occurring can be estimated directly. The dependent variable can assume two states (0 or 1), and there is a set of p independent variables X1, X2,…, Xp according to the following equation.

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}} \tag{1}$$

Where g (x) = $B_0 + B_1X_1 + \ldots + B_p X_p$. The coefficients B_0, B_1,… B_p are estimated from the dataset, using the maximum likelihood method, which determines the combination of coefficients that maximizes the probability.

The classic references of Logistic Regression are Cox & Snell and Hosmer & Lemeshow (Hosmer & Lemeshow 1999). Positive values for Logistic Regression coefficients represent an increase in the probability, i.e., a negative decrease in the probability.

An important concept was presented by Mandrekar (Mandrekar 2010), which defined as ROC curve a sensibility graph versus a test specificity, and the area above this curve expresses the model's measure performance. In this study, it is established that an area above the ROC – AUC curve (*Area Under Curve – AUC*) of 0.5 suggests no discrimination; 0.7 to 0.8 is considered acceptable; 0.8 to 0.9 is considered excellent, and over 0.9 is considered exceptional.

Logistic Regression is a powerful classifier by providing probabilities and by extending to multi-class classification problems. The advantages of using Logist Regression are that it has been extensively studied, and it does not make assumptions about the distribution of the independent variables (Maalouf & Siddiqi 2014).

In an imbalanced dataset context, there are implementations of Logistic Regression for algorithm-level solutions (Maalouf & Siddiqi 2014)(Maalouf & Trafalis 2011) (Brownlee 2020), and for data-level solutions (Torgo et al. 2013) (Brownlee 2020).

4 Computational Results

In this research, after pre-processing the dataset, two approaches were applied for Logistic Regression to obtain better performance: data-level solutions and algorithm-level solutions. In this section, both solutions will be detailed.

According to Brownlee (Brownlee 2020), an extreme imbalance is challenging for modeling that requires specialized techniques. Therefore, in this investigation, the two possibilities of data imbalance treatment were addressed.

4.1 Data-Level Solutions

This approach is the treatment of data to make the classes balanced. In the literature, the technique widely used for imbalanced data is SMOTE and its variations/extensions, as explained in item 3.2.

The values obtained as an application of the SMOTE technique and extensions are summarized in Table 4.

4.2 Algorithm-Level Solutions

In a different approach, without applying the sampling method, the Logistic Regression algorithm was employed with specific characteristics to deal with the imbalanced data. (Brownlee 2020) The performance was obtained concerning the area of the ROC curve (AUC).

Initials tests were executed using different machine learning techniques, but the low performance on AUC takes to abandon these approaches. For the Decision Tree, the result for AUC was 0.578. Applying Support Vector Machine, the AUC was 0.621. The neural network reveals the worst value, 0.500, and the better performance in this

Table 4. Data-Level Solutions - Results of the area on the ROC curve with SMOTE

Techniques	Area under ROC curve (AUC)	Package
SMOTE e random Under-sampling	0.658	from imblearn.over_sampling & from imblearn.under_sampling (Python)
SMOTE and Tomek Links sampling	0.534	imblearn.combine & imblearn.under_sampling (Python)
SMOTEENN	0.601	imblearn.combine (Python)
SMOTE	0.422	SmoteFamily (R)
DBSMOTE	0.534	SmoteFamily (R)
ADAS	0.548	SmoteFamily (R)
ANS	0.534	SmoteFamily (R)
SLS	0.491	SmoteFamily (R)

Source: Smote Family documentation available in https://cran.r-project.org/web/packages/smotefamily/smotefamily.pdf

test was for Logistic Regression, at least 0.647. All of the tests were applied by the same technique - Predefined Weights, and this type of approach will be detailed in the following paragraphs.

Two models were applied. The first was defining weight for the different classes, which is the proportion of cases of the minority class concerning the majority class.

In this approach, the focus is on modifying the classifier learning procedure.

An important property has to be addressed by the algorithm. Not all classification errors are equal. For this research, a false negative is worse or more costly than a false positive.

This issue is settled by cost-sensitive learning that takes the costs of prediction errors into account when training a machine learning model.

In cost-sensitive learning, each class is given a misclassification cost instead of each instance being either correctly or incorrectly classified. Thus, instead of optimizing the accuracy, the problem is then to minimize the total misclassification cost. Thus, a penalty associated with an incorrect prediction is named cost.

Weighted Logistic Regression implements cost-sensitive learning for Logistic Regression in Python (library Scikit-Learn), supporting class weighting.

This relation in target attribute (C.CorruptionTG) was 428 cases True and 302,608 cases False, the ratio value is $428/302{,}608 = 0.0014$. Adopting this value was employed as a class weight for the weighted Logistic Regression algorithm.t

The second model was a heuristic implementation of best practices used in the weighting of classes available in the library Scikit-Learn implemented in Python.

In this model, the weight assigned is dividing the population quantity by the product of the number of classes by the population quantity of the majority class, and the algorithm calculates it.

Both implementations were for Logistic Regression (LOGIT) through the Scikit-Learn library implemented in Python.

The results were calculated by averages of the AUC curve calculated using cross-validation with ten folders. That is, the partitions were remade three times to represent the most appropriate value for measurement.

For addressed adequately, the imbalanced classes were used as a library resource to ensure that the cross-validation partitions contain proportional samples from the minority class that is the class of interest in the research. Table 5 shows a summary of the algorithm-level computational results.

Table 5. Algorithm-level solutions

Model	Area under ROC curve (AUC)
I - Predefined weights	0.692
II - Heuristic weights	0.647

The best result for algorithm-level solutions was the Predefined Weights with a fixed weight assignment of 0.0014 according to the proportion of True and False cases of the target attribute.

The values obtained as an application of the Data-level solutions by applying SMOTE technique were not better than applying the Algorithm-level Solution. For example, the value range for AUC was from 0.491 to 0.658.

From this modeling result of Predefined Weights, the coefficients of the logistic regression attributes were established.

5 Discussion

The last section has presented the results of different approaches, and the best result, Predefined Weights, was used for modeling. Finally, the coefficients of the Logistic Regression attributes were established, and Table 6 lists the main attributes.

The coefficient ($b_{\#}$) is the estimated increase in the natural logarithm odds of the outcome per unit increase in the exposure value. In other words, the exponential function of the regression coefficient ($e^{b\#}$) is the odds ratio associated with a one-unit increase in exposure. (Szumilas 2010).

$P(Y = 1) = \frac{1}{1+e^{-g(x)}}$, Where $g\ (x) = -4.07 + 1.606\ X_1 + \ldots - 1.230\ X_{68}$.

The value of the Intercept is - 4.07, so, for $x = 0$; $g(0) = -4.07$; $P(Y = 1) = 0.01679$, i.e., the meaning that the target outcome (e.g., a correct response - corrupt) was about 1.68%.

Observing the relatively low value of the Intercept constant and seeing several attributes with positive coefficients, we see that the model operates with a low risk of corruption increased by the attributes with the highest coefficient.

Table 6. Main attributes and coefficients of logistic regression in descending order

A#	Attribute	Coeff. ($b_{\#}$)	e(b#)	A#	Attribute	Coeff. ($b_{\#}$)	$e^{(b\#)}$
A1	W.QtySIGRHfunc	1.606	4.987	A13	B.CodFirmAct accounting	0.539	1.715
A2	C.CEIS	1.124	3.079	A14	P.CandMaritalSt.1	0.533	1.705
A3	B.CodFirmAct HighEducInst	0.827	2.287	A15	P.CandEducation.6	0.531	1.701
A4	B.CodFirmAct ClinicalLab	0.725	2.065				
A5	W.QtySIAPEOff	0.709	2.033				
A6	B.CodFirmAct legal Services	0.673	1.961	A62	W.Salary	(0.129)	0.879
A7	B.CodFirmAct furniture trade	0.657	1.929	A63	C.TCDFrestriction	(0.142)	0.867
A8	B.OwnershipPerc	0.625	1.869	A64	B.CodFirmSize.1	(0.224)	0.798
A9	B.CodFirmAct souvenir trade	0.615	1.850	A65	B.QtFirmAct	(0.229)	0.7951
A10	B.CodFirmTaxOpt.6	0.594	1.811	A66	B.DaysOwnership	(0.291)	0.746
A11	B.CodFirmAct technical	0.565	1.760	A67	W.QtySIAPEfunc	(0.363)	0.695
A12	B.CodFirmAct book edition	0.561	1.753	A68	W.QtySIGRHOff	(1.230)	0.292

However, an attribute with a negative coefficient, such as F.DaysOwnership, indicates that the lack of this attribute, the corporate bond of a civil servant/military, reduces corruption risk.

The simplest way to interpret the Logistic Regression coefficient is to understand that the e(x), odds ratio, represents the proportion of increasing or decreasing the attribute related to the target (C.CorruptionTG).

The attribute "W.QtySIGRHfunc"[A1] presents e(x) = 4.9870. If possible, to increase the target (binary attribute) for a specific value less than one, the attribute "W.QtySIGRHfunc" most increased by 498,7% in a proportion way. This idea is not precise due to the impossibility of the target increased by different values. However, it could explain how to interpret the behavior of the attributes related to the target.

Next, the main rules obtained by analyzing the final model's attributes and coefficients will be outlined. The effects were divided into an increased or decreased risk of corruption and increased or decreased probability.

The highest risk scenario would represent civil servants/military with the following characteristics. With the highest risk of corruption, the following characteristics were listed, whether cumulative or not:

- A civil servant with several functions changes, excluding Public Security workers (policemen and firefighters) (W.QtySIGRHfunc [A1]);
- Civil servant partner of a company present on the list of Registration of Unfaithful and Suspended Companies (C.CEIS [A2]) or that has a high percentage of share capital or that the company has cadastral activity in the specific areas of higher education institution [A3], clinical laboratories [A4], legal services [A6], furniture [A9] and souvenir trade [A10] technical [A11] and accounting activities [A13], and, book edition [A12];
- Civil servant that was a candidate for elective office (political office) [A14] e [A15].On the other hand, the scenario for civil servant/military with a low risk of corruption could be composed by these attributes:
- A civil servant with several positions changes, excluding Public Security workers (police and firefighters) (W.QtySIGRHOff [A68]);
- Policemen and firefighters (Public Security Workers) with several function changes positions (W.QtySIAPEfunc [A67]).

6 Conclusions

Data enrichment and data cleansing were applied in this study by integrating eight different databases that could be delt into four domains. It started with 28 attributes, some of them were transformed to meet business needs, and the final 68 attributes were set for the data mining process.

As a result, the dataset represents extremely imbalanced classes that could be challenging to create a better model at a ratio of 1:707 or, in percentage terms, 0.14% of the interest class to the population.

In the process of reducing the dimensionality without loss performance (Area under ROC), attributes with a correlation greater than 0.8, variances less than 0.0001, and with a p-value greater than 0.05 were excluded. As a result, the dimensionality reduces from 1,127 attributes to 68 attributes.

Two approaches were utilized: data-level solutions by SMOTE and extensions, and algorithm-level solutions, by Predefined Weights and Heuristic Weights. In this case, applying an algorithm-level solution resulted in a better performance than a data-level solution (area value on the ROC curve of around 0.69).

The impact of knowledge discovery in corruption literature was the Logistic Regression coefficients representing the risk factor for each attribute concerning the possibility of corruption. The numerical representativeness of this coefficient is related to the response variable or target attribute of the investigation.

These identified risk factors for corruption can assist in the definition of overseen planning on the most significant risk for Public Administration, so cases with a high probability of occurrence and a high financial or social impact.

It is difficult to obtain similarly published works with machine learning applications in corruption or fraud, possibly because of resistance from those who work with this activity. Furthermore, the high resilience and dynamics of fraudsters can be helped by the available publications.

Acknowledgments. L.Cavique would like to thank the FCT Projects of Scientific Research and Technological Development in Data Science and Artificial Intelligence in Public Administration, 2018–2022 (DSAIPA/DS/0039/2018), for its support.

References

Fernandez, A., Garcia, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15th anniversary. J. Artif. Intell. Res. **61**, 863–905 (2018)

Brasil: Lei nº 8429, de 2 de julho de 1992, DOU (1992). http://www.planalto.gov.br/ccivil_03/leis/l8429.htm

Brownlee, J.: Imbalanced Classification with Python Choose Better Metrics, Balance Skewed Classes, and Apply Cost-Sensitive Learning. Machine Learning Mastery, vol. V1.2, pp. 1–22 (2020)

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**(1), 321–357 (2002). https://doi.org/10.1613/jair.953

Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Kaufmann, M., (ed.), 3 (2012)

Hosmer, D., Lemeshow, S.: Applied Survival Analysis - Regression Modeling of Time to Event Data, John Wiley, New York, pp. 386 (1999)

Knapp, E.D., Langill, J.T. Industrial network security: securing critical infrastructure networks for smart grid, SCADA, and other industrial control systems. In: Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and Other Industrial Control Systems, Second Edition (2014). https://doi.org/10.1016/B978-0-12-420114-9.00018-6

Maalouf, M., Siddiqi, M.: Weighted logistic regression for large-scale imbalanced and rare events data. Knowl.-Based Syst. **59**, 142–148 (2014). https://doi.org/10.1016/j.knosys.2014.01.012

Maalouf, M., Trafalis, T.B.: Robust weighted kernel logistic regression in imbalanced and rare events data. Comput. Stat. Data Anal. **55**(1), 168–183 (2011). https://doi.org/10.1016/j.csda.2010.06.014

Mandrekar, J.N.: Receiver operating characteristic curve in diagnostic test assessment. J. Thorac. Oncol. **5**(9), 1315–1316 (2010). https://doi.org/10.1097/JTO.0b013e3181ec173d

Mauro, P.: Corruption and growth. Source: Q. J. Econ. **110**(3), 681–712 (1995)

Olken, B.A.: Monitoring corruption : evidence from a field experiment in Indonesia. J. Polit. Econ. **115**(2), 200–249 (2007)

Padula, A.J.A., Albuquerque, P.H.M.: Government corruption on Brazilian capital markets: a study on Lava Jato (Car Wash) investigation. Revista de Administração de Empresas **58**(4), 405–417 (2018). https://doi.org/10.1590/S0034-759020180406

Szumilas, M.: Explaining odds ratios. J. Can. Acad. Child. Adolesc. Psychiatry, **341**(19:3), 227–229 (2010). https://doi.org/10.1136/bmj.c4414

Torgo, L., Ribeiro, R.P., Pfahringer, B., Branco, P.: SMOTE for regression. In: Correia, L., Reis, L.P., Cascalho, J. (eds.) EPIA 2013. LNCS (LNAI), vol. 8154, pp. 378–389. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40669-0_33

Transparency International. (n.d.).: Transparency International - What is Corruption?, 16 June 2019. https://www.transparency.org/what-is-corruption

Vimalraj, S., Rajendran, P.: A review on handling imbalanced data. In: International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1–11. IEEE (2018)
Zhu, B., Baesens, B., Backiel, A., Vanden Broucke, S.K.L.M.: Benchmarking sampling techniques for imbalance learning in churn prediction. J. Oper. Res. Soc. **69**(1), 49–65 (2018). https://doi.org/10.1057/s41274-016-0176-1