mathematics

Article

A Robust Approach for Identifying the Major Components of the Bribery Tolerance Index

Daniel Homocianu [1], Aurelian-Petruș Plopeanu [2] and Rodica Ianole-Calin [3,*]

[1] Department of Accounting, Business Information Systems and Statistics, Faculty of Economics and Business Administration, Alexandru Ioan Cuza University of Iasi, 700505 Iaşi, Romania; daniel.homocianu@uaic.ro
[2] Humanities and Social Sciences Research Department, Institute of Interdisciplinary Research, Alexandru Ioan Cuza University of Iasi, 700107 Iaşi, Romania; aplopeanu@gmail.com
[3] Faculty of Administration and Business, University of Bucharest, 030018 Bucharest, Romania
* Correspondence: rodica.ianole@faa.unibuc.ro

Abstract: The paper aims to emphasize the advantages of several advanced statistical and data mining techniques when applied to the dense literature on corruption measurements and determinants. For this purpose, we used all seven waves of the World Values Survey and we employed the Naive Bayes technique in SQL Server Analysis Services 2016, the LASSO package together with logit and melogit regressions with raw coefficients in Stata 16. We further conducted different types of tests and cross-validations on the wave, country, gender, and age categories. For eliminating multicollinearity, we used predictor correlation matrices. Moreover, we assessed the maximum computed variance inflation factor (VIF) against a maximum acceptable threshold, depending on the model's R squared in Ordinary Least Square (OLS) regressions. Our main contribution consists of a methodology for exploring and validating the most important predictors of the risk associated with bribery tolerance. We found the significant role of three influences corresponding to questions about attitudes towards the property, authority, and public services, and other people in terms of anti-cheating, anti-evasion, and anti-violence. We used scobit, probit, and logit regressions with average marginal effects to build and test the index based on these attitudes. We successfully tested the index using also risk prediction nomograms and accuracy measurements (AUCROC > 0.9).

Keywords: bribery tolerance index; Naive Bayes; LASSO; maximum acceptable VIF; correlation matrices; cross-validations; minimum accuracy loss; mixed-effects; average marginal effects; risk prediction nomograms

check for updates

Citation: Homocianu, D.; Plopeanu, A.-P.; Ianole-Calin, R. A Robust Approach for Identifying the Major Components of the Bribery Tolerance Index. *Mathematics* **2021**, *9*, 1570. https://doi.org/10.3390/math9131570

Academic Editor: David Carfì

Received: 14 June 2021
Accepted: 1 July 2021
Published: 3 July 2021

1. Introduction

The current massive increase in data about people's attitudes and behaviors raises both opportunities and challenges for economics and social sciences, on different levels [1]. One major area of innovation is reflected in the advanced statistical methodologies used to capture as accurately as possible the most relevant and actionable insights for private and public use [2]. In this spirit, there is a growing tendency to define comprehensive measures which are able to integrate various aspects of individual behaviors or socio-economic phenomena (e.g., development [3], poverty [4], and sustainability [5]). Under this umbrella, the use of composite indices appears as a common practice, with a high degree of heterogeneity concerning the many different computational techniques employed to obtain them. Namely, they vary from additive approaches (e.g., the tax morale index [6]) and ad-hoc selection of variables to more complex procedures, like principal component analysis and selection techniques using different correlation coefficients (e.g., the sustainable development index for European economies [7]).

As reported by [8], even if there are many available methods for variable selection (ridge or partial least-squares regressions [9]), the least absolute shrinkage and selection operator (LASSO) regression is desirable because it ensures sparsity of coefficients and interpretability of the covariate effects by directly reducing the regression variance [10]. Moreover, when constructing an area-based index, the LASSO avoids the disadvantages of the dimension reduction methods like not always being able to extract the common unique dimension for a composite index and inaccurate weights corresponding to variable importance [11]. Not last, LASSO is considered to be more resistant to "*p*-value mining" and thus, it provides a higher level of scientific reliability [12].

There has also been an increase in the application of machine learning and advances in statistics to solve similar problems and more [13–15].

Our paper aims to illustrate the strengths of this methodological framework for the case of corruption. Corruption is one of the topics highly investigated through composite index proxies [16,17], with the corruption perception index and control of corruption being among the most popular measures [18]. The edge of such indexes comes when considering various perspectives including organizational pathologies [19], and also how they combine "multiple data sources in a single index, lowering the probability of misrepresenting a country's level of corruption" [20].

In the spirit of this logic, our approach provides an in-depth understanding of variables within the same dataset (World Value Survey) and tests specific combinations of attitudes on different controversial actions (interpersonal and political violence, fiscal cheating, prostitution, etc.), as predictors for bribery tolerance. Through a set of rigorous statistical procedures, we retain the combination that offers the most robust prediction for considering bribery an acceptable practice, under the proposed name of the bribery tolerance index. Thus, we contribute to (1) the corruption literature by emphasizing potential areas of intervention in terms of public awareness and education on the importance of individual attitudes in shaping tolerant/intolerant behaviors; and (2) we advanced the methodological recommendations on how to build composite indexes, consolidating the positive evidence for the use of penalized regression techniques in social sciences and pointing to the need for comprehensive approaches that pay more attention to non-normal distributions, cross-validation tests, multicollinearity, objectively argued weighting, and prediction accuracy.

2. Materials and Methods

We used data from World Values Survey (the WVS TimeSeries 1981 2020 Stata v1 6.zip file available at https://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp and accessed on 1 February 2021) in the Stata format. We started from all seven waves, namely 1981–1984, 1989–1993, 1994–1998, 1999–2004, 2005–2009, 2010–2014, and 2017–2020, and 105 countries from all continents, more precisely from a dataset with 426,452 observations and 1045 variables. We also exported this as.csv (comma-separated values format, 1.75 GB) to serve for first-round mining.

For identifying the main influences of bribery tolerance, many methods, techniques, and instruments have been used. A schematic representation is also included (Appendix A—Figure A1). Among them:

- The Naive Bayes algorithm in the Microsoft Data Mining (DM) add-in available in Microsoft Excel and working with SQL Server Analysis Services 2016 Enterprise Edition x64 (SSAS) which acted as first-round mining (maximum ~10.7 GB of RAM (Random Access Memory) consumed only by SSAS and maximum ~5.3 additionally used by Excel on a Windows Server 2016 Datacenter virtual machine with six Intel Xeon Gold 6240 Cascadelake CPU cores and 24.4 GB of RAM (maximum ~18.5 GB consumed including the operating system) running on one of the private clouds (https://cloud.raas.uaic.ro/; https://tinyurl.com/4wn9huwj—last accessed on 9 June 2021) of Alexandru Ioan Cuza University, currently managed using OpenStack on Ubuntu and dedicated to research;
- Two forms of LASSO, originally documented in geophysics as an L1-regularization approach for a problem called sparse spike deconvolution [21–23], namely RLASSO (rigorous and penalizing LASSO to control overfitting), and CVLASSO (the LASSO which performs time-consuming cross-validation—here with the LSE option = largest

lambda for which the Means Squared Prediction Error-MSPE is within one standard error of the minimal MSPE) in Stata 16 MP-64 bit (second round mining);

- Elimination from correlated pairs (values near or less than 0.5 indicating low correlation according to [24]) using additional considerations—e.g., amount of missing values, AUCROC [25] for logit [26] regressions (Equation (1)), and VIF for OLS ones;
- Cross-validations using the mixed-effects technique (melogit—variables to select from as fixed effects) and checks of significance loss using the value of p—it was additionally considered the significance level of 1‰ starting from the fact that in large samples, p-values tend to decrease quickly to zero [27] when considering six binary forms (derived also in Stata) of the dependent variable (F117—Tables 1 and 2) and four criteria (random effects): wave, country, sex, and age;
- Smaller values of AIC-Akaike Information Criterion, and BIC-Bayesian Information Criterion [28] for a better fit of the chosen model [29], where the goodness of fit describes how well a statistical model fits a set of observations; larger ones for both p (contradiction of the Ho hypothesis) in the case of GOF (Goodness of Fit) test and chi^2 for the same GOF as additional indications of a better fit of the model; and
- Larger R-squared [30] for a better explaining power of the models.

$$\text{Logit}(p) = \ln(p/(1-p)) = \beta_0 + \sum(\beta_i \times X_i) + \varepsilon \tag{1}$$

where:

- p is the probability (risk) of bribery tolerance;
- $(1 - p)$ is the probability of not considering bribes acceptable;
- $p/(1 - p)$ represents the odds of bribery tolerance;
- i = 2, . . . , n and n is the total number of independent variables;
- β_0 is the bias (intercept) term;
- β_i measures the effect of a change in variable X_i on the risk of bribery tolerance;
- X_i is one explanatory variable from the array ($\sum$) of features selected after using LASSO; and
- ε represents the error term.

Table 1. WVS variables (and the corresponding original codes) used in our study.

Variable	Question	Coding
E290	Justifiable: Political violence (DK/NA as blanks)	1–10 scale
F114A	Justifiable: Claiming government benefits to which you are not entitled(DK/NA as blanks)	1–10 scale
F114B	Justifiable: Stealing property (DK/NA as blanks)	1–10 scale
F114D	Justifiable: Violence against other people (DK/NA as blanks)	1–10 scale
F114E	Justifiable: Terrorism as a political, ideological, or religious mean(DK/NA as blanks)	1–10 scale
F115	Justifiable: Avoiding a fare on public transport (DK/NA as blanks)	1–10 scale
F116	Justifiable: Cheating on taxes (DK/NA as blanks)	1–10 scale
F119	Justifiable: Prostitution (DK/NA as blanks)	1–10 scale
F199	Justifiable: For a man to beat his wife (DK/NA as blanks)	1–10 scale
Y010	SACSECVAL—Welzel Overall Secular Values (DK/NA as blanks)	0–1 scale
F117 (original outcome)	Justifiable: Someone accepting a bribe (DK/NA as blanks)	1–10 scale
justif_bribe_5_10	1 if F117 != Blank AND F117 >= 5; 0 if F117 != Blank AND F117 < 5	0 or 1
justif_bribe_6_10	1 if F117 != Blank AND F117 >= 6; 0 if F117 != Blank AND F117 < 6 AND F117 > 0	0 or 1
justif_bribe_7_10	1 if F117 != Blank AND F117 >= 7; 0 if F117 != Blank AND F117 < 7 AND F117 > 0	0 or 1
justif_bribe_8_10	1 if F117 != Blank AND F117 >= 8; 0 if F117 != Blank AND F117 < 8 AND F117 > 0	0 or 1
justif_bribe_9_10	1 if F117 != Blank AND F117 >= 9; 0 if F117 != Blank AND F117 < 9 AND F117 > 0	0 or 1
justif_bribe_10	1 if F117 == 10; 0 if F117 != Blank AND F117 < 10 AND F117 > 0	0 or 1

Source: WVS's data available at https://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp (accessed on 1 February 2021).

Table 2. Descriptive statistics for the WVS variables that were used in this study.

Variable	*n*	Mean	Std.Dev.	Min	0.25	Median	0.75	Max
E290	67,841	1.96	1.91	1	1	1	2	10
F114A	396,038	2.65	2.52	1	1	1	4	10
F114B	158,481	1.8	1.8	1	1	1	2	10
F114D	158,284	1.95	1.88	1	1	1	2	10
F114E	66,858	1.79	1.81	1	1	1	2	10
F115	388,305	2.6	2.45	1	1	1	4	10
F116	394,728	2.26	2.22	1	1	1	3	10
F119	360,755	2.63	2.48	1	1	1	4	10
F199	230,248	1.94	1.99	1	1	1	2	10
Y010	411,064	0.35	0.18	0	0.22	0.35	0.47	1
F117 (original outcome)	410,100	1.84	1.84	1	1	1	2	10
justif_bribe_5_10	410,100	0.09	0.29	0	0	0	0	1
justif_bribe_6_10	410,100	0.06	0.24	0	0	0	0	1
justif_bribe_7_10	410,100	0.04	0.2	0	0	0	0	1
justif_bribe_8_10	410,100	0.03	0.17	0	0	0	0	1
justif_bribe_9_10	410,100	0.02	0.14	0	0	0	0	1
justif_bribe_10	410,100	0.02	0.12	0	0	0	0	1

Source: Own calculation in Stata 16MP 64-bit using WVS's data.

Source: Own calculation in Stata 16MP 64-bit using WVS's data.

To correct for any form of heteroskedasticity, robust standard errors have been used for all types of regressions.

In terms of building the index of bribery tolerance, we additionally used average marginal effects calculated after performing scobit regressions to determine the weights of the resulting components. Scobit usually serves as an alternative [31] for the two most common techniques for estimation of models with a binary dependent variable, namely logit and probit, in case of disturbances to the normal or logistic distributions.

In addition, logit-based risk-prediction nomograms [32] served for confirming the approximation of these weights. Moreover, they ensured visual comparability because of supporting the overall view of the magnitude of effects and interpretations directly in risk terms when predicting the bribery tolerance.

The proposed framework's complexity (Appendix A—Figure A1) is also due to its dual nature. The latter means that it includes both automatic/unsupervised (first two data mining rounds—top of Figure A1) and supervised steps (derivations, further selections, building and testing the index—central part and bottom of Figure A1), which is an accepted practice in the field [33].

3. Results

After running the Naive Bayes algorithm, a dependency network has been obtained (Figure 1). Using a slider operating from the strongest to all links, the latter suggests the best predictors identified under the assumption of their independence, which will be subjected to further tests in Stata, responsible for eliminating redundant variables.

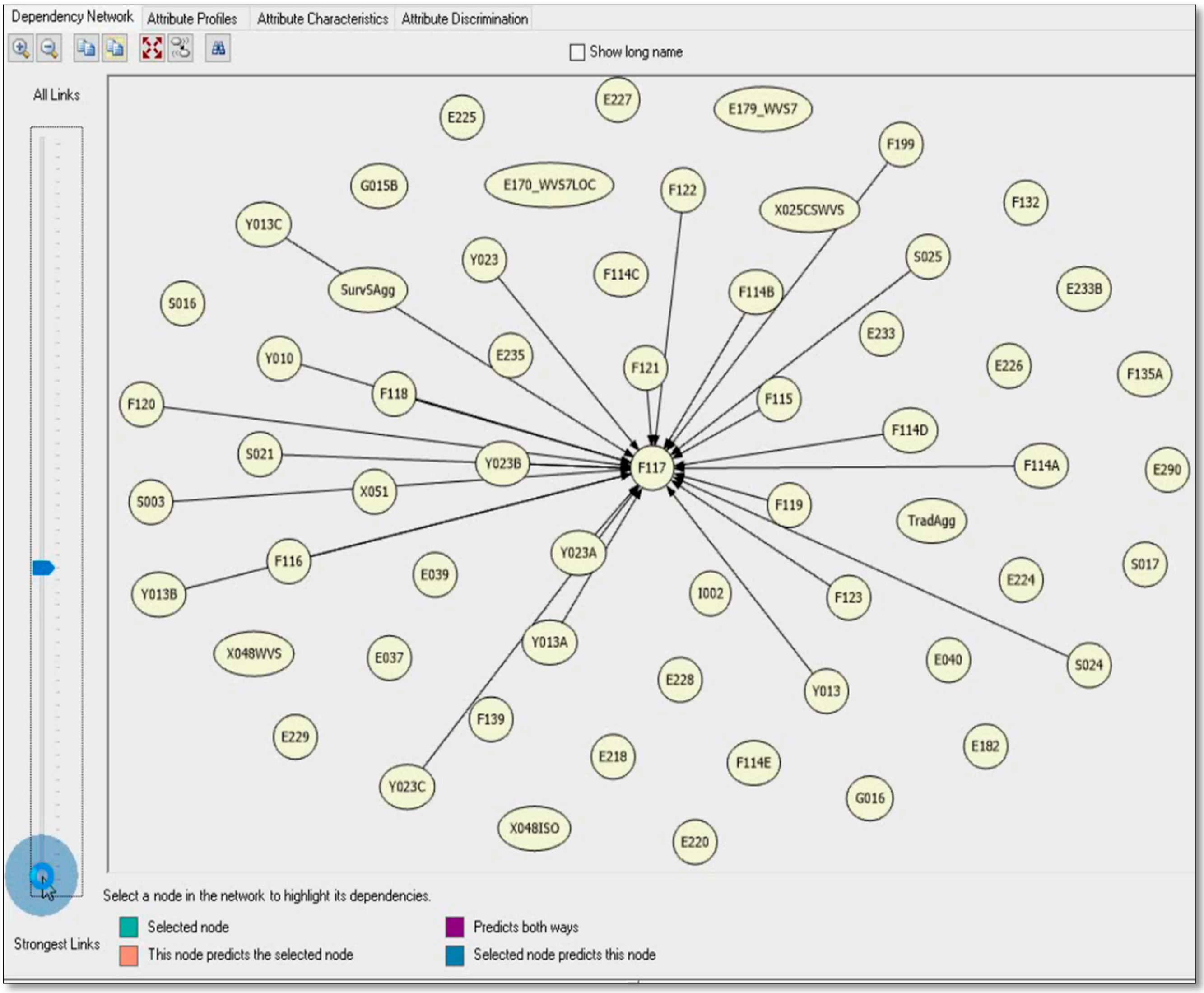

Figure 1. Top predictors as identified by using the Naive Bayes algorithm in Microsoft Data Mining add-in for spreadsheets. Source: Own computation in Microsoft Excel 2013 (Data Mining add-in) and SQL Server Analysis Services 2016.

After running both RLASSO and CVLASSO many times (depending on the variable category and overall, on a concatenation of the resulting selections) starting from all possible predictors available in Figure 1 (more than 50), only 13 have been retained, namely E290, F114A, F114B, F114D, F114E, F115, F116, F119, F199, Y010, Y013A, Y013B, and Y013C (Tables 1 and 2). From this list above, the last three have been identified as directly derived from F115, F116, F117 (Welzel relativism—https://www.worldvaluessurvey.org/WVSContents.jsp?CMSID=welzelidx&CMSID=welzelidx—accessed on 1 February 2021) and have been dropped (only 10 remaining—Tables 1 and 2).

The next step was to perform derivation operations (binary alternatives of the original outcome—F117) in Stata (Figure 2).

```
use "F:\d.p.desktop\dm\WVS_TimeSeries_stata_v1_6.dta"
generate justif_bribe_5_10 =.
replace justif_bribe_5_10 = 1 if F117!=. & F117>=5
replace justif_bribe_5_10 = 0 if F117!=. & F117<5 & F117>0
generate justif_bribe_6_10 =.
replace justif_bribe_6_10 = 1 if F117!=. & F117>=6
replace justif_bribe_6_10 = 0 if F117!=. & F117<6 & F117>0
generate justif_bribe_7_10 =.
replace justif_bribe_7_10 = 1 if F117!=. & F117>=7
replace justif_bribe_7_10 = 0 if F117!=. & F117<7 & F117>0
generate justif_bribe_8_10 =.
replace justif_bribe_8_10 = 1 if F117!=. & F117>=8
replace justif_bribe_8_10 = 0 if F117!=. & F117<8 & F117>0
generate justif_bribe_9_10 =.
replace justif_bribe_9_10 = 1 if F117!=. & F117>=9
replace justif_bribe_9_10 = 0 if F117!=. & F117<9 & F117>0
generate justif_bribe_10 =.
replace justif_bribe_10 = 1 if F117==10
replace justif_bribe_10 = 0 if F117!=. & F117<10 & F117>0
```

Figure 2. Stata processing script for performing derivations needed for cross-validations and starting from the original form of the dependent variable. Source: Own calculation in Stata 16MP 64-bit.

From the very beginning (Tables 1 and 2), the study site indicates the lowest number of responses for the questions corresponding to variables E290 and F114E.

After consecutively checking the correlation matrices when starting from this set of 10 remaining input variables (top of Figure 3—issues for correlation coefficients larger than 0.5 as absolute value), some intercorrelated ones have been eliminated when considering additional criteria. For instance, the accuracy one (Area under the Curve of the Receiver Operating Characteristics—AUCROC) for Logit models with just one predictor and an alternate binary outcome for F117 (justif_bribe_6_10). For instance, when considering F114D vs. E290 (AUCROC of 0.8099 for 157,239 observations vs. 0.7818 for 67,532). The same (first) for F114D vs. F114E, the first was kept (AUCROC of 0.8099 for 157,239 observations vs. 0.7863 for 66,517). At this point, only eight possible predictors remained, by eliminating E290 and F114E.

In the case of F114D vs. F199 (AUCROC of 0.8099 for 157,239 observations vs. 0.7853 for 228,185), the first one was also preserved. The reason was that the first one is also a more general variable (violence against other people) than the second (violence against wife) and brings a better explaining power (Pseudo R^2 of 0.2418 vs. 0.2126). When compared using the same basis (156,793 observations) obtained with a filtering condition (NOT NULL) on the opposite (IF var.!=.), the results confirmed again the choice of F114D (AUCROC of 0.8103) over F1119 (0.7919). At this point, only seven possible predictors remained.

	E290	F114A	F114B	F114D	F114E	F115	F116	F119	F199	Y010
E290	1.0000									
F114A	0.2965	1.0000								
F114B	0.5239	0.3677	1.0000							
F114D	**0.6312**	0.2872	0.5475	1.0000						
F114E	0.6632	0.3107	0.5576	**0.6591**	1.0000					
F115	0.3194	0.4547	0.4428	0.3220	0.3342	1.0000				
F116	0.4501	0.3639	**0.5973**	0.4640	0.4595	0.4685	1.0000			
F119	0.3477	0.1602	0.3327	0.3449	0.3077	0.2295	0.3238	1.0000		
F199	0.5668	0.2655	0.5388	**0.6528**	0.6060	0.3007	0.4478	0.2886	1.0000	
Y010	0.2968	0.2334	0.3380	0.2889	0.2539	0.3550	0.4074	0.3470	0.2335	1.0000

	F114A	F114D	F115	F116	F119	Y010
F114A	1.0000					
F114D	0.3203	1.0000				
F115	0.4785	0.3576	1.0000			
F116	0.4048	0.4943	0.4995	1.0000		
F119	0.2029	0.3745	0.2649	0.3520	1.0000	
Y010	0.2760	0.2977	0.3954	0.4415	0.3570	1.0000

Figure 3. Correlation matrices for 10 predictors obtained after performing R and CV LASSO selections and six remaining ones after removing the redundancy. Source: Own calculation in Stata 16MP 64-bit.

In the case of F116 vs. F114B (AUCROC of 0.8448 for 392,236 observations vs. 0.8495 for 157,569), additional arguments have been considered. They resulted when comparing the values for AUCROC when dropping each variable in Logit models (justif_bribe_6_10 as alternate binary outcome) with six predictors: AUCROC of 0.9271 for 122,545 observations when keeping F116 vs. 0.9186 for 124,711 when keeping F114B). More, when compared using the same basis (153,512 observations) obtained with a filtering condition (NOT NULL) on the opposite (IF var.!=.), the results confirmed again the choice of F116 (AUCROC of 0.8841) over F114B (0.8497). At this point, only six variables remained to be considered, namely F114A, F114D, F115, F116, F119, and Y010 (bottom of Figure 3—all correlation coefficients less than 0.5 as absolute value).

When evaluated against Max. Acceptable VIF (Equation (2), the maximum computed one [34–36] for the overall OLS model (same alternate binary outcome, namely justif_bribe_6_10) including only these six influences still indicates multicollinearity issues (1.743 vs. 1.5479 as Max. Computed VIF vs. Max. Acceptable VIF). The maximum absolute value of the correlation coefficient (a correlation command only for these six remaining input variables—bottom of Figure 3) is 0.4995 (~0.5 corresponding to F115 and F116), which also suggests low to medium correlation among the predictors. If additionally using a correlation command only for these two predictors (F115 vs. F116), the correlation coefficient (0.5068) better suggests the redundancy. If checking the corresponding average marginal effects of these two in a Logit model, the one of F116 is more than three times larger than that of F115. Besides, when comparing the accuracy of two Logit models with five predictors resulting when eliminating each but using a comparable basis (122,545 records if using the not null condition for the one eliminated) the AUCROC values also favored F116 against F115 (0.9248 vs. 0.9072). Moreover, avoiding a fare on public transportation (F115) is a particular form of cheating on taxes (F116) because it represents cheating on public transportation taxes.

$$\text{Max. Acceptable VIF} = 1/(1 - \mathrm{R}^2), \quad (2)$$

where R^2 is the model's coefficient of determination.

Moreover, if performing further cross-validations using a mixed-effects [37] method (melogit—those six above as fixed effects and the following as random ones, namely S002-

the wave, S003-the country code, X001-the respondents' sex, and X003R-the respondents' age for six intervals) only three resisted without losing significance (Tables 3–6) regardless of the binary format of the analyzed variable (justif_bribe_5_10, justif_bribe_6_10, justif_bribe_7_10, justif_bribe_8_10, justif_bribe_9_10, and justif_bribe_10). It is about F114A, F114D, and F116. If considering only these three and performing again the cross-validations above, the results obtained confirm no loss of significance. From the models in the first set of cross-validations (Tables 3 and 5—fourth model) predicting the second alternative binary form of the outcome (justif_bribe_6_10) has been discovered as the most accurate. In addition, the latter form corresponds to the symmetric upper band from the entire original scale (1–10). The choice above is also confirmed when using only the remaining three predictors. Therefore, we used this form with only three components to construct the index, taking into account only their marginal effects (not multiplied with the amplitudes) since the scales of the three variables are identical (Table 7, model 1, Figure 4—lines 1 and 2, and top of Figure 5).

In terms of data support for all the variables involved in the final selection (these three predictors above and the variable to analyze), we found coverage as follows: 151,636 distinct non-null observations (Table 7); last two waves (2010–2014 and 2017–2020) as the only two in which the question corresponding to F114D (Justifiable: Violence against other people) was included (the other three, namely F114A, F116, and F117, appear in all seven waves); 75 countries from all continents; a ratio of 1.1:1 between female and male respondents; and all six age intervals considered by WVS, namely 15–24, 25–34, 35–44, 45–54, 55–64, and 65 years or more.

Scobit, Logit, and Probit regressions [31] (justif_bribe_6_10 as chosen binary format based on maximum accuracy) have been used to confirm this form of the index (Table 7—models 1 vs. 2, 3 vs. 4, and 5 vs. 6) obtained from the most powerful three variables unveiled and underlined in the WVS's dataset. As observed above (Table 7), when compared with the base models having those three predictors, the alternate ones based only on this index as single input does not lose at all in terms of explaining power (Pseudo R^2) and accuracy of classification (AUCROC).

In terms of further validations using Zlotnik and Abraira's prediction nomograms [32] based on Logit regressions (Table 7, models 3 and 4 with excellent accuracy of classification—AUCROC = 0.9169), two such visual outputs have been generated. The first one (top of Figure 5) corresponds to the original model with those three predictors (positive influences) and clearly shows that as the respondents' attitudes towards cheating taxes (F116), violence against other people (F114D), and accepting government benefits not entitled (F114A) are more permissive, the risk of accepting bribes increases. The second (bottom of Figure 5) is based on the model having the index as the only predictor (also positive influence). Moreover, in terms of maximum theoretical probability (Zlotnik and Abraira's nomolog [32] for Logit regressions—Table 7, models 3 vs. 4) that corresponds to the most favorable value or combination of values of the predictors, it seems that there is no loss. Actually, in both cases, the score (value of 10) or the sum of scores (about 22.2) indicates a maximum probability with a value slightly above the 95% threshold.

Table 3. Cross-validations on country and wave using the first three alternate binary forms of the dependent variable.

Input Variable/Model	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
Dependent Variable	**justif_bribe_ 5_10**	**justif_bribe_ 5_10**	**justif_bribe_ 5_10**	**justif_bribe_ 6_10**	**justif_bribe_ 6_10**	**justif_bribe_ 6_10**	**justif_bribe_ 7_10**	**justif_bribe_ 7_10**	**justif_bribe_ 7_10**
Model Type	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**
F114A	0.1472 ***	0.1481 ***	0.1309 ***	0.1549 ***	0.1560 ***	0.1407 ***	0.1606 ***	0.1612 ***	0.1488 ***
	(0.0048)	(0.0101)	(0.0130)	(0.0058)	(0.0176)	(0.0147)	(0.0067)	(0.0194)	(0.0203)
F114D	0.2969 ***	0.2960 ***	0.2468 ***	0.2690 ***	0.2673 ***	0.2218 ***	0.2329 ***	0.2314 ***	0.1935 ***
	(0.0055)	(0.0134)	(0.0209)	(0.0060)	(0.0132)	(0.0211)	(0.0067)	(0.0032)	(0.0215)
F115	0.0995 ***	0.0999 ***	0.0851 ***	0.1095 ***	0.1101 ***	0.1008 ***	0.1052 ***	0.1057 ***	0.0988 ***
	(0.0053)	(0.0042)	(0.0140)	(0.0063)	(0.0068)	(0.0116)	(0.0074)	(0.0121)	(0.0129)
F116	0.3692 ***	0.3685 ***	0.3501 ***	0.3606 ***	0.3594 ***	0.3384 ***	0.3622 ***	0.3613 ***	0.3380 ***
	(0.0052)	(0.0446)	(0.0265)	(0.0058)	(0.0495)	(0.0243)	(0.0068)	(0.0554)	(0.0266)
F119	0.1367 ***	0.1375 ***	0.1672 ***	0.1448 ***	0.1460 ***	0.1661 ***	0.1369 ***	0.1378 ***	0.1548 ***
	(0.0049)	(0.0130)	(0.0157)	(0.0058)	(0.0070)	(0.0150)	(0.0067)	(0.0135)	(0.0179)
Y010	1.8602 ***	1.8565 ***	2.9867 ***	0.9335 ***	0.9239 ***	1.8382 ***	0.1470	0.1372 *	1.0877 ***
	(0.0794)	(0.2581)	(0.2827)	(0.0934)	(0.1112)	(0.2746)	(0.1061)	(0.0697)	(0.2936)
_cons	−6.3675 ***	−6.3686 ***	−6.8198 ***	−6.7649 ***	−6.7639 ***	−7.1348 ***	−6.8293 ***	−6.8259 ***	−7.2325 ***
	(0.0440)	(0.1273)	(0.1517)	(0.0537)	(0.1819)	(0.1774)	(0.0602)	(0.1926)	(0.1882)
var(_cons[S002])	N/A	0.0024 ***	N/A	N/A	0.0053 ***	N/A	N/A	0.0033 ***	N/A
	N/A	(0.0001)	N/A	N/A	(0.0006)	N/A	N/A	(0.0008)	N/A
var(_cons[S003])	N/A	N/A	0.3525 ***	N/A	N/A	0.3266 ***	N/A	N/A	0.3086 ***
	N/A	N/A	(0.0730)	N/A	N/A	(0.0641)	N/A	N/A	(0.0701)
N	122,545	122,545	122,545	122,545	122,545	122,545	122,545	122,545	122,545
Chi^2	18,816.6205	N/A	2364.8383	15,184.0298	N/A	1535.5780	12,985.2662	N/A	1418.4369
p	0.0000	N/A	0.0000	0.0000	N/A	0.0000	0.0000	N/A	0.0000
Pseudo R^2	0.4489	N/A	N/A	0.4585	N/A	N/A	0.4504	N/A	N/A

Table 3. *Cont.*

Input Variable/Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	**justif_bribe_ 5_10**	**justif_bribe_ 5_10**	**justif_bribe_ 5_10**	**justif_bribe_ 6_10**	**justif_bribe_ 6_10**	**justif_bribe_ 6_10**	**justif_bribe_ 7_10**	**justif_bribe_ 7_10**	**justif_bribe_ 7_10**
Model Type	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**
maxAbsCorCoefPredMtrx	0.4995	N/A	N/A	0.4995	N/A	N/A	0.4995	N/A	N/A
AUCROC	0.9220	N/A	N/A	0.9271	N/A	N/A	0.9251	N/A	N/A
pGOF	0.0363	N/A	N/A	0.2091	N/A	N/A	0.0000	N/A	N/A
Chi^2GOF	82,080.56	N/A	N/A	81,681.32	N/A	N/A	83,315.83	N/A	N/A
AIC	48,401.9455	48,379.8688	46,746.5493	35,258.8082	35,228.5612	34,194.5529	28,095.9260	28,077.7890	27,351.2618
BIC	48,469.9591	48,399.3012	46,824.2792	35,326.8219	35,247.9937	34,272.2828	28,163.9396	28,097.2215	27,428.9916
maxPnomologBiggerThan	0.9500	N/A	N/A	0.9000	N/A	N/A	0.9000	N/A	N/A

Source: Own calculation in Stata 16MP 64-bit. Notes: Robust standard errors are presented in parentheses. All raw coefficients above parentheses are significant at 5% (*), 1% (**), and 1‰ (***), respectively.

Table 4. Cross-validations on country and wave using the last three alternate binary forms of the dependent variable.

Input Variable/Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	**justif_bribe_ 8_10**	**justif_bribe_ 8_10**	**justif_bribe_ 8_10**	**justif_bribe_ 9_10**	**justif_bribe_ 9_10**	**justif_bribe_ 9_10**	**justif_bribe_ 10**	**justif_bribe_ 10**	**justif_bribe_ 10**
Model Type	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**
F114A	0.1617 ***	0.1620 ***	0.1528 ***	0.1683 ***	0.1685 ***	0.1603 ***	0.1743 ***	0.1743 ***	0.1677 ***
	(0.0078)	(0.0153)	(0.0282)	(0.0096)	(0.0256)	(0.0379)	(0.0116)	(0.0386)	(0.0446)
F114D	0.1953 ***	0.1941 ***	0.1628 ***	0.1728 ***	0.1712 ***	0.1456 ***	0.1491 ***	0.1489 **	0.1376 ***
	(0.0076)	(0.0097)	(0.0211)	(0.0090)	(0.0273)	(0.0223)	(0.0105)	(0.0469)	(0.0206)
F115	0.0987 ***	0.0990 ***	0.0985 ***	0.0899 ***	0.0902 ***	0.1003 ***	0.0756 ***	0.0756	0.1003 ***
	(0.0087)	(0.0173)	(0.0154)	(0.0106)	(0.0242)	(0.0175)	(0.0125)	(0.0402)	(0.0198)

Table 4. *Cont.*

Input Variable/Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	**justif_bribe_ 8_10**	**justif_bribe_ 8_10**	**justif_bribe_ 8_10**	**justif_bribe_ 9_10**	**justif_bribe_ 9_10**	**justif_bribe_ 9_10**	**justif_bribe_ 10**	**justif_bribe_ 10**	**justif_bribe_ 10**
Model Type	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**	**Logit—No Random Effects**	**MeLogit—Random Effects on Wave (S002)**	**MeLogit—Random Effects on Country (S003)**
F116	0.3742 ***	0.3736 ***	0.3450 ***	0.3917 ***	0.3911 ***	0.3460 ***	0.4122 ***	0.4122 ***	0.3521 ***
	(0.0084)	(0.0613)	(0.0331)	(0.0108)	(0.0769)	(0.0419)	(0.0136)	(0.0896)	(0.0497)
F119	0.1170 ***	0.1177 ***	0.1376 ***	0.0937 ***	0.0948	0.1229 ***	0.0656 ***	0.0658	0.1058 **
	(0.0079)	(0.0304)	(0.0255)	(0.0095)	(0.0529)	(0.0350)	(0.0111)	(0.0660)	(0.0405)
Y010	−0.5917 ***	−0.6018 ***	0.4050	−1.0694 ***	−1.0876 ***	−0.1236	−1.2979 ***	−1.3014 ***	−0.9532 *
	(0.1200)	(0.0938)	(0.3318)	(0.1412)	(0.1836)	(0.3513)	(0.1602)	(0.3919)	(0.4145)
_cons	−6.8296 ***	−6.8253 ***	−7.3027 ***	−7.0661 ***	−7.0586 ***	−7.6234 ***	−7.2663 ***	−7.2647 ***	−7.6797 ***
	(0.0674)	(0.1518)	(0.2112)	(0.0824)	(0.1287)	(0.2835)	(0.0999)	(0.1797)	(0.3525)
var(_cons[S002])	N/A	0.0022 *	N/A	N/A	0.0037 ***	N/A	N/A	0.0003	N/A
	N/A	(0.0009)	N/A	N/A	(0.0010)	N/A	N/A	(0.0005)	N/A
var(_cons[S003])	N/A	N/A	0.3415 ***	N/A	N/A	0.5071 ***	N/A	N/A	0.5515 ***
	N/A	N/A	(0.0730)	N/A	N/A	(0.1222)	N/A	N/A	(0.1209)
N	122,545	122,545	122,545	122,545	122,545	122,545	122,545	122,545	122,545
Chi^2	10,868.6143	N/A	1163.2037	8177.9460	N/A	829.7018	6141.2345	N/A	445.1360
p	0.0000	N/A	0.0000	0.0000	N/A	0.0000	0.0000	N/A	0.0000
Pseudo R^2	0.4373	N/A	N/A	0.4336	N/A	N/A	0.4188	N/A	N/A
maxAbsCorCoefPredMtrx	0.4995	N/A	N/A	0.4995	N/A	N/A	0.4995	N/A	N/A
AUCROC	0.9176	N/A	N/A	0.9145	N/A	N/A	0.9071	N/A	N/A
pGOF	0.0000	N/A	N/A	0.0000	N/A	N/A	0.0000	N/A	N/A
Chi^2GOF	89,944.11	N/A	N/A	99,321.47	N/A	N/A	1.1×10^5	N/A	N/A
AIC	22,303.0334	22,289.6962	21,626.3242	16,841.7315	16,827.2517	16,085.5189	13,327.1104	13,315.0487	12,459.7028
BIC	22,371.0470	22,309.1287	21,704.0541	16,909.7451	16,846.6842	16,163.2487	13,395.1240	13,324.7649	12,537.4326
maxPnomologBiggerThan	0.8000	N/A	N/A	0.7000	N/A	N/A	0.7000	N/A	N/A

Source: Own calculation in Stata 16MP 64-bit. Notes: Robust standard errors are presented in parentheses. All raw coefficients above parentheses are significant at 5% (*), 1% (**), and 1‰ (***), respectively.

Table 5. Cross-validations on sex and age using the first three alternate binary forms of the dependent variable.

Input Variable/Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	**justif_bribe_ 5_10**	**justif_bribe_ 5_10**	**justif_bribe_ 5_10**	**justif_bribe_ 6_10**	**justif_bribe_ 6_10**	**justif_bribe_ 6_10**	**justif_bribe_ 7_10**	**justif_bribe_ 7_10**	**justif_bribe_ 7_10**
Model Type	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**
F114A	0.1472 ***	0.1472 ***	0.1463 ***	0.1549 ***	0.1549 ***	0.1541 ***	0.1606 ***	0.1606 ***	0.1598 ***
	(0.0048)	(0.0079)	(0.0105)	(0.0058)	(0.0093)	(0.0102)	(0.0067)	(0.0042)	(0.0073)
F114D	0.2969 ***	0.2973 ***	0.2946 ***	0.2690 ***	0.2691 ***	0.2676 ***	0.2329 ***	0.2330 ***	0.2322 ***
	(0.0055)	(0.0063)	(0.0072)	(0.0060)	(0.0028)	(0.0084)	(0.0067)	(0.0017)	(0.0083)
F115	0.0995 ***	0.0993 ***	0.0980 ***	0.1095 ***	0.1095 ***	0.1083 ***	0.1052 ***	0.1052 ***	0.1046 ***
	(0.0053)	(0.0012)	(0.0065)	(0.0063)	(0.0029)	(0.0064)	(0.0074)	(0.0009)	(0.0083)
F116	0.3692 ***	0.3692 ***	0.3689 ***	0.3606 ***	0.3605 ***	0.3605 ***	0.3622 ***	0.3621 ***	0.3620 ***
	(0.0052)	(0.0077)	(0.0080)	(0.0058)	(0.0075)	(0.0073)	(0.0068)	(0.0049)	(0.0069)
F119	0.1367 ***	0.1370 ***	0.1370 ***	0.1448 ***	0.1448 ***	0.1448 ***	0.1369 ***	0.1368 ***	0.1372 ***
	(0.0049)	(0.0091)	(0.0045)	(0.0058)	(0.0038)	(0.0043)	(0.0067)	(0.0036)	(0.0060)
Y010	1.8602 ***	1.8689 ***	1.8400 ***	0.9335 ***	0.9331 ***	0.9220 ***	0.1470	0.1447	0.1397
	(0.0794)	(0.2259)	(0.0702)	(0.0934)	(0.2408)	(0.1019)	(0.1061)	(0.2630)	(0.0989)
_cons	−6.3675 ***	−6.3719 ***	−6.3628 ***	−6.7649 ***	−6.7638 ***	−6.7622 ***	−6.8293 ***	−6.8276 ***	−6.8279 ***
	(0.0440)	(0.1941)	(0.1284)	(0.0537)	(0.1950)	(0.1370)	(0.0602)	(0.1811)	(0.1155)
var(_cons[X001])	N/A	0.0008 **	N/A	N/A	0.0000 **	N/A	N/A	0.0000	N/A
	N/A	(0.0002)	N/A	N/A	(0.0000)	N/A	N/A	(0.0000)	N/A
var(_cons[X003R])	N/A	N/A	0.0096	N/A	N/A	0.0075	N/A	N/A	0.0060
	N/A	N/A	(0.0054)	N/A	N/A	(0.0045)	N/A	N/A	(0.0046)
N	122,545	122,474	122,166	122,545	122,474	122,166	122,545	122,474	122,166
Chi^2	18,816.6205	N/A	N/A	15,184.0298	N/A	N/A	12,985.2662	N/A	N/A
p	0.0000	N/A	N/A	0.0000	N/A	N/A	0.0000	N/A	N/A
Pseudo R^2	0.4489	N/A	N/A	0.4585	N/A	N/A	0.4504	N/A	N/A
maxAbsCorCoefPredMtrx	0.4995	N/A	N/A	0.4995	N/A	N/A	0.4995	N/A	N/A
AUCROC	0.9220	N/A	N/A	0.9271	N/A	N/A	0.9251	N/A	N/A

Table 5. *Cont.*

Input Variable/Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	**justif_bribe_ 5_10**	**justif_bribe_ 5_10**	**justif_bribe_ 5_10**	**justif_bribe_ 6_10**	**justif_bribe_ 6_10**	**justif_bribe_ 6_10**	**justif_bribe_ 7_10**	**justif_bribe_ 7_10**	**justif_bribe_ 7_10**
Model Type	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**
pGOF	0.0363	N/A	N/A	0.2091	N/A	N/A	0.0000	N/A	N/A
Chi^2GOF	82,080.56	N/A	N/A	81,681.32	N/A	N/A	83,315.83	N/A	N/A
AIC	48,401.9455	48,374.7190	48,169.0829	35,258.8082	35,236.9275	35,097.2370	28,095.9260	28,074.4344	27,968.0995
BIC	48,469.9591	48,394.1503	48,217.6486	35,326.8219	35,246.6431	35,145.8027	28,163.9396	28,084.1501	28,016.6651
maxPnomologBiggerThan	0.9500	N/A	N/A	0.9000	N/A	N/A	0.9000	N/A	N/A

Source: Own calculation in Stata 16MP 64-bit. Notes: Robust standard errors are presented in parentheses. All raw coefficients above parentheses are significant at 5% (*), 1% (**), and 1‰ (***), respectively.

Table 6. Cross-validations on sex and age using the last three alternate binary forms of the dependent variable.

Input Variable/Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	**justif_bribe_ 8_10**	**justif_bribe_ 8_10**	**justif_bribe_ 8_10**	**justif_bribe_ 9_10**	**justif_bribe_ 9_10**	**justif_bribe_ 9_10**	**justif_bribe_ 10**	**justif_bribe_ 10**	**justif_bribe_ 10**
Model Type	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**	**Logit— No Random Effects**	**MeLogit— Random Effects on Sex (X001)**	**MeLogit— Random Effects on Age (X003R)**
F114A	0.1617 ***	0.1617 ***	0.1613 ***	0.1683 ***	0.1683 ***	0.1685 ***	0.1743 ***	0.1743 ***	0.1745 ***
	(0.0078)	(0.0001)	(0.0110)	(0.0096)	(0.0057)	(0.0113)	(0.0116)	(0.0095)	(0.0118)
F114D	0.1953 ***	0.1952 ***	0.1952 ***	0.1728 ***	0.1727 ***	0.1731 ***	0.1491 ***	0.1490 ***	0.1498 ***
	(0.0076)	(0.0027)	(0.0083)	(0.0090)	(0.0182)	(0.0106)	(0.0105)	(0.0193)	(0.0141)
F115	0.0987 ***	0.0987 ***	0.0981 ***	0.0899 ***	0.0899 ***	0.0890 ***	0.0756 ***	0.0756 ***	0.0748 ***
	(0.0087)	(0.0001)	(0.0100)	(0.0106)	(0.0093)	(0.0139)	(0.0125)	(0.0033)	(0.0162)
F116	0.3742 ***	0.3742 ***	0.3738 ***	0.3917 ***	0.3917 ***	0.3905 ***	0.4122 ***	0.4123 ***	0.4111 ***
	(0.0084)	(0.0029)	(0.0069)	(0.0108)	(0.0038)	(0.0092)	(0.0136)	(0.0048)	(0.0127)
F119	0.1170 ***	0.1170 ***	0.1174 ***	0.0937 ***	0.0937 **	0.0942 ***	0.0656 ***	0.0656	0.0663 ***

Table 6. *Cont.*

Input Variable/Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	justif_bribe_ 8_10	justif_bribe_ 8_10	justif_bribe_ 8_10	justif_bribe_ 9_10	justif_bribe_ 9_10	justif_bribe_ 9_10	justif_bribe_ 10	justif_bribe_ 10	justif_bribe_ 10
Model Type	Logit—No Random Effects	MeLogit—Random Effects on Sex (X001)	MeLogit—Random Effects on Age (X003R)	Logit—No Random Effects	MeLogit—Random Effects on Sex (X001)	MeLogit—Random Effects on Age (X003R)	Logit—No Random Effects	MeLogit—Random Effects on Sex (X001)	MeLogit—Random Effects on Age (X003R)
	(0.0079)	(0.0164)	(0.0102)	(0.0095)	(0.0294)	(0.0108)	(0.0111)	(0.0362)	(0.0133)
Y010	22,120.5917 ***	22,120.5913 *	22,120.5779 ***	22,121.0694 ***	22,121.0696 ***	22,121.0437 ***	22,121.2979 ***	22,121.2985 ***	22,121.2715 ***
	(0.1200)	(0.3007)	(0.1168)	(0.1412)	(0.1595)	(0.1381)	(0.1602)	(0.1402)	(0.1848)
_cons	22,126.8296 ***	22,126.8290 ***	22,126.8319 ***	22,127.0661 ***	22,127.0653 ***	22,127.0730 ***	22,127.2663 ***	22,127.2656 ***	22,127.2673 ***
	(0.0674)	(0.0927)	(0.0854)	(0.0824)	(0.0659)	(0.0750)	(0.0999)	(0.0957)	(0.1012)
var(_cons[X001])	N/A	0.0000	N/A	N/A	0.0000	N/A	N/A	0.0000	N/A
	N/A	(0.0000)	N/A	N/A	(0.0000)	N/A	N/A	(0.0000)	N/A
var(_cons[X003R])	N/A	N/A	0.0046	N/A	N/A	0.0064	N/A	N/A	0.0078
	N/A	N/A	(0.0038)	N/A	N/A	(0.0043)	N/A	N/A	(0.0043)
N	122,545	122,474	122,166	122,545	122,474	122,166	122,545	122,474	122,166
Chi^2	10,868.6143	N/A	N/A	8177.9460	N/A	N/A	6141.2345	N/A	N/A
p	0.0000	N/A	N/A	0.0000	N/A	N/A	0.0000	N/A	N/A
Pseudo R^2	0.4373	N/A	N/A	0.4336	N/A	N/A	0.4188	N/A	N/A
maxAbsCorCoefPredMtrx	0.4995	N/A	N/A	0.4995	N/A	N/A	0.4995	N/A	N/A
AUCROC	0.9176	N/A	N/A	0.9145	N/A	N/A	0.9071	N/A	N/A
pGOF	0.0000	N/A	N/A	0.0000	N/A	N/A	0.0000	N/A	N/A
Chi^2GOF	89,944.11	N/A	N/A	99,321.47	N/A	N/A	1.1×10^5	N/A	N/A
AIC	22,303.0334	22,288.2163	22,222.0906	16,841.7315	16,827.9683	16,795.8192	13,327.1104	13,315.8519	13,287.8886
BIC	22,371.0470	22,297.9320	22,270.6563	16,909.7451	16,837.6839	16,844.3849	13,395.1240	13,335.2832	13,336.4543
maxPnomologBiggerThan	0.8000	N/A	N/A	0.7000	N/A	N/A	0.7000	N/A	N/A

Source: Own calculation in Stata 16MP 64-bit. Notes: Robust standard errors are presented in parentheses. All raw coefficients above parentheses are significant at 5% (*), 1% (**), and 1‰ (***), respectively.

Table 7. Determining and testing the bribe index using justif_bribe_6_10 as a binary form of the outcome.

Input Variable/Model	(1)	(2)	(3)	(4)	(5)	(6)
Model Type	**Scobit**	**Scobit**	**Logit**	**Logit**	**Probit**	**Probit**
F114A	0.0075 ***		0.0075 ***		0.0073 ***	
	(0.0002)		(0.0002)		(0.0002)	
F114D	0.0129 ***		0.0129 ***		0.0131 ***	
	(0.0002)		(0.0002)		(0.0002)	
F116	0.0166 ***		0.0164 ***		0.0171 ***	
	(0.0002)		(0.0002)		(0.0002)	
brb_idx_scobit3x		0.0370 ***		0.0367 ***		0.0377 ***
		(0.0003)		(0.0003)		(0.0003)
N	151,636	151,636	151,636	151,636	151,636	151,636
Chi^2	N/A	N/A	18,188.9132	18,198.9719	20,218.0620	20,074.9742
p	N/A	N/A	0.0000	0.0000	0.0000	0.0000
Pseudo R^2	N/A	N/A	0.4291	0.4291	0.4294	0.4294
maxAbsCorCoefPredMtrx	0.4804	N/A	0.4804	N/A	0.4804	N/A
AUC	N/A	N/A	0.9169	0.9169	0.9169	0.9169
pGOF	N/A	N/A	0.0000	0.0000	0.0000	0.0000
Chi^2GOF	N/A	N/A	2790.76	2790.11	2556.99	2560.84
AIC	43,478.2267	43,474.2266	43,625.4103	43,621.7071	43,597.3988	43,596.4702
BIC	43,527.8729	43,504.0143	43,665.1273	43,641.5656	43,637.1158	43,616.3287
maxPnomologBiggerThan	N/A	N/A	0.9000	0.9000	N/A	N/A

Source: Own calculation in Stata 16MP 64-bit. Notes: Robust standard errors are presented in parentheses. All coefficients above parentheses are computed as average marginal effects and are significant at 1‰ (***).

```
scobit justif_bribe_6_10 F114A F114D F116, vce(robust)
margins, dydx(*)
generate sum_coeff=0.0075054+0.0129164+0.0166124
generate brb_idx_scobit3x=0.0075054/sum_coeff*F114A+0.0129164/sum_coeff*F114D+0.0166124/sum_coeff*F116
```

Figure 4. Stata processing script for deriving the proposed bribe index based on a Scobit regression with resulting coefficients expressed as average marginal effects (Table 7, Model 1). Source: Own calculation in Stata 16MP 64-bit.

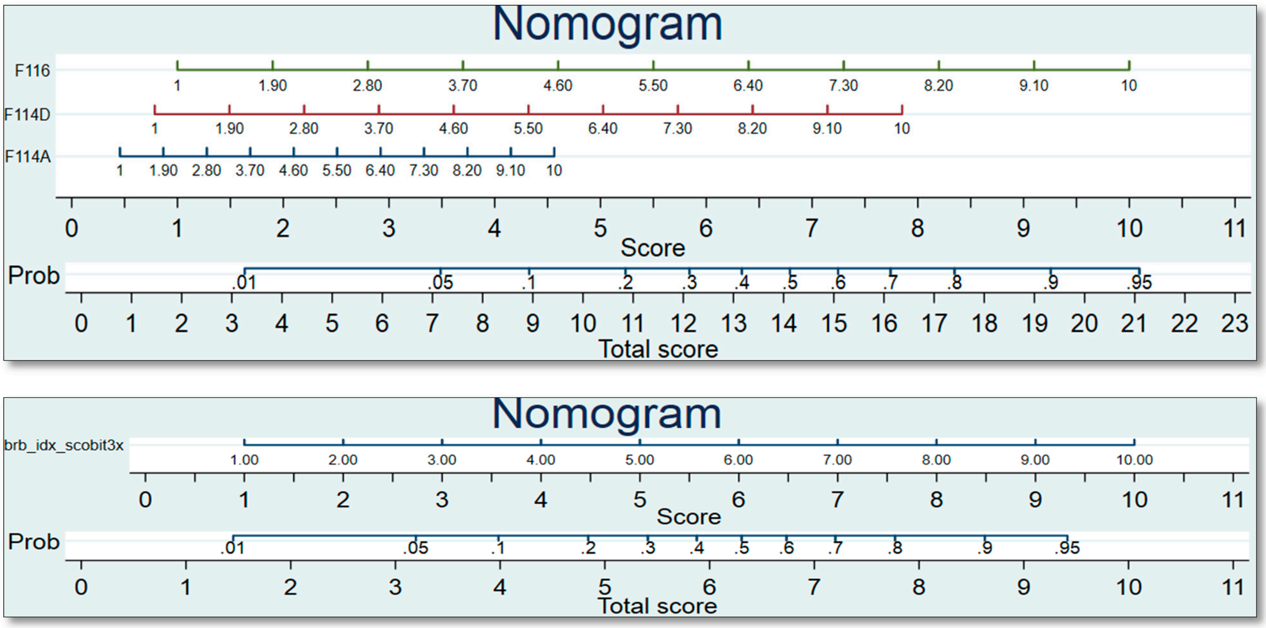

Figure 5. Two comparable prediction nomograms for estimating the bribery tolerance risk. Source: Own calculations using the nomolog command in Stata 16 MP 64-bit.

4. Discussion and Conclusions

This paper puts forward a robust methodology for identifying the major determinants of considering bribes an acceptable practice and subsequently computing a corresponding index, namely the one of bribery tolerance.

Our empirical model is highly parsimonious and indicates three major influences (from an initial list of 13 variables from the WVS dataset) that can predict bribery tolerance with a 91.69% accuracy of classification: the attitude towards cheating on taxes, the attitude towards claiming government benefits to which a person is not entitled, and the attitude towards violence against other people. The first two individual attitudes reflect upon people's relationship with the state and are considered "the usual suspects" for actual participation in corrupt practices in numerous studies conducted in transition [38,39] and developing economies [40]. Thus, our findings align with the literature for the narrower case of bribery tolerance and they further expand their predictive validity by covering a much larger spectrum of 105 worldwide countries. The fact that these practices are positioned in some countries, in terms of people's perception, at the lower and middle spectrum of crimes [41–43], but they pave the way for the serious case of corruption, makes them even more relevant in terms of government priorities on changing public views on critical topics. The third factor related to interpersonal violence was not previously examined in association with corruption, thus further research is needed to unveil the psychological mechanisms supporting this link.

The entire methodological approach proposed and described in this paper is based on the scientific principles of triangulation, cross-validation, and reproducibility [44]. Triangulation means that we used various methods, techniques, and applications (Naive Bayes, LASSO variable selection techniques, different types of regression analysis, mixed-effects modeling, average marginal effects, post-estimations of accuracy and goodness of fit, dynamic thresholds for variance inflation factors, risk prediction nomograms, etc.) and we have got results that agree across them. Cross-validation [45] means that we tested using many subsamples randomly (CVLASSO) and not randomly (clusters using four criteria) extracted from the entire dataset. Reproducibility [46] means that our findings are 100% replicable starting from the public dataset we have used and from all metadata and steps described as detailed as possible in this paper. To that extent, the procedure can be useful to better understand other complex socio-economic phenomena, beyond our case example of bribery tolerance.

Author Contributions: Conceptualization, A.-P.P., D.H. and R.I.-C. Data curation, A.-P.P. and D.H., Formal analysis, A.-P.P. and D.H. Investigation, A.-P.P. and D.H. Methodology, A.-P.P., D.H. and R.I.-C. Project administration, R.I.-C. Resources, A.-P.P., D.H. and R.I.-C. Software, A.-P.P. and D.H. Supervision, A.-P.P. and R.I.-C. Validation, A.-P.P., D.H. and R.I.-C. Visualization, D.H. Writing—original draft, D.H. and R.I.-C. Writing—review, and editing, A.-P.P., D.H. and R.I.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data belongs to World Values Survey (the WVS TimeSeries 1981 2020 Stata v1 6.zip file available at https://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp and accessed on 1 February 2021).

Acknowledgments: For allowing us to explore the dataset and publish this paper, we would like to thank the World Values Survey and supporting projects. In terms of technical aid (https://cloud.raas.uaic.ro/—last accessed on 9 June 2021), this paper is supported by the Competitiveness Operational Programme Romania under project number SMIS 124759—RaaS-IS (Research as a Service Iasi) id POC/398/1/124759, coordinated by Marin FOTACHE, to whom we are grateful.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

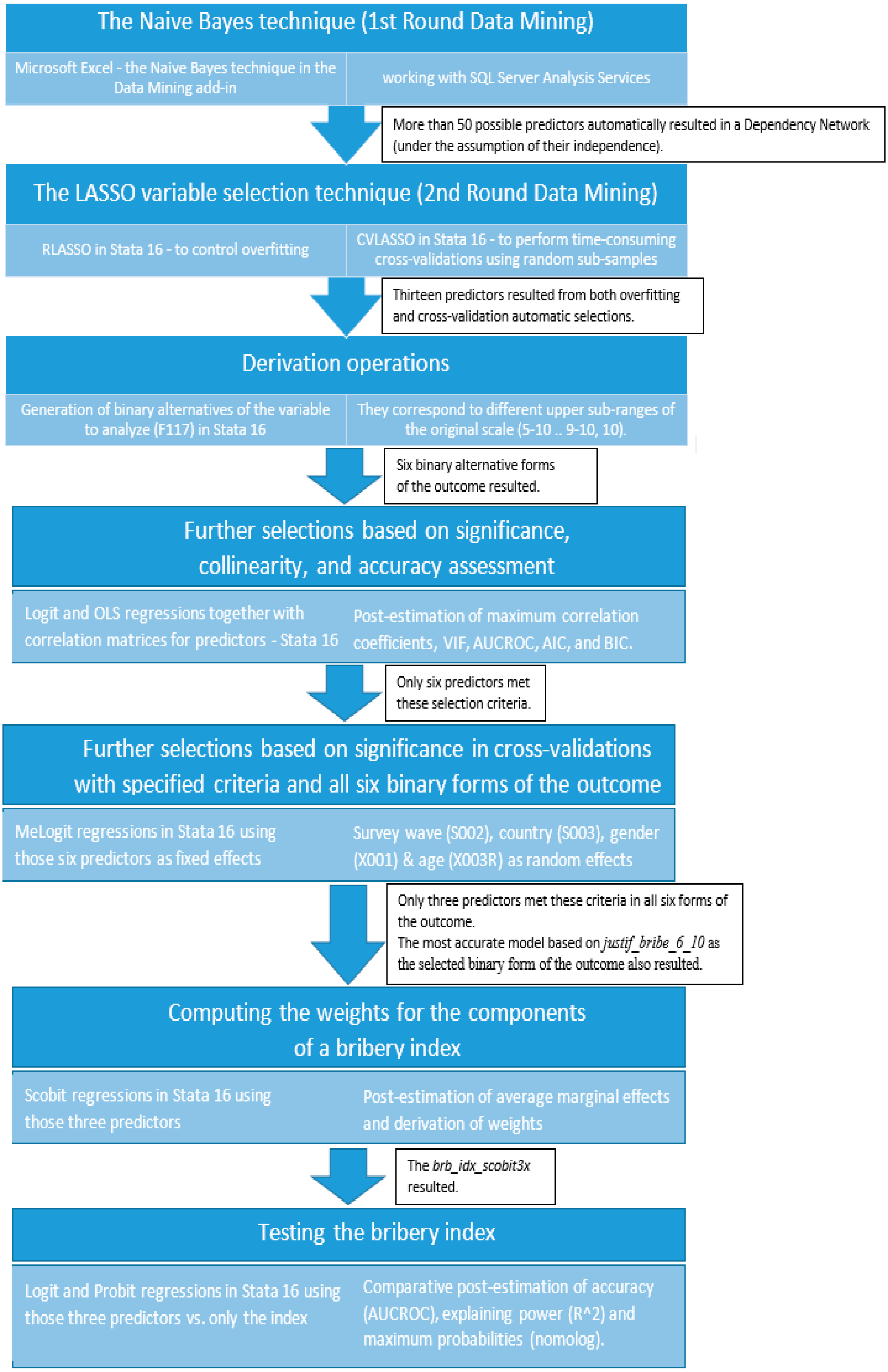

Figure A1. Schematic representation of the techniques used. Source: The authors' own projection.

References

1. King, G. Ensuring the data-rich future of the social sciences. *Science* **2011**, *331*, 719–721. [CrossRef]
2. Einav, L.; Levin, J. Economics in the age of big data. *Science* **2014**, *346*, 1243089. [CrossRef]
3. Booysen, F. An overview and evaluation of composite indices of development. *Soc. Indic. Res.* **2002**, *59*, 115–151. [CrossRef]
4. De Muro, P.; Mazziotta, M.; Pareto, A. Composite indices of development and poverty: An application to MDGs. *Soc. Indic. Res.* **2011**, *104*, 1–18. [CrossRef]
5. Shaker, R.R.; Zubalsky, S.L. Examining patterns of sustainability across Europe: A multivariate and spatial assessment of 25 composite indices. *Int. J. Sustain. Dev. World Ecol.* **2015**, *22*, 1–13. [CrossRef]
6. Horodnic, I.A.; Williams, C.C.; Ianole-Calin, R. Does higher cash-in-hand income motivate young people to engage in under-declared employment? *East. J. Eur. Stud.* **2020**, *11*, 48–69.
7. Turturean, C.I.; Asandului, L.A.; Chirila, C.; Homocianu, D. Composite index of sustainable development of EU countries'economies (ISDE-EU). *Transform. Bus. Econ.* **2019**, *18*, 586–605.
8. Yoneoka, D.; Saito, E.; Nakaoka, S. New algorithm for constructing area-based index with geographical heterogeneities and variable selection: An application to gastric cancer screening. *Sci. Rep.* **2016**, *6*, 26582. [CrossRef] [PubMed]
9. Druică, E.; Vâlsan, C.; Ianole-Călin, R.; Mihail-Papuc, R.; Munteanu, I. Exploring the Link between Academic Dishonesty and Economic Delinquency: A Partial Least Squares Path Modeling Approach. *Mathematics* **2019**, *7*, 1241. [CrossRef]
10. Wheeler, D.C. Simultaneous coefficient penalization and model selection in geographically weighted regression: The geographically weighted lasso. *Environ. Plan.* **2009**, *41*, 722–742. [CrossRef]
11. Nakaya, T. Evaluating socioeconomic inequalities in cancer mortality by using areal statistics in Japan: A note on the relation between the municipal cancer mortality and the areal deprivation index. *Proc. Inst. Stat. Math.* **2011**, *59*, 239–265.
12. Hindman, M. Building better models: Prediction, replication, and machine learning in the social sciences. *Ann. Am. Acad. Political Soc. Sci.* **2015**, *659*, 48–62. [CrossRef]
13. Requejo-Castro, D.; Giné-Garriga, R.; Pérez-Foguet, A. Data-driven Bayesian network modelling to explore the relationships between SDG 6 and the 2030 Agenda. *Sci. Total. Environ.* **2020**, *710*, 136014. [CrossRef]
14. Imani, M.; Ghoreishi, S.F. Two-Stage Bayesian Optimization for Scalable Inference in State-Space Models. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef] [PubMed]
15. Dixon, M.F.; Halperin, I.; Bilokon, P. *Machine Learning in Finance. From Theory to Practice*; Springer Nature: Cham, Switzerland, 2020. [CrossRef]
16. Chabova, K. Measuring corruption in Europe: Public opinion surveys and composite indices. *Qual. Quant.* **2017**, *51*, 1877–1900. [CrossRef]
17. Fazekas, M.; Tóth, I.J.; King, L.P. Anatomy of grand corruption: A composite corruption risk index based on objective data. In *Corruption Research Center Budapest Working Papers No. CRCB-WP/2013, 2*; Institute of Economics, Centre for Economic and Regional Studies: Budapest, Hungary, 2013.
18. Villarino, J.M.B. Measuring corruption: A critical analysis of the existing datasets and their suitability for diachronic transnational research. *Soc. Indic. Res.* **2021**, 1–39. [CrossRef]
19. Dobrowolski, Z. *Combating Corruption and Other Organizational Pathologies*; Peter Lang: Frankfurt Am Main, Germany, 2016; pp. 100–106. [CrossRef]
20. Lambsdorff, J.G. The Methodology of the Corruption Perceptions Index 2007. Internet Center for Corruption Research. 2007. Available online: http://www.icgg.org/corruption.cpi_2006.html (accessed on 1 June 2021).
21. Levy, S.; Fullagar, P.K. Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics* **1981**, *46*, 1235–1243. [CrossRef]
22. Santosa, F.; Symes, W.W. Linear Inversion of Band-Limited Reflection Seismograms. *SIAM J. Sci. Stat. Comput.* **1986**, *7*, 1307–1330. [CrossRef]
23. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Society. Ser. B* **1996**, *58*, 267–288. [CrossRef]
24. Mukaka, M. A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **2012**, *24*, 69–71. [PubMed]
25. Jiménez-Valverde, A. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Glob. Ecol. Biogeogr.* **2012**, *21*, 498–507. [CrossRef]
26. Bewick, V.; Cheek, L.; Ball, J. Review. Statistics review 14: Logistic regression. *Crit. Care* **2005**, *9*, 112–118. [CrossRef]
27. Lin, M.; Lucas, H.; Shmueli, G. Too big to fail: Large samples and the p-value problem. *Inf. Syst. Res.* **2013**, *24*, 906–917. [CrossRef]
28. Dziak, J.J.; Coffman, D.L.; Lanza, S.T.; Li, R.; Jermiin, L.S. Sensitivity and Specificity of Information Criteria. *Brief. Bioinform.* **2020**, *21*, 553–565. [CrossRef]
29. Kéry, M.; Royle, J.A. Modeling Static Occurrence and Species Distributions Using Siteoccupancy Models. *Appl. Hierarchical Modeling Ecol.* **2016**, 551–629. [CrossRef]
30. Miles, J.R. Squared, adjusted r squared. In *Encyclopedia of Statistics in Behavioral Science*; Wiley: New York, NY, USA, 2005. [CrossRef]
31. Nagler, J. An alternative estimator to Logit and Probit. *Am. J. Political Sci.* **1994**, *38*, 230–255. [CrossRef]
32. Zlotnik, A.; Abraira, V. A general-purpose nomogram generator for predictive logistic regression models. *Stata J.* **2015**, *15*, 537–546. [CrossRef]

33. Yao, Z.; Holmbom, A.H.; Eklund, T.; Back, B. Combining Unsupervised and Supervised Data Mining Techniques for Conducting Customer Portfolio Analysis. In *Advances in Data Mining. Applications and Theoretical Aspects*; Perner, P., Ed.; ICDM 2010. Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6171. [CrossRef]
34. Vatcheva, K.P.; Lee, M.J.; McCormick, J.B.; Rahbar, M.H. Multi-collinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology* **2016**, *6*, 227. [PubMed]
35. Shrestha, N. Detecting Multicollinearity in Regression Analysis. *Am. J. Appl. Math. Stat.* **2020**, *8*, 39–42. [CrossRef]
36. Freund, R.J.; Wilson, W.J.; Sa, P. *Regression Analysis: Statistical Modeling of a Response Variable*, 2nd ed.; Academic Press: Cambridge, MA, USA, 2006.
37. Ten Have, T.R.; Kunselman, A.R.; Tran, L. A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. *Stat. Med.* **1999**, *18*, 947–960. [CrossRef]
38. Horodnic, I.; Rodgers, P.; Williams, C.; Momtazian, L. (Eds.) *The Informal Economy: Exploring Drivers and Practices*; Routledge: London, UK, 2017.
39. Vâlsan, C.; Druică, E.; Ianole-Călin, R. State capacity and tolerance towards tax evasion: First evidence from Romania. *Adm. Sci.* **2020**, *10*, 33. [CrossRef]
40. Shafiq, M.N. Aspects of Moral Change in India, 1990–2006: Evidence from Public Attitudes toward Tax Evasion and Bribery. *World Dev.* **2015**, *68*, 136–148. [CrossRef]
41. James, S.; McGee, R.W.; Benk, S.; Budak, T. How seriously do taxpayers regard tax evasion? A survey of opinion in England. *J. Money Laund. Control.* **2019**. Available online: https://www.emerald.com/insight/content/doi/10.1108/JMLC-09-2018-0056/full/html (accessed on 1 June 2021).
42. McGee, R.W.; Devos, K.; Benk, S. Attitudes towards tax evasion in Turkey and Australia: A comparative study. *Soc. Sci.* **2016**, *5*, 10. [CrossRef]
43. Aljaaidi, K.S.Y.; Manaf, N.A.A.; Karlinsky, S.S. Tax evasion as a crime: A survey of perception in Yemen. *Int. J. Bus. Manag.* **2011**, *6*, 190. [CrossRef]
44. Munafò, M.R.; Smith, G.D. Robust research needs many lines of evidence. *Nature* **2018**, *553*, 399–401. [CrossRef]
45. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
46. Baker, M. 1500 Scientists Lift the Lid on Reproducibility. *Nature* **2016**, *533*, 452–454. [CrossRef] [PubMed]