Department of Economics - FEA/USP

Applying Machine Learning Algorithms to Predict the Size of the Informal Economy

JOÃO FELIX

MICHEL ALEXANDRE

GILBERTO TADEU LIMA

WORKING PAPER SERIES Nº 2023-10

DEPARTMENT OF ECONOMICS, FEA-USP
WORKING PAPER Nº 2023-10

Applying Machine Learning Algorithms to Predict the Size of the Informal Economy

João Felix (joaosilvafelix19@gmail.com)

Michel Alexandre (michel.alsilva@gmail.com)

Gilberto Tadeu Lima (giltadeu@usp.br)

Abstract:

The use of machine learning models and techniques to predict economic variables has been growing lately, motivated by their better performance when compared to that of linear models. Although linear models have the advantage of considerable interpretive power, efforts have intensified in recent years to make machine learning models more interpretable. In this paper, tests are conducted to determine whether models based on machine learning algorithms have better performance relative to that of linear models for predicting the size of the informal economy. The paper also explores whether the determinants of such size detected as the most important by machine learning models are the same as those detected in the literature based on traditional linear models. For this purpose, observations were collected and processed for 122 countries from 2004 to 2014. Next, eleven models (four linear and seven based on machine learning algorithms) were used to predict the size of the informal economy in these countries. The relative importance of the predictive variables in determining the results yielded by the machine learning algorithms was calculated using Shapley values. The results suggest that (i) models based on machine learning algorithms have better predictive performance than that of linear models and (ii) the main determinants detected through the Shapley values coincide with those detected in the literature using traditional linear models.

Keywords: Informal economy; machine learning; linear models; Shapley values.

JEL Codes: C52; C53; O17.

Applying machine learning algorithms to predict the size of the informal economy

João Felix[1†], Michel Alexandre[2*†], Gilberto Tadeu Lima[2†]

[1*]Department of Economics, Federal University of Paraíba, Via Ipê Amarelo, João Pessoa, 58059-356, PB, Brazil.
[2]Department of Economics, University of São Paulo, Av. Prof. Luciano Gualberto 908, São Paulo, 05508-210, SP, Brazil.

*Corresponding author(s). E-mail(s): michel.alsilva@gmail.com;
Contributing authors: joaosilvafelix19@gmail.com; giltadeu@usp.br;
†These authors contributed equally to this work.

Abstract

The use of machine learning models and techniques to predict economic variables has been growing lately, motivated by their better performance when compared to that of linear models. Although linear models have the advantage of considerable interpretive power, efforts have intensified in recent years to make machine learning models more interpretable. In this paper, tests are conducted to determine whether models based on machine learning algorithms have better performance relative to that of linear models for predicting the size of the informal economy. The paper also explores whether the determinants of such size detected as the most important by machine learning models are the same as those detected in the literature based on traditional linear models. For this purpose, observations were collected and processed for 122 countries from 2004 to 2014. Next, eleven models (four linear and seven based on machine learning algorithms) were used to predict the size of the informal economy in these countries. The relative importance of the predictive variables in determining the results yielded by the machine learning algorithms was calculated using Shapley values. The results suggest that (i) models based on machine learning algorithms have better predictive performance than that of linear models and (ii) the main determinants detected through the Shapley values coincide with those detected in the literature using traditional linear models.

Keywords: informal economy, machine learning, linear models, Shapley values

1 Introduction

Informal economies—also referred to as informality, underground economies, hidden economies, or even shadow economies—play an important role in several countries. Despite being a segment that provides employment opportunities for many people who for different (and sometimes involuntary) reasons are unable to enter the formal labor market, the informal economy is also a source of problems. This is because, for example, the informal economy can distort the degree of competition in markets where it is significant (Ranis & Stewart, 1999; Ulyssea, 2020) and, by definition, reduces tax collection (Schneider, Raczkowski, & Mróz, 2015; Ulyssea, 2020; Vousinas, 2017), harming the macroeconomic environment of a country.

In Brazil, for example, Alm and Embaye (2013) estimated that for the period between 1984 and 2006, the average size of the informal economy was 38.76% of that of the formal economy. Medina and Schneider (2018), for the period between 2004 and 2015, obtained a very similar average value of 37.63% for the size of the Brazilian informal economy. These figures illustrate how relevant this segment of economic activity is for developing countries such as Brazil. In fact, there is evidence that formal and informal firms coexist in Brazil even in narrowly defined sectors (Ulyssea, 2018). Therefore, it is of considerable importance to understand the main determinants of the existence of informal economies, as well as to make better predictions about their size. Currently, linear regression models are widely used for this purpose. However, despite being very useful for the analysis of causal inference, they are sometimes criticized for not being efficient models for the task of predicting nonlinear datasets.

Currently, different machine learning (ML) models are being increasingly used for prediction purposes in the most diverse areas of knowledge (Dabiri, Kheyroddin, & Faramarzi, 2022; Gambhir, Jain, Gupta, & Tomer, 2020; Goldstein, Navar, & Carter, 2017). Studies have shown that methods based on ML algorithms are highly promising alternatives for problems related to forecasting different economic variables. Therefore, these models are a very interesting alternative for a more adequate prediction of the size of the informal economy. In fact, ML algorithms have already been used to forecast several economic variables, such as the real gross domestic product (GDP).Yoon (2021), for example, used the gradient boosting and random trees models to forecast real GDP growth in Japan between 2001 and 2018 and to compare the performance of these models to that of the models used by the International Monetary Fund (IMF) and the Bank of Japan. Employing the mean absolute percentage error (MAPE), Yoon (2021) found that the random trees model and, especially, the gradient boosting model outperformed those institutional models used as a reference for comparison.

On the other hand, although ML methods perform better than traditional econometric techniques regarding prediction, they have often been criticized for their black-box nature, i.e., it is not easy to determine and interpret how such a prediction was made. To overcome this limitation and potential problem, several techniques have been developed with the purpose of increasing the interpretability of ML models, showing the importance of predictive variables in determining the prediction result of the model. Examples of these techniques include Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) and Shapley Additive Explanations (SHAP) (Lundberg & Lee, 2017).

In addition to the above-mentioned methodologies, some ML models have also been used for causal inference, as in Guo, Huang, Wang, and Wang (2022). The authors explored the initial impacts of COVID-19 on offline micro business merchants in the informal sector in the initial periods of stricter control in China due to the virus outbreak, employing a decision tree model (a gradient boosting decision tree). The authors found that offline micro business merchants had a drop in their activities that reached a peak of around 50% between December 31, 2019 and April 2, 2020.

Therefore, even for problems where the main purpose is to predict the value of a particular target variable, understanding how the dependent variable relates to the independent variables is useful for the analysis of the problem at hand. Some notable attributes cited in the literature as the main determinants of the size of the informal economy include tax evasion, institutional factors (such as corruption, institutional strength, and democracy), international trade, country income, and bureaucratic issues.

Schneider and Klinglmair (2004), analyzing data from 110 countries, found that the main determinants of the informal economy are the increase in the tax burden and social security contributions, as well as the increase in regulatory activities. In another study using data from 28 European Union countries in the years between 2003 and 2014, Schneider et al. (2015) found that the main determinants of the informal economy are unemployment and self-employment together with the willingness of individuals to comply with their tax obligations (tax morale). On the other hand, Goel and Nelson (2016) observed that bureaucratic variables tend to be more significant than tax severity, although both are positively related to the size of the informal economy.

Ivașcu and Ștefoni (2023) used algorithms based on ML and linear models to investigate the relationship between the size of the informal economy and the most important government expenditures (social protection, health, and education) using data from 28 European Union countries between 1995 and 2020. The authors employed four different ML models (Support Vector Regression, Neural Networks, Random Forest, and XGBoost) and found that these non-linear models performed better (in terms of R^2 and root mean squared error) at explaining the variation in the size of the informal economy when compared to linear models. When analyzing the relationship between internet use and the size of the informal economy, Elgin (2013), found that both variables are strongly related to the GDP per capita. Countries with higher incomes had, on average, smaller informal economies. A similar relationship between the size of the informal economy and the GDP per capita was found by Lyulyov et al. (2021) and Zhanabekov (2022).

Regarding international trade, Canh and Dinh Thanh (2020) observed that trade liberalization has a negative effect on the size of the informal economy, suggesting that an increase in trade liberalization tends to decrease the informal sector. The authors also found that the quality and diversity of exports tend to influence the size of the informal economy. However, a nonlinear relationship was found when using the quadratic form of these same variables. Regarding trade liberalization, a similar result was found by Elbahnasawy (2021) and Zhanabekov (2022). The degree of democracy also appears to be an important determinant of the size of the informal economy.

Teobaldelli and Schneider (2013) found that as direct democracy gains force in the decision-making process, fiscal policies better reflect the preferences of citizens, thus reducing their incentive to operate in the informal sector. The same effect of democracy on the size of the informal economy was found by Elbahnasawy (2021).

Due to the enormous importance of the informal segment of economic activity, in this paper, tests were conducted to determine whether models based on ML algorithms perform better than linear models that have been used for the same purpose. In addition, tests were conducted to determine whether the most important attributes indicated by ML models as the main determinants of the size of the informal economy are the same as those detected in the literature based on linear models. For this purpose, observations were collected and properly treated regarding the size of the informal economy as a proportion of the GDP for 122 countries from 2004 to 2014 together with observations for 15 potential predictive variables. This large and broad set of observations was then used to determine which of the eleven linear and ML-based models had the best predictive performance. Finally, the Shapley values were calculated according to the SHAP technique (Lundberg & Lee, 2017) in the case of ML models to measure the importance of each attribute in determining the predictions that were made.

Our results show that ML models perform significantly better than linear models in terms of predicting the size of the informal economy. In addition, the main determinants detected by the SHAP technique, such as the per capita income, degree of democracy, and bureaucratic elements, coincide with those detected in the literature using traditional linear models. Our paper contributes to the scarce literature on the application of ML techniques to the prediction of the size of the informal economy. To the best of our knowledge, there are only two papers on this topic, Shami and Lazebnik (2023) and Ivașcu and Ștefoni (2023). In relation to these papers, we compare the performance of a higher number of linear and ML-based models (four and seven, respectively). Shami and Lazebnik (2023) compare a Random Forest model to a linear regression model and Ivașcu and Ștefoni (2023) compare four ML-based models against one linear model. Moreover, we apply the Shapley values methodology to measure the importance of the determinants in predicting the target variable, which further increases the interpretability of our estimates. Our results corroborate the findings of these previous studies, according to which ML models are better predictors of the size of the informal economy than linear models.

In addition to this introduction, this paper consists of three additional sections. The data used and the methodological issues involved are described and clarified in Section 2. In Section 3, the various results obtained are presented and discussed. Finally, concluding remarks are presented in Section 4.

2 Methodology and data

2.1 Data

Data regarding the target variable were the same as those used in Medina and Schneider (2018). The authors estimated the size of the informal economy as a proportion of the GDP for more than 158 countries from 1991 to 2015 using the Multiple Causes

Multiple Indicators (MIMIC) model. The predictive variables were the attributes used by Goel and Nelson (2016) to analyze the determinants of the size of the informal economy: inflation, unemployment, trade liberalization, foreign direct investment (net inflow), final consumption expenditure of the general government, start-up procedures to register a business, cost of business start-up procedures, time required to start a business, time required to register property, time to prepare and pay taxes, democracy index, tax burden and quality and diversity of exports. In addition to these variables, the GDP per capita (on a logarithmic scale) was included.

More specifically, the predictive variables were inflation (GDP deflator, % annual), total unemployment (% of total labor force, national estimate), trade liberalization (referred to as exchange in this paper, % of GDP), foreign direct investment (referred to as FDI in this paper, % of GDP), general government final consumption expenditure (referred to as government expenditure in this paper, % of GDP), start-up procedures to register a business (number), cost of business start-up procedures (% of gross national income (GNI) per capita), time required to start a business (days), time required to register property (days), time to prepare and pay taxes (hours) and GDP per capita (current US$) were collected through the World Bank's *World Development Indicators* (WDI). The tax burden was obtained using *The Heritage Foundation*, the democracy index was collected using *Polity5: Regime Authority Characteristics and Transitions*, and finally, the quality and diversity of exports were collected from the statistics database of the directorate of trade of the International Monetary Fund (IMF). The data source and the codes/variable names are provided in Table A3 in the Appendix.

All the attributes mentioned above are measured annually and differ in their availability over time, leading to a high proportion of missing observations. Therefore, the period chosen for analysis was between 2004 and 2014. For the time span of the variables, all attributes were collected in a single dataset. It is noteworthy that the data from the World Bank, Heritage Foundation, Polity5, and IMF are for several countries over time and were collected independently. Therefore, data from some countries were not available from all data sources. Data from countries that were not represented in all series were removed from the final dataset. Data from countries that did not have at least one observation in any of the attributes were also removed. Finally, data from 122 countries were included in the dataset. The list of countries can be found in Table A2 of the Appendix.

Table 1 shows the descriptive statistics for each attribute used to predict the size of the informal economy, as well as the target variable of the study. As will be explained below, the variable "democracy" was categorized using the *one-hot encoding* technique and therefore does not appear in this table.

Table 1: Descriptive statistics

Variable	No.	Minimum	Mean	Median	Maximum	Standard deviation
GDP Per Capita (in US$)	1342	128.53	14711.48	5175.19	123678.70	20684.00
Quality	1342	0.24	0.83	0.86	0.86	0.16
Diversity	1342	1.42	3.20	2.96	6.33	1.20
Tax burden	1342	32.00	74.90	77.00	99.94	12.78
Unemployment	1342	0.20	7.47	6.35	35.46	5.18
Inflation	1342	-24.21	6.98	4.81	95.40	8.82
Exchange	1342	22.10	85.55	74.60	437.32	49.20
Governement spending	1342	2.04	15.33	14.88	43.48	5.26
Tribute Time	1342	27.00	307.68	252.88	2600.00	290.52
Business Procedure	1342	1.00	8.78	9.00	20.00	3.48
Business Time	1342	0.50	36.89	22.00	697.00	61.35
Cost Procedures	1342	0.00	48.87	14.80	1491.00	120.25
Ownership Time	1342	1.00	58.62	36.00	690.00	70.41
FDI	1342	-57.53	5.59	3.04	279.36	14.88
Shadow Economy	1342	6.16	28.73	28.78	69.08	13.13

As shown in Table 1, there are 1342 observations for each attribute. As explained below, imputation was performed for the missing data. The average informal economy, as a proportion of the GDP, was 28.73% for the period under analysis, with a minimum of 6.16% and a maximum of 69.08%.

After the data merging stage, the missing observations were treated, and the democracy variable was categorized. Missing data were imputed by the mean of the corresponding attribute for each country individually. For example, if there was a missing observation for attribute x of country i, the average of this attribute would be calculated using only data from this country over time as a reference, rather than considering the average of this attribute over all other countries. Regarding the democracy variable, although the attribute was numerical, each number represents a stage of democracy. Thus, it was categorized by creating dummy variables for each stage using the *one-hot encoding* technique (one if the democracy stage corresponds to that value, and zero otherwise).

Table 2 explains how this variable was categorized. The column named "Democracy stage" indicates the stage of democracy of a country, as classified by Polity5.[1] As indicated in the table above, some stages share the same description, as is the case for open anocracy. To avoid creating dummy variables in excess, stages with the same description were allocated to the same dummy variable. After treating the missing data and categorizing the democracy variables, the final dataset contained 22 variables (one dependent variable and 21 predictive variables).

Table 2: Stage of democracy

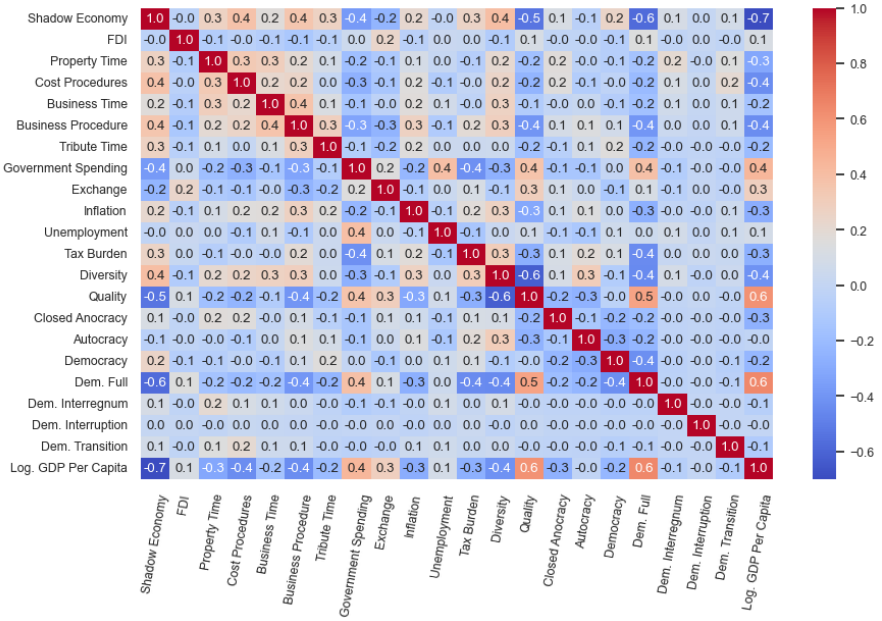
Democracy stage	**Description**	**Dummy**
-88	Transtion	Dem. Transition
-77	Interregnum	Dem. Interregnum
-66	Interruption	Dem. Interruption
0	Autocracy	Autocracy
1	Closed Anocracy	Closed Anocracy
2	Closed Anocracy	
3	Open Anocracy	Open Anocracy
4	Open Anocracy	
5	Open Anocracy	
6	Open Anocracy	
7	Democracy	Democracy
8	Democracy	
9	Democracy	
10	Full democracy	Dem. Full

Figure 1 describes how the size of the informal economy correlates with the other variables. This figure shows that the GDP per capita, full democracy, and quality

[1]The Polity5 project codifies the characteristics of state authorities to perform comparative and quantitative analyses for most independent states between 1800 and 2018.

are the variables that have a stronger negative correlation with the target variable. In turn, bureaucratic attributes such as business time, cost procedures, and tribute time are positively correlated with the size of the informal sector. In addition to these variables, diversity has a slightly stronger negative correlation.

Fig. 1: Correlation chart



After the completion of the entire process, all independent variables were normalized within the range 0 and 1. The variable x was replaced by $z = (x - x_{min})/(x_{max} - x_{min})$, where x_{min} (x_{max}) is the minimum (maximum) value of x.

2.2 Algorithms

Eleven linear and ML-based forecasting algorithms were chosen to predict the size of the informal economy for each country. The models based on ML algorithms are Random Forest, Support Vector Regression (SVR), Bagging, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), and Stacking. The four linear models are Linear Regression, Lasso, Ridge, and Elastic Net.

The Linear Regression model is one of the more traditional models used for the prediction task through regression, establishing a linear relationship between the target variable and the set of independent variables. The Lasso, Ridge, and Elastic Net models are regularized linear models that utilize weight constraints to reduce overfitting through regularization (Géron, 2022).

The Random Forest model is a decision tree-based model that uses a combination of random trees to obtain more robust predictions. The SVR model (Smola & Schölkopf, 2004) is an adaptation to regression problems of the Support Vector Machine (SVM) model, which has been widely applied in classification problems. The Bagging model is an ensemble method that generates multiple versions of predictors and uses them to create an aggregate predictor (Breiman, 1996). The Stacking model follows a logic similar to that of the Bagging method. In this approach, a meta-model combines the predictions of several individually trained ML models (the base models) and creates a general model from them for prediction, as described by Wolpert (1992). In our study, the base models are Linear Regression, Random Forest, and Gradient Boosting. The Linear Regression is used as meta-model.

The XGBoost model (Chen & Guestrin, 2016) combines decision trees ("weak models") to build stronger models. The LightGBM (Ke et al., 2017) and XGBoost models are based on gradient boosting, known for their ability to quickly adapt to the training data. Compared with other models, these models were developed to improve the performance based on gradient boosting. Finally, the CatBoost model (Dorogush, Ershov, & Gulin, 2018)) is an algorithm developed to automatically deal with categorical attributes and, like the other models described above, also deals with regression and classification problems.

2.3 Validation and evaluation of algorithms

To obtain the best possible prediction of the size of the informal economy, before implementing any of the models, the dataset used was separated into training and testing datasets, with 20% of the data (269 observations) allocated for testing and 80% (1073 observations) for training. The random seed was equal to 0. The training data were used for the development of the model, and the test data were used to evaluate the performance of the model. In addition, all models were validated using k-fold cross-validation with 5 folds.

The root mean square error (RMSE) and the coefficient of determination (R^2) were used as performance measures, and their algebraic expressions are given by

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

and

$$R^2 = 1 - \frac{SSE}{SST}, \tag{2}$$

where $y_i$is the size of the informal economy, $\hat{y}_i$ is the value predicted by the prediction algorithm, SSE is the sum of squared errors and SST is the sum of total squares.

Table A1 shows the hyperparameters used for the optimization of the chosen models and the search values that were tested for each hyperparameter. The method to perform the optimization was the *RandomizedSearchCV*. For the Stacking model, the models used were Random Forest, Gradient Boosting, and Linear Regression.

3 Results

3.1 Performance of the algorithms

Table 3 shows the RMSE and R^2 values of the models with the test data. As shown in this table, the linear models (Linear Regression, Lasso, Ridge, and Elastic Net) had higher RMSE values and lower coefficient of determination values. On the other hand, the ML models performed significantly better, with RMSE values below 4.5 and R^2 values greater than 90%. The only exception was SVR, which was not as superior to the linear models as the other ML-based models according to those statistical measures. The same relative performance of the SVR algorithm has been found in Ivașcu and Ștefoni (2023).

Table 3: Models' performance

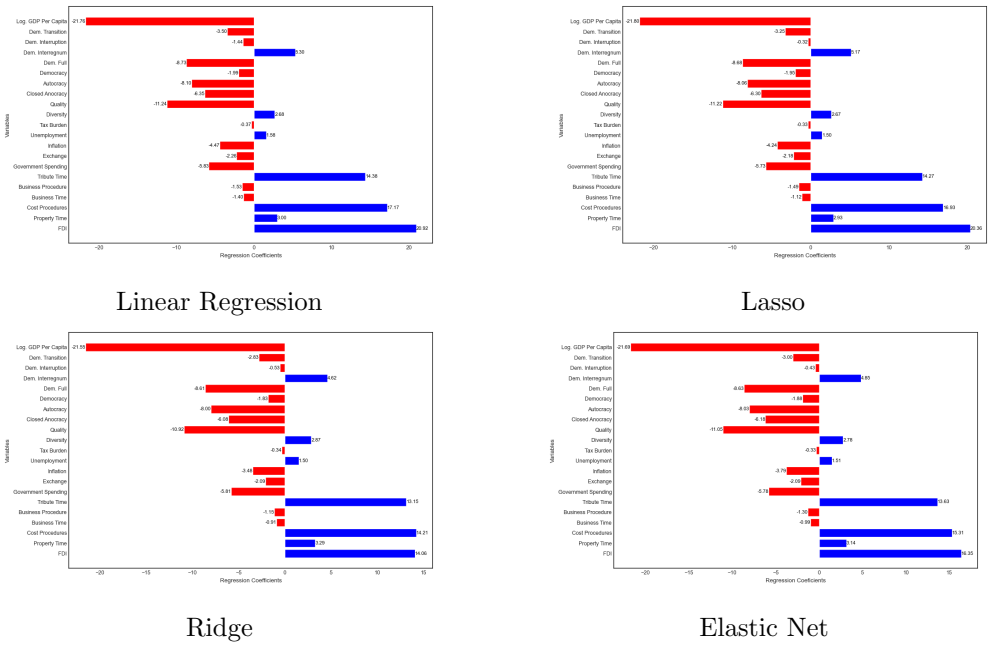
Model	**RMSE**	$\boldsymbol{R^2}$
Linear Regression	8.415	0.579
Lasso	8.409	0.579
Ridge	8.404	0.580
Elastic Net	8.403	0.579
Random Forest	3.62	0.933
SVR	7.314	0.696
XGBoost	3.592	0.921
LightGBM	3.582	0.938
CatBoost	3.446	0.957
Bagging	4.173	0.921
Stacking	4.45	0.916

These results show a considerable advance in terms of the coefficient of determination (R^2) compared to those of other studies using only linear models. For example, when analyzing the relationship between the different uses of the internet and the size of the informal economy, Elgin (2013) reported a coefficient of determination for all countries of at most 56%. Using almost the same variables but from a different time period, Goel and Nelson (2016), employing ordinary least squares (OLS) estimation, reported a R^2 of 49%. Canh and Dinh Thanh (2020), using the same dataset and variables, except the GDP per capita, employed a variety of specifications and found R^2 values ranging from 22.8% to 41.6%. In addition, the results presented in Table 3 are in line with those obtained by other studies. Shami and Lazebnik (2023) found a better performance of the Random Forest algorithm when compared to that of the Linear Regression model for the task of predicting the size of the informal economy. Considering the same task, in Ivașcu and Ștefoni (2023), the SVR also performed worse than other ML-based models.

3.2 Relative importance of the considered predictive variables

Although linear models usually do not yield better predictions when compared to those of models based on ML algorithms, they have the advantage of being more easily interpretable. More specifically, by using linear models, we know which variables positively or negatively influence the size of a dependent variable. To verify this information, the regression coefficients of the linear models were analyzed. Figure 2 presents the regression coefficients of the ten variables that most influenced the size of the informal sector for the four linear models.

Fig. 2: Regression coefficients of the considered linear models

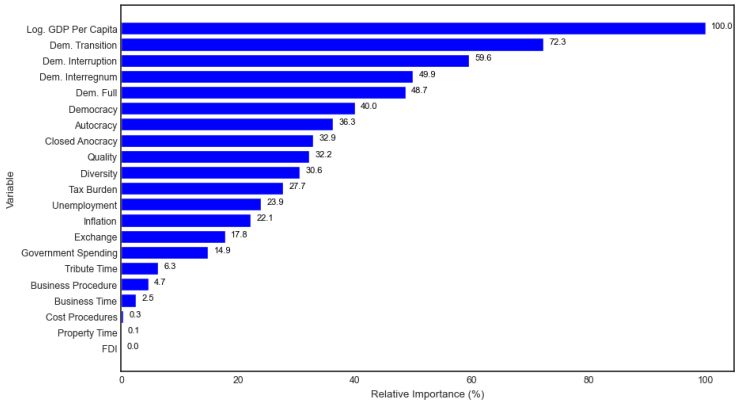


For all four models, the GDP per capita variable was the most important variable in reducing the size of the informal economy. In addition to this variable, the quality, Dem. Full, Autocracy, Closed Anocracy and government spending variables also contributed to decreasing the size of the informal economy, while the FDI, procedure cost, tribute time, and Dem. Interregnum contributed to increasing the size of the informal economy.

Figure 3 shows the relative importance of each of the predictive variables for the prediction of the target variable according to the Random Forest algorithm. The GDP per capita was the most important variable in the forecast of the size of the informal economy (as the most important variable, the relative importance of the GDP per capita is 100%). The dummy variables created to represent democracy were the next most relevant. Specifically, Dem. Transition, which was 62.9% as important as GDP

per capita for predicting the size of the informal economy, was the variable with the second greatest explanatory power. Dem. Interruption was 50.3% as important as the most important attribute for the prediction of the target variable. On the other hand, procedure cost and the FDI, which had large coefficient values in the linear models, had almost zero importance for the prediction in question.

Fig. 3: Relative importance of each predictive variable – Random Forest model



Unlike linear regression models, models based on ML algorithms are commonly described as black-box models because the interpretability or explanation of how the algorithms arrived at a certain result/prediction is difficult to understand. However, several techniques have been recently developed to obtain a better understanding of how such models obtain their predictions, as demonstrated by the advances achieved by Ribeiro et al. (2016), Lundberg and Lee (2017) and Ribeiro, Singh, and Guestrin (2018).

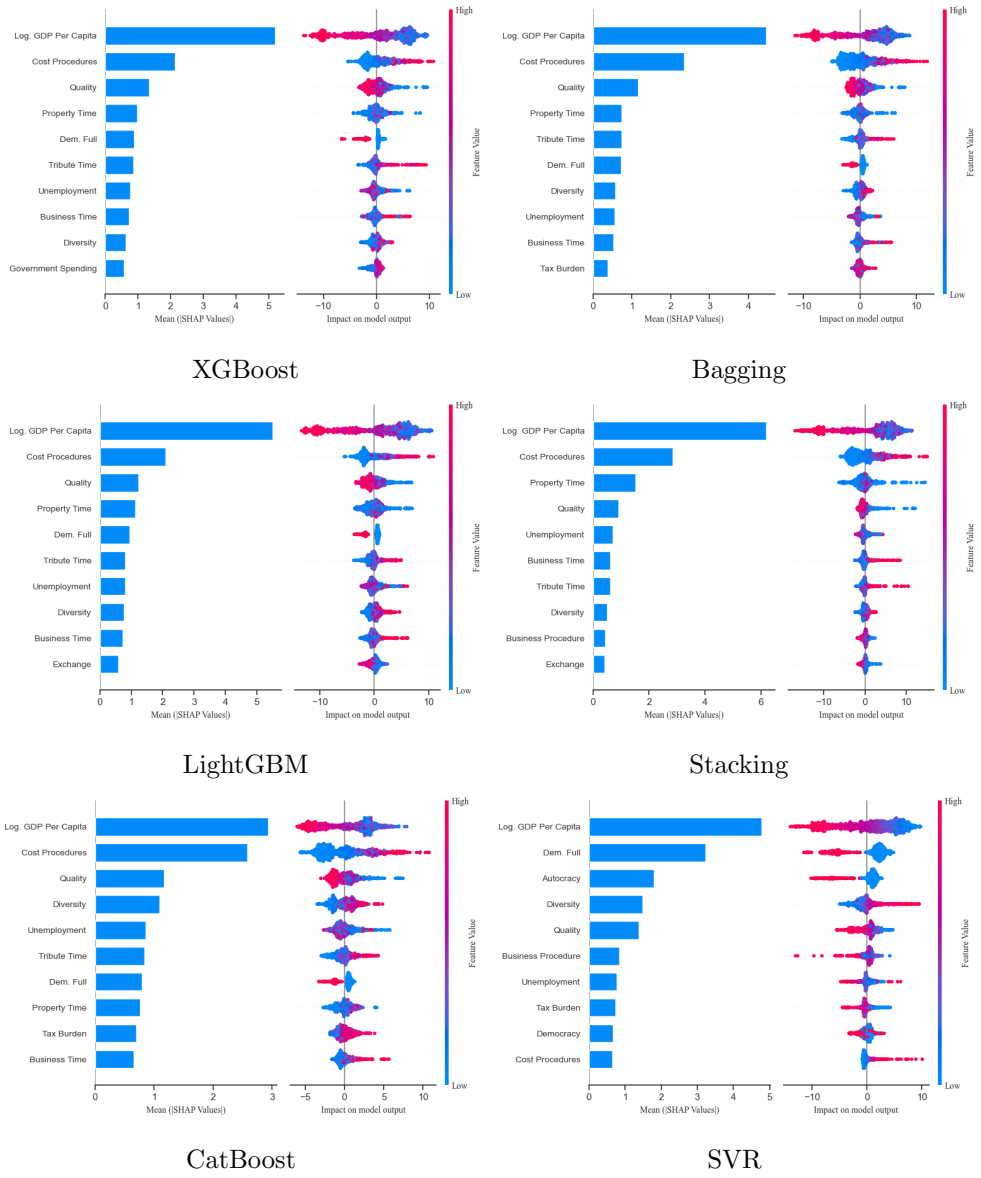
One of these techniques, the Shapley values technique (Shapley, 1953), was applied to the data in this paper to verify how each attribute contributes to the prediction of the value of the target variable. In general, the Shapley values compute the importance of each attribute for each of the observations in the training set and then indicate the final contribution of each independent variable in the final prediction. In this study, we computed the Shapley values using the **SH**apley **A**dditive ex**P**lanation (SHAP) methodology developed by Lundberg and Lee (2017). A SHAP value is calculated for each predictive variable/observation pair. A positive (negative) SHAP value means that the inclusion of that predictive variable in the model positively (negatively) influences the value of the target variable predicted by that model for that specific observation.[2]

Figure 4 shows the SHAP values for the six remaining ML algorithms. On the left side of each panel are the averages of the absolute SHAP values per observation for the ten most important attributes in descending order. On the right side of the

[2]For more details on SHAP values, see Lundberg and Lee (2017).

subgraphs are the SHAP values for each observation of these ten variables. The x-axis shows the SHAP values, that is, how the inclusion of the attribute influences (positively or negatively) the predicted value of the size of the informal economy for each observation. The closer to red (blue) the observation color is, the higher (lower) the attribute value for that observation.

Fig. 4: SHAP values of the considered ML-based models



XGBoost Bagging

LightGBM Stacking

CatBoost SVR

The left side of the panels in Figure 4 shows the importance of each attribute in determining the result. Therefore, as in the linear and Random Forest models, the GDP per capita was the most important variable for predicting the size of the informal sector. Procedure cost was the second most important variable for five of the six algorithms (in the case of the SVR, the second most important variable was Dem. Full), which is an important variable to understand the dynamics of the dependent attribute. This was also the case in the linear models.

The right side of the panels shows whether the attribute is positively or negatively related to the target variable. In the case of the GDP per capita variable, most of the observations in red color (i.e., with higher values of GDP per capita) have a negative SHAP value. This means that the inclusion of the predictive variable GDP per capita in the six models contributed, in most cases, to reducing the predicted value of the size of the informal economy. Therefore, the GDP per capita is negatively related to the size of the informal economy, corroborating the existing results. Following the same reasoning, it can be concluded that the attribute procedure cost (full debt) is positively (negatively) related to the size of the informal economy.

4 Conclusion

The use of models based on ML algorithms, which have performed better than linear models for prediction purposes, has been increasing as of late. In addition, although linear models have the advantage of considerable interpretive power, research efforts have intensified in recent years to make ML models more interpretable.

In this paper, we tested whether models based on ML algorithms have better performance relative to that of linear models for predicting the size of the informal economy. We also identified the most important determinants of the size of the informal economy in both types of models. The predictive variables used in the ML models were the same as those used in the literature based on traditional linear models.

The results suggest that models based on ML algorithms perform better in predicting the size of the informal economy. On average, the linear models obtained a coefficient of determination of 57.90% versus 89.74% for the ML models (excluding the SVR, this average rises to 93.10%). A likely plausible explanation for this superior performance is that ML models deal more adequately with nonlinear data. In fact, the expressive performance of ML models obtained in this paper is much higher compared to the results reported in the empirical studies by Elgin (2013), Goel and Nelson (2016) and Canh and Dinh Thanh (2020). These results are in line with those obtained by other studies (Ivașcu & Ștefoni, 2023; Shami & Lazebnik, 2023), who found a better performance of ML-based models when compared to that of linear models for the task of predicting the size of the informal economy. The results also suggest that the independent attributes indicated by the SHAP technique as the most important determinants of the size of the informal economy are essentially the same as those found in the literature based on linear models, namely, per capita income, as in Elgin (2013), Lyulyov et al. (2021) and Zhanabekov (2022), stage of democracy, as in Teobaldelli and Schneider (2013) and Elbahnasawy (2021), and bureaucratic elements, as in Goel and Nelson (2016).

Therefore, the various results obtained in this paper demonstrate that models that best deal with nonlinear data are a promising alternative for forecasting important economic variables, such as the size of the informal economy. In fact, the ML models obtained coefficient of determination values that were on average approximately 32 percentage points higher than those of the linear models, which suggests that such models should be recommended for the task of prediction. In the case of the size of the informal economy, the variable of interest in this paper, predicting and detecting its main determinants as accurately as possible is beneficial in the design, implementation, and management of economic policy in general and specific public policies that are complementary in a localized manner.

Appendix A Tables

Table A1: Hyperparameter optimization

Model	**Hyperparameter**	**Values**
Linear Regression	fit_intercept	[True, False]
	copy_X	[True, False]
Lasso	fit_intercept	[True, False]
	copy_X	[True, False]
	alpha	[0.001, 0.01, 0.1, 1.0, 10.0]
	max_iter	[1000, 2000, 3000]
Ridge	fit_intercept	[True, False]
	copy_X	[True, False]
	alpha	[0.001, 0.01, 0.1, 1.0, 10.0]
	max_iter	[1000, 2000, 3000]
Elastic Net	alpha	[0.001, 0.01, 0.1, 1.0, 10.0]
	l1_ratio	[0.25, 0.5, 0.75]
	fit_intercept	[True, False]
	max_iter	[1000, 2000, 3000]
	tol	[1e-4, 1e-3, 1e-2]
Random Forest	n_estimators	[100, 200, 300]
	criterion	[‘friedman_mse’, ‘squared_error’, ‘poisson’, ‘absolute_error’]
	max_depth	[None, 5, 10]
	min_samples_split	[2, 5, 10]
	min_samples_leaf	[1, 2, 4]
	max_features	[‘auto’, ‘sqrt’, ‘log2’]
	bootstrap	[True, False]
SVR	kernel	[‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’]
	C	[0.1, 1.0, 10.0]
	epsilon	[0.1, 0.2, 0.5]

Table A1: Hyperparameter optimization

Model	Hyperparameter	Values
XGBoosting	learning_rate	[0.1, 0.01, 0.001]
	max_depth	[3, 5, 10]
	n_estimators	[100, 200, 300]
	gamma	[0, 0.1, 0.2]
	subsample	[0.8, 1.0]
	colsample_bytree	[0.8, 1.0]
LightGBM	learning_rate	[0.1, 0.01, 0.001]
	max_depth	[3, 5, 10]
	n_estimators	[100, 200, 300]
	subsample	[0.8, 1.0]
	colsample_bytree	[0.8, 1.0]
CatBoost	learning_rate	[0.1, 0.01, 0.001]
	depth	[3, 5, 10]
	iterations	[100, 200, 300]
Bagging	n_estimators	[10, 50, 100]
	max_samples	[0.5, 0.7, 0.9]
	max_features	[0.5, 0.7, 0.9]

Table A2: List of countries

Angola	Chile	Fiji	Kazakhstan	Malaysia	Sierra Leone	Zambia
Albania	China	France	Kenya	Namibia	El Salvador	Zimbabwe
Argentina	Cameroon	Gabon	Kyrgyzstan	Niger	Suriname	
Armenia	Democratic Rep. Of Congo	United Kingdom	Cambodia	Nigeria	Slovakia	
Australia	Congo	Georgia	South Korea	Nicaragua	Slovenia	
Austria	Colombia	Ghana	Kuwait	Netherlands	Sweden	
Azerbaijan	Comoros	Greece	Laos	Norway	Syria	
Burundi	Costa Rica	Guatemala	Lebanon	New Zealand	Togo	
Belgium	Cyprus	Honduras	Sri Lanka	Pakistan	Thailand	
Benin	Czechia	Haiti	Lesotho	Peru	Tajikistan	
Burkina Faso	Germany	Hungary	Lithuania	Philippines	Tunisia	
Bangladesh	Denmark	Indonesia	Luxembourg	Poland	Türkiye	
Bulgaria	Dominican Republic	India	Latvia	Portugal	Tanzania	
Bahrain	Algeria	Ireland	Morocco	Paraguay	Uganda	
Belarus	Ecuador	Iran	Moldova	Qatar	Ukraine	
Bolivia	Egypt	Israel	Madagascar	Romania	Uruguay	
Brazil	Spain	Italy	Mexico	Russia	United States	
Botswana	Estonia	Jamaica	Mali	Saudi Arabia	Venezuela	
Canada	Ethiopia	Jordan	Mongolia	Senegal	Vietnam	
Switzerland	Finland	Japan	Mauritius	Singapore	South Africa	

Table A3: Data sources

Variable	Source	Code/Variable name
Inflation	WDI	NY.GDP.DEFL.KD.ZG
Unemployment	WDI	SL.UEM.TOTL.NE.ZS
Exchange	WDI	NE.TRD.GNFS.ZS
FDI	WDI	BX.KLT.DINV.WD.GD.ZS
Governement Spending	WDI	NE.CON.GOVT.ZS
Business Procedure	WDI	IC.REG.PROC
Cost Procedures	WDI	IC.REG.COST.PC.ZS
Business Time	WDI	IC.REG.DURS
Property Time	WDI	IC.PRP.DURS
Tribute Time	WDI	IC.TAX.DURS
GDP per capita	WDI	NY.GDP.PCAP.CD
Tax Burden	The Heritage Foundation	Tax Burden
Democracy	Polity5 Project	democ
Quality	IMF	Export Quality
Diversity	IMF	Export Diversity

Declarations

Funding

João Felix is grateful to Coordination for the Improvement of Higher Education Personnel (Capes) for research funding (grant 88887.668564/2022-00). Gilberto Tadeu Lima is grateful to the National Research Council of Scientific and Technological Development (CNPq) for research funding (grant 311811/2018-3).

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

Alm, J., & Embaye, A. (2013). Using dynamic panel methods to estimate shadow economies around the world, 1984–2006. *Public Finance Review*, *41*(5), 510–543,

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140,

Canh, P.N., & Dinh Thanh, S. (2020). Exports and the shadow economy: Non-linear effects. *The Journal of International Trade & Economic Development*, *29*(7), 865–890,

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Dabiri, H., Kheyroddin, A., Faramarzi, A. (2022). Predicting tensile strength of spliced and non-spliced steel bars using machine learning-and regression-based methods. *Construction and Building Materials*, *325*, 126835,

Dorogush, A.V., Ershov, V., Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, ,

Elbahnasawy, N.G. (2021). Can e-government limit the scope of the informal economy? *World Development*, *139*, 105341,

Elgin, C. (2013). Internet usage and the shadow economy: Evidence from panel data. *Economic Systems*, *37*(1), 111–121,

Gambhir, E., Jain, R., Gupta, A., Tomer, U. (2020). Regression analysis of COVID-19 using machine learning algorithms. *2020 international conference on smart electronics and communication (icosec)* (pp. 65–71).

Géron, A. (2022). *Hands-on machine learning with scikit-learn, keras, and tensorflow*. O'Reilly Media, Inc.

Goel, R.K., & Nelson, M.A. (2016). Shining a light on the shadows: Identifying robust determinants of the shadow economy. *Economic Modelling*, *58*, 351–364,

Goldstein, B.A., Navar, A.M., Carter, R.E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European Heart Journal*, *38*(23), 1805–1814,

Guo, F., Huang, Y., Wang, J., Wang, X. (2022). The informal economy at times of COVID-19 pandemic. *China Economic Review*, *71*, 101722,

Ivașcu, C.-F., & Ștefoni, S.E. (2023). Modelling the non-linear dependencies between government expenditures and shadow economy using data-driven approaches. *Scientific Annals of Economics and Business*, *70*(1), 97–114,

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, ,

Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, ,

Lyulyov, O., Paliienko, M., Prasol, L., Vasylieva, T., Kubatko, O., Kubatko, V. (2021). Determinants of shadow economy in transition countries: Economic and environmental aspects. *International Journal of Global Energy Issues*, *43*(2-3), 166–182,

Medina, L., & Schneider, M.F. (2018). *Shadow economies around the world: what did we learn over the last 20 years?* International Monetary Fund.

Ranis, G., & Stewart, F. (1999). V-goods and the role of the urban informal sector in development. *Economic Development and Cultural Change*, *47*(2), 259–288,

Ribeiro, M.T., Singh, S., Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M.T., Singh, S., Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the aaai conference on artificial intelligence* (Vol. 32).

Schneider, F., & Klinglmair, R. (2004). Shadow economies around the world: what do we know? *Available at SSRN 518526*, ,

Schneider, F., Raczkowski, K., Mróz, B. (2015). Shadow economy and tax evasion in the EU. *Journal of Money Laundering Control*, *18*(1), 34–51,

Shami, L., & Lazebnik, T. (2023). Implementing machine learning methods in estimating the size of the non-observed economy. *Computational Economics*, 1–18,

Shapley, L.S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317,

Smola, A.J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*, 199–222,

Teobaldelli, D., & Schneider, F. (2013). The influence of direct democracy on the shadow economy. *Public Choice*, *157*, 543–567,

Ulyssea, G. (2018). Firms, informality, and development: Theory and evidence from Brazil. *American Economic Review*, *108*(8), 2015-47,

Ulyssea, G. (2020). Informality: Causes and consequences for development. *Annual Review of Economics*, *12*(1), 525-546,

Vousinas, G.L. (2017). Shadow economy and tax evasion. The Achilles heel of Greek economy. Determinants, effects and policy proposals. *Journal of Money Laundering Control*, *20*(4), 386–404,

Wolpert, D.H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259,

Yoon, J. (2021). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, *57*(1), 247–265,

Zhanabekov, S. (2022). Robust determinants of the shadow economy. *Bulletin of Economic Research*, *74*(4), 1017–1052,