Asia-Pacific Journal of Business (아태비즈니스연구)
Vol. 12, No. 4, December 2021 (pp.13-40)
https://doi.org/10.32599/apjb.12.4.202112.13

APJB
ISSN 2233-5900 (Print)
ISSN 2384-3934 (Online)

Corporate Corruption Prediction Evidence From Emerging Markets*

Yang Sok Kim[a], Kyunga Na[b], Young-Hee Kang[c]

[a]Major in Management Information Systems, Keimyung University, South Korea
[b]Department of International Business, Chungbuk National University, South Korea
[c]Major in Business Administration, Keimyung University, South Korea

Received 30 November 2021, Revised 20 December 2021, Accepted 23 December 2021

Abstract

Purpose - The purpose of this study is to predict corporate corruption in emerging markets such as Brazil, Russia, India, and China (BRIC) using different machine learning techniques. Since corruption is a significant problem that can affect corporate performance, particularly in emerging markets, it is important to correctly identify whether a company engages in corrupt practices.
Design/methodology/approach - In order to address the research question, we employ predictive analytic techniques (machine learning methods). Using the World Bank Enterprise Survey Data, this study evaluates various predictive models generated by seven supervised learning algorithms: k-Nearest Neighbour (k-NN), Naïve Bayes (NB), Decision Tree (DT), Decision Rules (DR), Logistic Regression (LR), Support Vector Machines (SVM), and Artificial Neural Network (ANN).
Findings - We find that DT, DR, SVM and ANN create highly accurate models (over 90% of accuracy). Among various factors, firm age is the most significant, while several other determinants such as source of working capital, top manager experience, and the number of permanent full-time employees also contribute to company corruption.
Research implications or Originality - This research successfully demonstrates how machine learning can be applied to predict corporate corruption and also identifies the major causes of corporate corruption.

Keywords: BRIC, Corporate Corruption, Emerging Markets, Machine Learning
JEL Classifications: C14

Ⅰ. Introduction

Transparency International, an international anti-corruption organization, believes that corruption is '*the abuse of entrusted power for private gain*' (Transparency International, 2016a) while the World Bank defines it as '*the abuse of public power for private benefit*' (Tanzi, 1998). Corruption is a major problem because '*it corrodes the fabric of society. It undermines people's trust in political and economic systems, institutions and leaders. It can cost people their freedom, health, money – and sometimes their lives*' (Transparency International, 2016b). There

* This work was supported by the Chungbuk National University Research Fund of 2020.
[a] First Author, E-mail: yangsok.kim@gw.kmu.ac.kr
[b] Corresponding Author, E-mail: kna@chungbuk.ac.kr
[c] Co-Author, E-mail: kang02@gw.kmu.ac.kr

is much research on corruption from both economic and political viewpoints (Tanzi, 1998; Lambsdorff, 1999; Jain, 2001). In this light, prior research has focused more on country level analysis of corruption – addressing its economic, political and social impacts – than on the business level analysis of corruption.

However, it is also important to look at corruption from the business level to be able to identify the signs of corruption within businesses. With this knowledge, the government can well prepare its anti-corruption measures and thus promote fair competition among companies. The probability that a certain company is corrupt is information that is also important to potential clients in the financial market. For instance, as bribery tends to increase transaction costs (Wei, 2000) and decrease efficiency (Kaufmann and Wei, 1999), investors are more likely to place their money in ethical companies rather than corrupt ones in order to maximize returns.

In order to predict corporate corruption, we employ predictive analytic techniques, which are known as machine learning in computer science. These techniques have been used extensively to generate predictive models. Many machine learning algorithms have been developed to address different aspects of the learning task. For this research, we use the following algorithms: k-Nearest Neighbour (K-NN), Naïve Bayes (NB), Decision Tree (DT), Decision Rules (DR), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). We discuss the details of these techniques in Section 3.

To use machine learning techniques, it is necessary to have data that can represent our research question. To this end, we use data collected by the World Bank known as 'Enterprise Surveys' (http://www.enterprisesurveys.org/). The World Bank chooses several countries and conducts this survey annually; consequently, the survey data are not collected in the same year, and contain large amounts of information unrelated to corruption. We are specifically interested in businesses in Brazil, Russia, India and China (BRIC) because their economies have achieved fast growth while ethical entrepreneurship has not followed suite. Previous research shows that businesses in emerging markets such as BRIC are more likely to be corrupt compared to those in developed (Cuervo-Cazurra, 2006; Sanchez et al., 2008). Therefore, in order to address the research question of this paper, it is necessary to choose relevant features. We discuss the feature selection approach in Section 4.

This research has two objectives. First, we aim to find whether the predictive analytic techniques can be used to predict company corruption, and if so, to evaluate the predictive performances of the algorithms. Second, this research aims to provide a reasonable explanation to corruption prediction. To this end, we set up an experimental procedure and compare the performances of the created models, and extract explanations from applicable models. Evaluation results and prediction explanations are discussed in Section 5.

This study suggests a new approach for analysing business corruption. In particular, it addresses the question of how to predict corruption in individual companies. Although this study has some valuable results, it has limitations that should be addressed in future studies, which is discussed in Section 6.

II. Related Research

2.1 Consequences of Corruption

Corruption can be defined in many different ways. The most popular definition is the abuse

of public power for private benefit. According to (Tanzi, 1998), acts of corruption can be classified into the following categories – (1) bureaucratic or political, (2) cost-reducing or benefit-enhancing, (3) briber-initiated or bribee-initiated, (4) coercive or collusive, (5) centralized or decentralized, (6) predictable or arbitrary, and (7) involving cash payments or not. There are conflicting views on the consequences of corruption. On the one hand, corruption negatively impacts the economy. It creates bureaucratic hurdles to demand bribes and thus decrease efficiency (Myrdal, 1968), and adversely affects the provision of social services (Gupta et al., 2000). Lambsdorff (2006) lists the following as the consequences of corruption: inequality, low productivity (e.g. GDP per capita and GDP growth), reduction of investment (e.g., overall investment and composition of investment), misallocation of public resources (e.g., budget allocation distortions and reduced public sector quality) and low security of property rights and misallocation of private resources (e.g., distortion of markets, underground economies and tax cheating). Much empirical research supports these outcomes. Mauro(1995) reports that corruption reduces investment and thus reduces the rate of growth. Yanzi and Davoo아 (1998) find similar empirical results, and also find that corruption can reduce productivity. Gupta et al.(2002) report that corruption has a positive relationship with income inequality and poverty, supported by the increase of the Gini coefficient of income inequality. Gupta et al.(2000) state that high levels of corruption have adverse impact on provision of health care and education services. Habib and Zurawicki (2002) report that corruption has a negative relationship with foreign direct investment because foreign investors consider corruption wrong and operationally inefficient. Several researchers have reported negative relationships between corruption and tax revenues (Imam and Jacobs, 2014; Tanzi and Davoodi, 1998/2000; Brasoveanu and Brasoveanu, 2009). On the other hand, corruption can have positive effects on the economy. It improves social welfare, both because it helps avoid cumbersome regulations and because it can work as a rewarding system for badly paid bureaucrats (Leff, 1964). Contrary to Habib and Zurawicki (2002), Egger and Winner (2005) show positive relationships between corruption and direct foreign investment after analysing 73 developed and developing countries.

2.2 Causes of Corruption

2.2.1. Determining Factors of Corruption

Tanzi (1998) classifies the determining factors of corruption into two broad categories – direct factors and indirect factors. The direct factors include regulations and authorizations, taxations, public spending decisions (e.g., investment project, procurement spending, extra-budgetary accounts), provision of goods and services at discount (e.g., foreign exchange, credit, electricity and water, rationed goods, access to educational and health facilities, access to public land and housing), discretionary decisions on various resources, and financing of parties. The indirect factors relate to the quality of bureaucracy, public sector wages, penalty framework, institutional influences, transparency of rules, laws and processes, and leadership. Treisman (2000) identifies the following factors as determining factors – history and tradition (e.g., Protestant traditions, histories of British rule), economic development (e.g., proportion of imports), political institutions (e.g., federal states and their degree of democracy) and public policy. Following Treisman (2000), Pellegrini (2011) classifies the causes of corruption into two broad categories – historical roots and contemporary causes. He includes legal theory,

British colonization theory, Protestant religion theory and ethnolinguistic fractionalization as the historical roots, and economic development, rent-seeking behaviours, social institutions and political stability as the contemporary causes. In addition to these factors, several other factors have been considered as determining factors, suc as gender (Dollar et al., 2001; Swamy et al., 2001) and press freedom (Brunetti and Weder, 2003; Freille et al., 2007).

2.2.2. Empirical Analysis of Determining Factors

Ades and Di Tella (1999) demonstrate that the level of corruption can be controlled by economic openness, measured by the ratio of imports to GDP. However, Majeed (2014) argues that it is not just economic openness that can reduce corruption, but also the complementary policies. Elliott (1997) argues that government involvement in the economy, measured by the ratio of public budget to GDP, can be related to the level of corruption since it may promote monopolies and discourage open and fair competition among business players, which in turn may encourage corruption. Montalvo and Reynal-Querol (2005) show that ethnic diversity has a positive correlation with control of corruption. However, Churchill et al. (2013) argue that corruption has a negative relationship to ethnic diversity because people may put the interest of their ethnic group above the interest of the nation. Bhattacharyya and Rolder (2010) and Treisman (2000) report that the level of democracy and corruption are positively related. Freille et al. (2007) state that press freedom has a negative correlation to the level of corruption since it enables the media to publish balanced news about the government and promote accountability mechanism by the public. Leite and Weidman (1999) report high proportions of exports (economic openness) from natural resources have positive relationships with corruption. Fatic (2000) reports that the stability of politics, economy and society have positive relationships with successful corruption control, since they can promote transparent corruption monitoring, assessment and control within the government and citizenry. In fact, political stability ensures law enforcement and law enforcement ensures control of corruption. Eiras (2003) reports that the lack of economic freedom, which can be caused by the lack of rule of law, over-regulation and a large public sector, could force citizens to become involved in informal economic activities, instead of formal activities, a situation that is fertile for corruption. Churchill et al. (2013) suggest a list of factors that are related to the level of corruption as summarized in Table 1.We added additional references that discuss the same independent variables.

Table 1. Determining Factors of Corruption

Independent Variables	References	Category	Relationship
Economic openness	(Ades and Di Tella, 1999), (Majeed, 2014)	Economic	Negative / Positive
Public budget	(Delavallade, 2006), (Tanzi and Davoodi, 1998), (Arikan, 2004)	Economic	Negative
Natural resources	(Leite and Weidmann, 1999), (Bhattacharyya and Hodler, 2010)	Economic	Negative
Economic freedom	(Eiras, 2003), (Shen and Williamson, 2005), (Swaleheen and Stansel, 2007), (Graeff and Mehlkop, 2003), (Pieroni and d'Agostino, 2013)	Political	Positive
Quality of democracy	(Treisman, 2000), (Chowdhury, 2004), Shen, C. and J. B. Williamson (2005)	Political	Positive

Press freedom	(Brunetti and Weder, 2003), (Chowdhury, 2004)	Political	Positive
Political stability	(Fredriksson and Svensson, 2003), (Nur-tegin and Czap, 2012)	Political	Positive
Ethnic diversity	(Yehoue, 2007), (Stendahl, 2016)	Social	Negative
Quality of regulation	(Churchill et al., 2013),	Social	Positive
Religion	(North et al., 2013)	Social	Positive

2.2.3 Summary

In this section, we summarized previous research which discusses the causes of corruption. As De Graaf (2007) argues, most of the prior research do not try to address'actual, individual corruption cases'and thus need'more contextual research'. Following this claim, we mainly focus on actual, individual corruption cases and aim to reveal contextual patterns that explain each individual firm's possible corrupting sources.

Ⅲ. Techniques

3.1 RapidMiner - A Predictive Analytic Platform

Predictive analytic techniques help determine the class label of a given example based on the training data. There are many software platforms, including IBM Predictive Analytics, RapidMiner, TIBCO Analytics, Oracle Data Mining (ODM), KNIME, SAS Predictive Analytics, etc. Formerly known as YALE (Yet Another Learning Environment), RapidMiner was developed in 2001 by a group of researchers at the Technical University of Dortmund (Mierswa et al., 2006). We use RapidMiner to conduct predictive analytics because it has an easy-to-use graphical interface and powerful analytic capability. It is ranked as one of the top five leaders in the analytic platforms by Gartner Research (Kart et al., 2016).

3.2 Learning Algorithms

This section summarizes seven algorithms provided by RapidMiner. The following terms are used throughout this paper.

- **Examples:** A set of cases used for the analysis. In our analysis, each example is described by a set of attribute values and has a class attribute called'label'. Examples are divided into training and testing examples.
- **Training dataset:** A subset of examples selected for learning the predictive model.
- **Testing dataset:** A subset of examples selected for demonstrating the effectiveness of the predictive model in its predictions.
- **Predictive model:** The model learned from the training dataset by applying the learning algorithm. Given an example of the testing dataset, the predictive model predicts a class value (label) based on its attribute values. The predicted label can differ from the actual label of the given example.

3.2.1 k-Nearest Neighbour

K-Nearest Neighbour (k-NN) algorithm does not create an explicit predictive model. Instead, it predicts a class label using k examples of the training dataset that are the most similar to the given example, called the '*nearest neighbours*' of the given example. In order to apply k-NN, it is necessary to set three parameters. First, parameter k is used to specify the number of the nearest neighbours. Second, a distance (similarity) measure should be defined to find the nearest neighbours. Applicable similarity measures for different attribute value types are summarized in Table 2. Since the dataset contains mixed attribute value types, we employed Mixed Euclidean Distance as the similarity measure. Third, it is necessary to set the aggregation function. A simple approach is to use a simple majority voting approach, which counts the class labels of the nearest neighbours and chooses the label that has highest count. It is also possible to apply a weighted voting approach, which assigns different weights to different labels.

Table 2. Similarity Measures

Attribute Value Type	Similarity Measure
Numeric Type	Euclidean Distance, Camberra Distance, Chebychev Distance, Correlation Similarity, Cosine Similarity, Dice Similarity, Dynamic Time Warping Distance, Inner Product Similarity
Nominal Type	Nominal Distance, Dice Similarity, Jaccard Similarity, Kulczynski Similarity, Rogers Tanimoto Similarity, Russell Rao Similarity, Simple Matching Similarity
Mixed Type	Mixed Euclidean Distance

3.2.2 Naïve Bayes

Naïve Bayes (NB) is a simple probabilistic algorithm for classification. It is based on Bayes' theorem with strong (naive) independence assumptions. That is, NB assumes that the state a particular attribute of a class is unrelated to that of other attributes. The label with the highest posterior probability is the predicted label for a given example.

$$\hat{y}(x) = \arg\max(k)\{P(Y=k \mid X=x)\} = \arg\max(k)\{P(X=x \mid Y=k)P(Y=k)\}$$

In the above, Y is categorical target variable, k is a specific label, X is categorical predictor vector and x is a specific example. The advantage of the Naive Bayes algorithm is that it requires a small training dataset to estimate the means and variances of the variables relevant to the classification task.

3.2.3 Decision Tree

Decision Tree (DT) learns a decision tree, which is a tree-like graph or model. Many different algorithms have been suggested by researchers, including C4.5 (Quinlan, 1992), CART (Breiman et al., 1984), and CHAID (Biggs et al., 1991). In this research, we use the DT algorithm implemented in RapidMiner, which is close to C4.5. Generally, the algorithm works as follows:

- Step 1: Select an attribute A used for the dataset split. It is important to choose a good attribute at each stage for a useful decision tree. Information gain (IG), gain ratio (GR), and Gini index (GI) are the most popular criteria.
- Step 2: Divide the dataset into subsets by using the best attribute chosen. If the attribute is a nominal attribute, examples that have each value of the attribute A form a subset and if the best attribute is a numerical attribute, two subsets are formed by disjoint ranges of the attribute A.
- Step 3: Return a tree with one edge or branch for each subset.

A descendant subtree or a label value is created for every branch by applying the algorithm recursively. In general, the recursion stops if all the examples have the same label value, or if most examples are of the same label. In addition to this general stop condition, there are other stopping conditions such as:

- Number of examples in the current subtree is lower than a threshold, which is adjusted using the minimal size for split parameter.
- No attribute reaches a certain threshold, which is determined by the minimum gain parameter.
- The maximal depth is reached, which is defined by the maximal depth parameter.

Pruning is a technique to generalize an over-fitted tree and enhance its predictive power on unseen data. While pre-pruning is applied parallel to the tree creation process, post-pruning is applied after the tree creation process is complete. The DT algorithm tends to be more meaningful and easier to interpret compared to other algorithms.

3.2.4 Decision Rule

Decision Rules (DR) algorithm generates a set of rules based on the training dataset. The DR algorithm implemented in RapidMiner works in a similar fashion to the propositional rule learner named 'Repeated Incremental Pruning to Produce Error Reduction' (Cohen, 1995). It works as follows: starting with the less prevalent classes, the algorithm grows and prunes rules iteratively until there are no more positive examples or the error is greater than 50%. In the growing phase, for each rule, conditions are greedily added until 100% accuracy is reached. The procedure tries all possible values for all attributes to select the condition with the highest information gain. In the pruning phase, for each rule, any final sequence of the antecedents is pruned with the pruning metric.

3.2.5 Support Vector Machines

The Support Vector Machines (SVM) algorithm constructs a single or a set of hyperplanes in a high or infinite dimensional space, and can be used for a wide range of tasks. Intuitively, the hyperplane that has the largest distance to the nearest training data points of any class achieves a good separation. Although the original question may be stated in a finite dimensional space, the sets to be discriminated are often not linearly separable in that space. Consequently, the original space is mapped to a much higher-dimensional space through a kernel function K(x, y). Performance of the SVM algorithm can be affected by many parameters, including kernel functions, complexity constant, convergence epsilon, and factors for the SVM complexity for negative and positive examples.

3.2.6 Artificial Neural Network

The Artificial Neural Network (ANN) algorithm is a computational model that is inspired by the structure and functional aspects of biological neural networks (Zhang, 2000). A neural network consists of a group of interconnected artificial neurons. The structure of an ANN changes based on external or internal information passed to the network during the training phase. This research uses a feed-forward neural network trained by a back propagation algorithm (Rumelhart et al., 1985). This algorithm consists of two phases: propagation and weight update. With the given weights of attributes, a pre-difined error function is computed using the output values and the true values. The error is then propagated back through the network to reduce the error value. After a sufficiently large number of training cycles, the network usually converges to a point where the error is small.

IV. Empirical Setup

4.1 Data Sets

4.1.1 Feature Selection and Construction

Initial data are collected from the World Bank Enterprise Survey. The World Bank describes it as'*a firm-level survey of a representative sample of an economy's private sector. The surveys cover a broad range of business environment topics including access to finance, corruption, infrastructure, crime, competition, and performance measures.*' A detailed description on the survey methodology and data is available at http://www.enterprisesurveys.org/methodology. The data includes features for various purposes, not just for addressing corruption. The excessive number of features can cause huge computational burden on the machine learning algorithms. Therefore, it is necessary to choose the relevant features that best represent the problem domain. Based on prior research, we heuristically choose 28 attributes (questions) as independent variables (see Table 3).

Table 3. Selected Survey Questions for Independent Variables

Index	Question wording	Response type	Variables
	B. GENERAL INFORMATION		
B.1	What is this firm's current legal status? (Shareholding company with shares trade in the stock market, Shareholding company with non-traded shares or shares traded privately, Sole proprietorship, Partnership, Limited partnership)	Categorical	b1
B.2	What percentage of this firm is owned by each of the following: (Private domestic individuals, companies or organizations; Private foreign individuals, companies or organizations; Government/State; Other)	Percent	b2a,b2b
B.3	What percentage of this firm does the largest owner or owners own?	Percent	b3
B.4	Amongst the owners of the firm, are there any females?	Yes/No	b4
B.5	In what year did this establishment begin operations?	Year	b5
B.6	How many full-time employees did this establishment employ when it started operations? Please include all employees and	Number	b6

	managers		
B.6a	Was this establishment formally registered when it began operations?	Yes/No	b6a
B.6b	In what year was this establishment formally registered?	Year	b6b
B.7	How many years of experience working in this sector does the top manager have?	Number	b7
B.7a	Is the top manager female?	Yes/No	b7a
	D. SALES AND SUPPLIES		
D.2	In fiscal year [last complete fiscal year], what were this establishment's total annual sales? Please also write out the number (i.e. 50,000 as Fifty Thousand)	Number (LCUs)	d2
D.3	In fiscal year [last complete fiscal year], what percentage of this establishment's sales were: (National sales; Indirect exports; Direct exports)	Percent	d3a, d3b, d3c
	E. DEGREE OF COMPETITION		
E.1	In fiscal year [last complete fiscal year], which of the following was the main market in which this establishment sold its main product? (Local, National, International)	Categorical	e1
E.11	Does this establishment compete against unregistered or informal firms?	Yes/No	e11
	K. FINANCE		
K.3	Over fiscal year [last complete fiscal year], please estimate the proportion of this establishment's working capital that was financed from each of the following sources? (Internal funds/Retained earnings; Borrowed from banks (private and state-owned); Borrowed from non-bank financial institutions; Purchases on credit from suppliers and advances from customers; Other (moneylenders, friends, relatives, etc.))	Percent	k3a, k3bc, k3e, k3f, k3hd
K.21	In fiscal year [last complete fiscal year], did this establishment have its annual financial statements checked and certified by an external auditor?	Yes/No	k21
	J. BUSINESS-GOVERNMENT RELATIONS		
J.4	Over the last year, how many times was this establishment either inspected by tax officials or required to meet with them?	Number	j4
J.6a	Over the last year, has this establishment secured or attempted to secure a government contract?	Yes/No	j6a
J.30	As I list some factors that can affect the current operations of a business, please look at this card and tell me if you think that each factor is No Obstacle, a Minor Obstacle, a Moderate Obstacle, a Major Obstacle, or a Very Severe Obstacle to the current operations of this establishment. (Tax rates; Tax administration; Business licensing and permits; Political instability; Corruption; Courts)	Categorical	j30f
	L. LABOR		
L.1	At the end of fiscal year [last complete fiscal year], how many permanent, full-time employees did this establishment employ? Please include all employees and managers	Number	l1
L.2	Three fiscal years ago, at the end of fiscal year [three complete fiscal years ago], how many permanent, full-time employees did this establishment employ? Please include all employees and managers	Number	l2

Using a subset of questions, we construct the dependent variable (class attribute) that represents whether or not a company has committed corruption. The selected questions are summarized in Table 4. If the company answers positively for questions C.14, C.21, G.4, J.5, J.12, and J.15 or answers with greater than 0 for questions J.6 and J.7, we regard the company as being corrupt.

Table 4. Selected Survey Questions for Constructing Dependent Variable

Index	Question wording	Response type	Variables
	C. INFRASTRUCTURE AND SERVICES		
C.14	In reference to that application for a water connection, was an informal gift or payment expected or requested?	Yes/No	c14
C.21	In reference to that application for a telephone connection, was an informal gift or payment expected or requested?	Yes/No	c21
	G. LAND AND PERMITS		
G.4	In reference to that application for a construction-related permit, was an informal gift or payment expected or requested?	Yes/No	g4
	J. BUSINESS-GOVERNMENT RELATIONS		
J.5	In any of these inspections or meetings was a gift or informal payment expected or requested	Yes/No	j5
J.6	When establishments like this one do business with the government, what percent of the contract value would be typically paid in informal payments or gifts to secure the contract?	Percent	j6
J.7	It is said that establishments are sometimes required to make gifts or informal payments to public officials to"get things done"with regard to customs, taxes, licenses, regulations, services etc. On average, what percentage of total annual sales, or estimated total annual value, do establishments like this one pay in informal payments or gifts to public officials for this purpose?	Percent or Number (LCUs)	j7a or j7b
J.12	In reference to that application for an import license, was an informal gift or payment expected or requested?	Yes/No	j12
J.15	In reference to that application for an operating license, was an informal gift or payment expected or requested?	Yes/No	j15

4.1.2 Descriptive Statistics of Attributes

A total of 18,003 companies have responded to the survey. The dataset consists of companies from India (9,281 / 51.6%), Russia (4,220 / 23.4%), China (2,700 / 15.0%) and Brazil (1,802 / 10.0%). The sample businesses are distributed over 32 industries. Wholesale (1,696 / 9.4%), transport machines (1,391 / 7.7%), and machinery and equipment (1,235 / 6.9%) compose large portions of the sample, while refined petroleum product, recycling, other services, communication equipment and motor vehicles have proportions less than 0.5% of the total sample (see Table 5).

Table 5. Attribute: Industry

Description	Value	Number of Companies	Ratio	Description	Value	Number of Companies	Ratio
Wholesale	1	1,696	9.4	Wood	19	55	1.4
Machinery and equipment	2	1,235	6.9	Furniture	20	381	2.1
Plastics & rubber	3	971	5.4	Paper	21	195	1.1
Fabricated metal	4	927	5.2	Precision	22	157	0.9

products				instruments			
Retail	5	948	5.3	Recorded media	23	278	1.5
Chemicals	6	995	5.5	Auto parts	24	143	0.8
Food	7	964	5.4	Shoes and leather	25	247	1.4
Basic metals	8	810	4.5	Tobacco	26	112	0.6
Non-metallic mineral products	9	795	4.4	Post and telecommunications	27	107	0.6
Textiles	10	922	5.1	Other manufacturing	28	113	0.6
Electronics	11	855	4.8	Refined petroleum product	29	41	0.2
Services of motor vehicles	12	596	3.3	Recycling	30	28	0.2
Transport machines	13	1,391	7.7	Other services	31	28	0.2
Construction: Section F	15	795	4.4	Communication equipment	32	8	0.0
Garments	16	554	3.1	Motor vehicles	48	6	0.0
Hotel and restaurants	17	663	3.7	N/A	?	160	0.9
IT	18	612	3.4				

Table 6. Attributes

Attributes		Value	Attribute Values	Number of Companies	Ratio (%)
General Information	b1	1	Sole proprietorship	5,841	32.4
		2	Shareholding company with non-traded shares	4,676	26.0
		3	Limited partnership	5,684	31.6
		4	Shareholding company with shares traded	379	2.1
		5	Shareholding company with shares trade in the stock market	303	1.7
		6	Other	345	1.9
		?	N/A	775	4.3
	b2a	numeric	Min 0, Max 100, Average 96.4, Deviation 17.2		
	b2b	numeric	Min 0, Max 100, Average 1.6, Deviation 11.5		
	b3	numeric	Min 0, Max 100, Average 74.8, Deviation 27.0		
	b4	0	No	12,234	68.0
		1	Yes	4,870	27.1
		?	N/A	899	5.0
	b6	numeric	Min 1, Max 9010, Average 39.0, Deviation 173.3		
	b6a	0	No	896	5.0
		1	Yes	16,829	93.5
		?	N/A	278	1.5
	b7	numeric	Min 1, Max 70, Average 14.9, Deviation 9.6		
	b7a	0	No	15,813	87.8
		1	Yes	2,139	11.9
		?	N/A	51	0.3
Sales & Supplies	d2	numeric	Min 1, Max 90000000000, Average 269001453.8 Deviation 2007896795.3		
	d3a	numeric	Min 1, Max 100, Average 92.5 Deviation 22.2		

	d3b	numeric	Min 1, Max 100, Average 1.5 Deviation 9.2		
	d3c	numeric	Min 1, Max 100, Average 4.9 Deviation 17.8		
Degree of Competition	e1	1	Local	5,700	31.7
		2	National	7,706	42.8
		3	International	803	4.5
		?	N/A	3,794	21.1
	e11	0	No	9,842	54.7
		1	Yes	7,376	41.0
		?	N/A	785	4.4
Finance	k3a	numeric	Min 1, Max 100, Average 71.0 Deviation 34.1		
	k3bc	numeric	Min 1, Max 100, Average 19.3 Deviation 29.0		
	k3e	numeric	Min 1, Max 100, Average 9.9 Deviation 385.9		
	k3f	numeric	Min 0, Max 50000, Average 1.2 Deviation 8.1		
	k3hd	numeric	Min 0, Max 60000000, Average 15248.1 Deviation 741570.9		
Business-Government Relationship	k21	0	NO	6,756	37.5
		1	YES	10,736	59.6
		2	ANY NUMBER	174	1.0
		?	N/A	337	1.9
	j4	numeric	Min 0, Max 1, Average 0.0 Deviation 0.1		
	j6a	0	NO	14,323	79.6
		1	YES	3,134	17.4
		2	ANY NUMBER	37	0.2
		?	N/A	509	2.8
	j30f	0	NO	17,491	97.2
		1	YES	112	0.6
		?	N/A	400	2.2
Labor	l1	numeric	Min -9, Max 30000, Average 124.0 Deviation 558.2		
	l2	numeric	Min 0, Max 355652, Average 140.9 Deviation 3007.6		
Class		0	No	11,939	66.3
		1	Yes	6,064	33.7

The companies also have diverse legal status, including sole proprietorship (5,841 / 32.4%), shareholding company with non-traded shares (4,676 / 26.0%), limited partnership (5,684 /31.6%), shareholding company with shares traded (379 / 2.1%), shareholding company with shares traded in the stock market (303/1.7%), other (345 / 1.9%), and missing attribute (775 / 4.3%) (See b1 in Table 6). Owners tend to be male (68.0%) (See b4 in Table 6). Most companies have been formally registered when they were established (93.5%) (See b6a in Table 6). Average sales of the companies are about 269 million US dollars with most sales made domestically (92.5%) (See d2, d3a, d3b and d3c in Table 6). The companies sell their products in Local (5,700 / 31.7%), National (7,706 / 42.8%), and International (803, 4.5%) markets (See e1 in Table 6). The companies compete with both unregistered (54.7%) and registered companies (41.0%). On average, 71.0% of working capital has been financed from internal funds/retained earnings, 19.3% has been borrowed from banks, 9.9% has been borrowed from non-bank financial institutions, and 1.2% has been purchased on credit/advances

from suppliers and customers (See k3a, k3bc, k3e and k3f in Table 6). 59.6% of companies have had their financial statements checked and certified by external auditors in the last fiscal year whereas 37.5% have not been audited (See k21 in Table 6). 79.6% of companies have not been inspected by the tax officials over the last 12 months, while 17.4% have been inspected (See k21 in Table 6). Only 17.4% of companies have secured or have attempted to secure a government contract over the last year, while 79.6% have not (See j6a in Table 6). Finally, the proportion of corrupted and non-corrupted firms is 66.3% and 33.7%, respectively (See class in Table 6).

4.2 Data Pre-processing

Several issues become apparent if the raw data is used as-is. First, SVM and ANN cannot handle nominal values when learning their models, so it is necessary to convert nominal attribute values into numeric attribute values. Dummy coding is used for this purpose. This approach creates a new attribute for all values of the nominal attribute, excluding the comparison group. Except for missing values, the new attribute which corresponds to the actual nominal value of that example, becomes value 1 and all other new attributes become value 0. If the nominal value corresponds to the comparison group, the new attributes are all set to 0. The comparison group is an optional parameter for'dummy coding'. When there is no comparison group, the new attribute corresponding to the nominal value of the example gets value 1 and others get value 0. In this case, there will be no example with all new attributes set to 0. Second, it is necessary to handle missing values. Some algorithms such as DT and DR can handle missing values, while others such as SVM, and ANN, cannot. In order to solve this problem, we replace the missing values by the average values of each attribute. For the categorical value, most frequent value is used for replacing the missing value. Note that this transformation has been applied after converting all attributes into numeric values. Third, monetary values are not consistent among countries. Therefore, we convert monetary values into US Dollar using annual average foreign exchange rates.

4.3 Technique Setup

4.3.1 Evaluation Framework

We compare various models created by machine learning algorithms. For this purpose, the dataset is divided into two subsets - the training and the testing datasets. While the training dataset is used to generate the predictive models with various algorithms, the testing dataset is used to check the performance of the models. For measuring performance, each example of the testing dataset is processed by the model and then the predicted class label is compared with the actual label. In order to ensure the fairness of the evaluation, we employ x-fold cross validation. In this approach, the dataset is divided into x subsets. For each evaluation, one of the x subsets is used as the testing dataset and the remaining x-1 subsets are used as a training dataset. The average performance of all x trials is computed. Although this method requires the algorithms to be rerun x times, it reduces the dependency on data division since every example is included in a test dataset once, and is included in a training dataset x-1 times. The variance of the evaluation results is reduced as x is increased. We used RapidMiner for evaluating various algorithms in x-fold cross validation.

4.3.2 Parameter Setting

We used default parameter configuration provided by RapidMiner Operator provides for each modelling algorithm (see Table 7). However, for several operators, we make some special considerations as follows: k-NN requires to setup for k (number of nearest neighbour) since k significantly impacts the performance. We test different values of k (e.g., 3, 5, 7, 9 and 11) and compare their performance. The best attribute selection criteria, such as Gain Ratio, Information Gain, and Gini Index, are important in DT and thus we compare DT performance across different best attribute selection criteria.

Table 7. Default Parameter Settings

Algorithms	Parameter settings
K-NN	similarity measure = Mixed Euclidean Distance
NB	use laplace correction
DR	criterion=information gain, sample ratio = 0.9, pereness = 0.9, minimal prune benefit=0.25
DT	Maximal depth=10, confidence=0.1(for pruning), minimal gain=0.01, minimal leaf size=2
SVM	kernel type=dot, C=0.0, conversionce epsilon=0.001, L pos=1.0, L neg=1.0, eplsilon=0.0, epsilon plus=0.0, epsilon minus=0.0
ANN	hidden layer =1, hidden layer size=2, learning rate=0.01, momentum=0.9, error epsilon=1.0E-4

4.3.3 Confusion Matrix

A confusion matrix contains information about the actual versus the predicted classifications from a classification model. Classification performance is often evaluated using the data from the matrix. Table 8 shows the confusion matrix for a binary classification problem.

Table 8. Confusion Matrix

	True	False
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

The entries in the confusion matrix have the following meaning:
- True Positive (TP) is the number of correct predictions when an example is positive;
- False Positive (FP) is the number of incorrect predictions when an example is positive;
- False Negative (FN) is the number of incorrect of predictions when an example negative; and
- True Negative (TN) is the number of correct predictions when an example is negative.

4.3.4 Performance Metrics

Common metrics for the two-class matrix is defined as follows:
The accuracy is the proportion of the total number of predictions that are correct. It is defined

as:

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

The recall or true positive rate is the proportion of positive cases that are correctly identified. It is defined as

$$recall = \frac{TP}{TP+FN} \quad (2)$$

The false positive rate is the proportion of negatives cases that are incorrectly classified as positive. It is defined as

$$false\ positve\ rate = \frac{FP}{FP+TN} \quad (3)$$

The true negative rate is defined as the proportion of negatives cases that are classified correctly. It is defined as

$$true\ negativerate = \frac{TN}{FP+TN} \quad (4)$$

The false negative rate is the proportion of positive cases that are incorrectly classified as negative. It is defined as

$$falsenegativerate = \frac{FN}{TP+FN} \quad (5)$$

Finally, precision is the proportion of the predicted positive cases that are correct. It is defined as

$$precision = \frac{TP}{TP+FP} \quad (6)$$

If the number of negative cases is much greater than the number of positive cases, using accuracy alone can be misleading (Provost et al., 1998). Suppose there are 1000 cases, 995 of which are negative cases and five of which are positive cases. Now the classification model classifies them all as negative. Although the model missed all positive cases, the accuracy would be 99.5%. Geometric mean (g-mean) (Kubat et al., 1998) of recall and precision as defined in equations (7) and f-Measure (Lewis and Gale, 1994) as defined in equation (8) have been developed to overcome this limitation. In equation (8), β can range from 0 to infinity, which is used to control the weight assigned to recall and precision. β is usually set to 1, in which case the f-measure can be defined as (9). If all positive cases are incorrectly classified by any model, equations (7), (8) and (9) will all have a value of 0.

$$g-mean = \sqrt{recall \times precision} \quad (7)$$

$$f-measure = \frac{(\beta^2+1)\times precision \times recall}{\beta^2 \times precision + recall} \quad (8)$$

$$f-measure = \frac{2\times precision \times recall}{precision + recall} \quad (9)$$

V. Results

5.1 Performances

Table 9 summarizes the prediction performance metrics defined above. K-NNs are measured with different k while DTs are measured with different node selection criteria (Gain Ratio, Information Gain, and Gini Index).

Table 9. Predictive Performances

	TP	FP	FN	TN	accu-racy (%)	recall	false positive rate(%)	true negative rate(%)	Precision (%)	f-measure (%)
K-NN (k=3)	5029	4962	1035	6977	66.7	82.9	41.6	58.4	50.3	62.6
K-NN (k=5)	4996	4855	1068	7084	67.1	82.4	40.7	59.3	50.7	62.8
K-NN (k=7)	5010	4835	1054	7104	67.3	82.6	40.5	59.5	50.9	63.0
K-NN (k=9)	5019	4857	1045	7082	67.2	82.8	40.7	59.3	50.8	63.0
K-NN (k=11)	5036	4954	1028	6985	66.8	83.0	41.5	58.5	50.4	62.7
NB	4539	1071	1525	10868	85.6	74.9	9.0	91.0	80.9	77.8
DR	4698	131	1366	11808	91.7	77.5	1.1	98.9	97.3	86.3
DT (GR)	4638	20	1426	11919	92.0	76.5	0.2	99.8	99.6	86.5
DT (IG)	4546	8	1518	11931	91.5	75.0	0.1	99.9	99.8	85.6
DT (GI)	4885	384	1179	11555	91.3	80.6	3.2	96.8	92.7	86.2
SVM	4443	184	1621	11755	90.0	73.3	1.5	98.5	96.0	83.1
ANN	4742	477	1322	11462	90.0	78.2	4.0	96.0	90.9	84.1
Ave.					80.0	79.4	19.6	80.4	74.4	74.7

Average performances of the learning algorithms are summarized in the bottom line - accuracy is 80.0%, recall is 79.4%, false positive rate is 19.6%, true negative rate is 80.4%, precision is 74.4% and f-measure is 74.7%. Accuracy ranges from 66.8% (K-NN with k=11) to 92.0% (Decision Tree with Gain Ratio). In accuracy, K-NNs shows low performances (66.7% ~ 67.3%), while DR, DT, SVM and ANN display high performances (over 90%), and NB show moderate performance (85.6%). In recall, k-NNs and DR (over 80%) exhibit better performances compared to other approaches. DR and DTs demonstrate extremely low false positive rate and high true negative rate compared to other models. DR and DTs also display high precision compared to k-NNs. Finally, DR, DT, SVM and ANN reveal high performances in f-measure. Based on these performance results, we may choose different machine learning approaches for different objectives. For example, if the aim of the predictive analysis is to find as many potentially corruptive companies as possible, it is better to use k-NNs; inversely, if the aim is to predict the precise company that is corrupt, it is better to use DTs with Gain Ratio or with Information Gain.

5.2 Models

5.2.1 Decision Rules

As illustrated in Figure 1, DR generates a set of rules which are applied sequentially from top to bottom. Each rule is described by if ⟨condition⟩ then ⟨conclusion⟩ format. If a rule applies to a given example, the model stops and returns the predicted label; otherwise, the next rule is examined. If any rule is not satisfied, the final label is applied to the example. Significant rules are as follows: Rule 1 (if b6b 〉 1991.500 and j6a = 0 then 0) implies that"*if a company was established after 1992 and has not secured or has not attempted to secure government contract in the last 12 months, it does not commit corruption.*"A total of 11,153 examples satisfy this condition, and 879 (7.9%) out of them have committed corruption. Rule 2 (if b6b ≤ 1991.500 then 1) means that"*if a company was established before 1992, it commits corruption*". This means that if a company has existed for a long time, it probably has committed corruption in some way. This rule displays very high accuracy - a total of 4,175 examples satisfy this condition, and 4,149 (99.8%) out of them have committed corruption. Rule 7 (if b6b 〉 2011.500 then 1) means that"*if a company was established after June 2011, it commits corruption*". Finally, if a company does not satisfy any of above rules, the final rule (else 1) is applied to the example, meaning the company has committed corruption.

Fig. 1. Model of Decision Rules

```
[rule 1] if b6b > 1991.500 and j6a = 0 then 0  (879 / 10274)
[rule 2] if b6b ≤ 1991.500 then 1  (4149 / 8)
[rule 3] if b6b ≤ 2011.500 and b5 > 1999.500 and k3bc ≤ 37.500 and b7 ≤ 10.500 and l1 ≤ 99
and d2 ≤ 34250000 then 0  (21 / 196)
[rule 4] if b6b ≤ 2011.500 and b6b > 1995.500 and j6a =? then 0  (17 / 208)
[rule 5] if b6b ≤ 2011.500 and a1 = 1 and b5 > 2001.500 then 0  (52 / 262)
[rule 6] if b6b ≤ 2011.500 and d2 > 20200000 and k3bc ≤ 45 and l1 > 138 and b6 > 11 then 0  (3 / 53)
[rule 7] if b6b > 2011.500 then 1  (31 / 1)
[rule 8] if j6a = 1 and b7 ≤ 25.500 then 0  (449 / 791)
[rule 9] else 1  (416 / 131)
```

5.2.2 Decision Trees

As illustrated in Figure 1, DT generates a set of rules, which are constructed as a tree-like structure. Decisions are made by following the satisfied nodes. For example, assuming that an example has the following values: b6b = 2012, a7=0, d3a=78 and k3bc =30. In this case, the example satisfies decision nodes b6b 〉 1991.500, b6b ≤ 2012.500, b6b 〉 2011.500, a7 = 0, d3a 〉 73.500 and k3bc 〉 25 and therefore has predicted label of 1 (committed corruption). Three different decision trees that utilise Gain Ratio, Information Grain, and Gini Index as split criteria are illustrated in Figure 2. All decision trees choose b6b (the year of establishment) as a root node, but have different subtrees. If b6b = ? (Missing value) and if b6b ≤ 1991.500, they also make the'corruption'decision (class=1). However, they produce different subtrees if b6b 〉1991.5. If the Gain Ratio is used for selecting decision node attributes, detailed sub-decision nodes are created by b6b (the year of establishment), a7 (is a part of large company?), d3a

(% of national sales), k3bc (% of working capital borrowed from banks), d2 (total sales in last year), l1 (number of permanent, full-time employees at end of last fiscal year), and b6 (number of full-time employees of the establishment when the firm started operations) (See Figure 2 (a)). If Information Gain is used for selecting decision node attribute,'b6b 〉1991.5'node is further divided into a4a (Industry sampling sector), d3a (% of national sales) and b7 (top manager experience by years) (See Figure 2 (b)). Finally if Gini Index is used in selecting the decision node attribute, there is no further division of'b6b 〉1991.5'node (See Figure 2 (c)).

Fig. 2. Model of Decision Tree

```
b6b = ?: 1 [1=399, 0=0]
b6b > 1991.500
|   b6b > 2012.500: 1 [1=45, 0=0]
|   b6b ≤ 2012.500
|   |   b6b > 2011.500
|   |   |   a7 = 0
|   |   |   |   d3a > 73.500
|   |   |   |   |   k3bc > 25: 0 [1=3,
0=8]
|   |   |   |   |   k3bc ≤ 25
|   |   |   |   |   |   b4 = 0
|   |   |   |   |   |   |   d2 = ?: 1
[1=5, 0=0]
|   |   |   |   |   |   |   d2 >
35000000: 1 [1=11, 0=0]
|   |   |   |   |   |   |   d2 ≤
35000000
|   |   |   |   |   |   |   |   l1 >
52.500: 0 [1=1, 0=5]
|   |   |   |   |   |   |   |   l1 ≤
52.500
|   |   |   |   |   |   |   |   |   d2 >
2000000: 1 [1=8, 0=0]
|   |   |   |   |   |   |   |   |   d2 ≤
2000000
|   |   |   |   |   |   |   |   |   |   b6
> 3.500: 1 [1=3, 0=1]
|   |   |   |   |   |   |   |   |   |   b6
≤ 3.500: 0 [1=0, 0=2]
|   |   |   |   |   |   b4 = 1: 0 [1=1,
0=2]
|   |   |   |   d3a ≤ 73.500: 0 [1=0,
0=4]
|   |   |   a7 = 1: 1 [1=29, 0=0]
|   |   b6b ≤ 2011.500: 0 [1=1410,
0=11909]
b6b ≤ 1991.500: 1 [1=4149, 0=8]
```

Criterion = Gain Ratio

```
b6b = ?: 1 [1=399, 0=0]
b6b > 1991.500: 0 [1=1516,
0=11931]
b6b ≤ 1991.500
|   a4a = 1: 1 [1=130, 0=0]
|   a4a = 10: 1 [1=295, 0=0]
|   a4a = 11: 1 [1=225, 0=0]
|   a4a = 12
|   |   d3a > 85: 1 [1=115, 0=1]
|   |   d3a ≤ 85
|   |   |   b7 > 16.500: 0 [1=1,
0=2]
|   |   |   b7 ≤ 16.500: 1 [1=2,
0=0]
|   a4a = 13: 1 [1=276, 0=0]
|   a4a = 15: 1 [1=69, 0=0]
|   a4a = 16: 1 [1=152, 0=0]
|   a4a = 17: 1 [1=130, 0=0]
|   a4a = 18: 1 [1=118, 0=0]
|   a4a = 19: 1 [1=67, 0=0]
|   a4a = 2: 1 [1=352, 0=1]
|   a4a = 20: 1 [1=123, 0=0]
|   a4a = 21: 1 [1=63, 0=0]
|   a4a = 22: 1 [1=15, 0=0]
|   a4a = 23: 1 [1=53, 0=0]
|   a4a = 24: 1 [1=55, 0=0]
|   a4a = 25: 1 [1=93, 0=0]
|   a4a = 26: 1 [1=59, 0=0]
|   a4a = 27: 1 [1=9, 0=0]
|   a4a = 28: 1 [1=54, 0=2]
|   a4a = 29: 1 [1=19, 0=0]
|   a4a = 3: 1 [1=279, 0=1]
|   a4a = 30: 1 [1=3, 0=0]
|   a4a = 31: 1 [1=9, 0=0]
|   a4a = 4: 1 [1=221, 0=0]
|   a4a = 5: 1 [1=132, 0=0]
|   a4a = 6: 1 [1=275, 0=0]
|   a4a = 7: 1 [1=242, 0=1]
|   a4a = 8: 1 [1=228, 0=0]
|   a4a = 9: 1 [1=211, 0=0]
|   a4a = ?: 1 [1=70, 0=0]
|   a4a = ?: 1 [1=4, 0=0]
```

Criterion = Information Gain

```
b6b = ?: 1 [1=399, 0=0]
b6b > 1991.500: 0 [1=1516,
0=11931]
b6b ≤ 1991.500: 1 [1=4149, 0=8
```

Criterion = Gini Index

Decision trees show that the company is likely to be corrupt in the following situations: If the year of establishment is unknown (b6b = ? [1 = 399, 0 = 0]) (all decision trees); and if the company was established before 1991.5 (b6b ≤ 1991.500: 1 [1 = 4149, 0 = 8])(Gain Ratio, Gini Index). Note that when Information Gain is used for selecting decision node,'b6b

〉1991.5' node is further divided, but it only identifies small portion of non-corrupted companies by considering d3a and b7. This may be reduced if we use more conservative pruning parameters. Decision trees also show that a company is unlikely to be corrupt if it was established after 1991.50 (b6b 〉 1991.500). This decision node contains a large number of non-corrupt companies (class = 0, 11931), but also contains a substantial number of corrupt companies (class = 1, 1516). This implies that satisfying this decision rule may produce false predictions. Again, this problem may be overcome by considering more conservative pruning, such as limiting the depth of the decision tree or increasing the minimal leaf size.

5.2.3 Support Vector Machine Model

SVM learns the weights for attributes and the intercept, and uses them to predict a class label for a given example. Table 10 summarizes the top ten positive and negative weights. The following attributes are positively related to corruption: the year in which the company was established (b2b, 1.48), unknown major market (e1 = ?, 0.37), commencing year (d5, 0.19), external audit conducted in the last year (k21 = 1, 0.16), and China (a1=3, 0.08). The following attributes are negatively related to corruption: secured government contract (j6a = 1, 0.28), major market is local (e1 = 1, -0.27), no external audit conducted in the last year (k21 = 0, -0.25), obstacles is 1 (j30f = 1, -0.22) and secured government contract is unknown (j6a = 1, -0.17).

Table 10. Weights for Attributes of SVM

	Low Weight Attributes		High Weight Attributes	
	attribute	weight	attribute	weight
1	j6a = 1	-0.28	b6b	1.48
2	e1 = 1	-0.27	e1 = ?	0.37
3	k21 = 0	-0.25	b5	0.19
4	j30f = 1	-0.22	k21 = 1	0.16
5	b6a = ?	-0.17	a1 = 3	0.08
6	j30f = ?	-0.15	j6a = 0	0.07
7	j6a = 2	-0.14	j3 = 0	0.06
8	a7 = ?	-0.12	e1 = 3	0.06
9	e11 = ?	-0.07	k3hd	0.04
10	k3a	-0.07	e11 = 1	0.04
11	j3 = 1	-0.05	a7 = 1	0.03
12	d3b	-0.04	b7a = 1	0.03
13	k21 = ?	-0.04	a4a = 4	0.02
14	d3c	-0.03	d2	0.02
15	k3f	-0.03	b1 = 1	0.02

5.2.4 Others

K-NN does not provide a model for classification. Instead, when it classifies an example the model finds most k similar examples from the training dataset and predicts the class label using voting mechanism. Naïve Bayes also does not create a model for classification. Instead, it predicts the class label of a given example using class distribution information (probability) of each attribute value. The model generated by ANN consists of attribute weights between

input nodes and hidden nodes, and between hidden nodes and output nodes. It is difficult to interpret these weights by themselves, and may require other techniques for interpretation.

VI. Discussion

6.1 Corruption Prediction without Firm Age

The most significant result from this research is that the firm's age reveals a strong relationship with corruption. This means that older companies demonstrate a greater potential to commit corruption compared to relatively younger companies. As firms can be relieved from heavy regulations and complicated bureaucracy with corruptive behaviors (Rose-Akerman, 1999), which are main features of emerging markets, older firms are more likely to engage in corruption than young ones. This notion is supported by several models, such as decision rules, decision trees, and logistic regression. This outcome is also aligned with common sense in that the longer a company lasts, the higher the probability that the company will commit at least one of the corruption factors described in Table 4, either deliberately or by chance. Obviously, other attributes may contribute to possible corruption as identified by the models. Therefore, if the attributes related to firm age (e.g., b5 and b2b) are removed, it may give further insight on which factors determine corruption.

6.1.1 Performances

When we remove age related attributes, the performances of all models decrease significantly for all performance measures. The decrease seems to roughly scale with the original (accuracy decreasing by 10% of 70% = 63%; of 90% = 81% etc.). SVM exhibits optimum accuracy (accuracy = 81.8% and f-measure = 70.6%) both in accuracy and f-measure.

Table 11. Predictive Performances without Firm Age

	TP	FP	FN	TN	accu-racy (%)	recall	false positive rate(%)	true negative rate(%)	Precision (%)	f-measure (%)
K-NN(k=3)	5073	5743	991	6196	62.6	83.7	48.1	51.9	46.9	60.1
K-NN(k=5)	5078	5746	986	6193	62.6	83.7	48.1	51.9	46.9	60.1
K-NN(k=7)	5068	5816	996	6123	62.2	83.6	48.7	51.3	46.6	59.8
K-NN(k=9)	5086	5856	978	6083	62.0	83.9	49.0	51.0	46.5	59.8
K-NN (k=11)	5088	5994	976	5945	61.3	83.9	50.2	49.8	45.9	59.3
NB	3601	2070	2463	9869	74.8	59.4	17.3	82.7	63.5	61.4
DR	3371	856	2693	11083	80.3	55.6	7.2	92.8	79.7	65.5
DT (GR)	1361	118	4703	11821	73.2	22.4	1.0	99.0	92.0	36.1
DT (IG)	3161	473	2903	11466	81.2	52.1	4.0	96.0	87.0	65.2
DT (GI)	3139	481	2925	11458	81.1	51.8	4.0	96.0	86.7	64.8
SVM	3938	1146	2126	10793	81.8	64.9	9.6	90.4	77.5	70.6
ANN	3860	1334	2204	10605	80.3	63.7	11.2	88.8	74.3	68.6
Ave.					71.8	67.2	25.9	74.1	65.1	61.3

6.1.2 Models

The DR model from after removing firm age attributes is illustrated in Figure 3. The number of rules significantly increases when the business age attributes are removed, and more attributes are used to identify rule condition. Most frequently used attributes are external audit (k21), secured government contract (j6a), top manager's experience (b7), major market (e1), and percentage of working capital financed from internal funds/retained earnings (k3a).

- The rules that display which company is more likely to commit corruption include:
- The company's major market is local and its top manager's experience is greater than 18.5 years ([rule 2]);
- The company's major market is local and the company's financial statements have not been checked and certified by an external auditor in last financial year ([rule 5]) and
- The company's total annual sale is less than 1350 USD ([rule 13] if d2 ≤ 1350 then 1 (136 / 0)).

Fig. 3. Model of Decision Rules without Firm Age

[rule 1] if k21 = 1 and j6a = 0 then 0 (1353 / 7270)
[rule 2] if e1 = 1 and b7 > 18.500 then 1 (1350 / 167)
[rule 3] if e11 = 1 and j6a = 0 then 0 (522 / 1531)
[rule 4] if a1 = 2 and e1 = ? then 0 (52 / 266)
[rule 5] if e1 = 1 and k21 = 0 then 1 (903 / 252)
[rule 6] if b7 ≤ 17.500 and l1 > 31.500 and j6a = 0 and d2 ≤ 45500000 then 0 (30 / 214)
[rule 7] if b7 ≤ 16.500 and j30f = 0 and e1 = ? then 0 (67 / 355)
[rule 8] if b7 ≤ 19.500 and a1 = 2 and l1 > 40.500 then 0 (60 / 213)
[rule 9] if k3a > 98.500 and b7 > 20.500 and k21 = 0 and a1 = 4 then 1 (75 / 7)
[rule 10] if k3a ≤ 84.500 and j6a = 0 and b7 ≤ 20.500 then 0 (95 / 308)
[rule 11] if k21 = 0 and j3 = 1 then 1 (373 / 121)
[rule 12] if d2 > 4250000 and j6a = ? and l2 ≤ 1743.500 then 0 (10 / 143)
[rule 13] if d2 ≤ 1350 then 1 (136 / 0)
[rule 14] if e11 = 1 and l1 > 102 then 0 (26 / 86)
[rule 15] if b3 ≤ 52.500 and b7 > 20.500 and k21 = 0 and a1 = 1 then 1 (52 / 7)
[rule 16] if d2 ≤ 2305000000 and a1 = 3 and j6a = 0 then 0 (23 / 109)
[rule 17] if k3a > 99 and l1 ≤ 21.500 and d2 > 156200000 and e1 = 2 then 1 (6 / 0)
[rule 18] if d2 > 2305000000 and b7 > 38.500 then 1 (3 / 0)
[rule 19] if k3a > 98.500 and b7 > 11.500 and l1 ≤ 30.500 and a7 = 0 and e1 = 1 then 1 (34 / 4)
[rule 20] if e1 = 2 and b7 > 15.500 then 1 (285 / 187)
[rule 21] if k21 = 1 and a1 = 1 then 0 (189 / 298)
[rule 22] if d2 > 2305000000 and b7 > 18.500 then 1 (1 / 0)
[rule 23] if b6a = ? then 1 (38 / 0)
[rule 24] if b6 ≤ 17.500 and k3hd > 7.500 and b7 ≤ 16 then 0 (1 / 12)
[rule 25] if b7 > 18.500 and l1 > 11 and b6 ≤ 45 then 0 (6 / 29)
[rule 26] if b6 > 29 and l1 ≤ 15.500 and k3a > 15 and b7 ≤ 19 then 1 (32 / 4)
[rule 27] if d2 ≤ 950000000 and d2 > 113500000 and l1 > 209 and b7 ≤ 9 then 0 (1 / 12)
[rule 28] if d2 ≤ 635000000 and a7 = 1 and k21 = 1 then 0 (4 / 29)
[rule 29] if j6a = 1 and b3 ≤ 99.500 then 1 (104 / 69)
[rule 30] if d2 ≤ 11500000 and e1 = 1 and b7 > 11.500 then 0 (0 / 8)
[rule 31] if l1 > 17.500 and j6a = 0 and b7 ≤ 9.500 then 0 (9 / 25)
[rule 32] if k3a > 77.500 and l1 ≤ 22.500 and d2 > 12950000 and b7 ≤ 18.500 and d2 ≤ 50000000 then 1 (20 / 3)
[rule 33] if k3a ≤ 76.500 and d2 ≤ 12300000 and k3a ≤ 32.500 and k3bc ≤ 72.500 then 0 (3 / 21)
[rule 34] if l2 > 83.500 and d2 ≤ 282460000 then 1 (17 / 2)
[rule 35] if l1 > 13.500 and l2 ≤ 19.500 and b1 = 1 and l1 > 17.500 then 0 (2 / 12)
[rule 36] if l1 ≤ 17.500 and d2 > 22350000 and b4 = 0 then 1 (11 / 4)
[rule 37] if l1 ≤ 10.500 and l2 > 9.500 and j6a = 1 then 1 (7 / 0)
[rule 38] if b6 ≤ 36.500 and k3f ≤ 2.500 and k21 = 1 and k3a > 47.500 then 0 (18 / 46)

```
[rule 39] if b7 ≤ 19 and l2 > 24.500 and k3a ≤ 82.500 and b7 > 7.500 and b4 = 0 then 1  (14 / 3)
[rule 40] if k3a > 76.500 and l1 ≤ 29 and d2 > 6450000 and d2 ≤ 12350000 and b1 = 1 then 1  (7 / 0)
[rule 41] if k3a > 76.500 and l1 ≤ 29 and j6a = 1 and b6 ≤ 27.500 then 1  (13 / 0)
[rule 42] if b4 = 0 and b1 = 2 then 0  (14 / 34)
[rule 43] if b6 > 36.500 and k3bc ≤ 12.500 and b6 > 105 then 1  (8 / 0)
[rule 44] else 1  (81 / 79)
```

Fig. 4. Model of Decision Tree without Firm Age

```
d2 = ?
|   e1 = 1: 1 [1=909, 0=60]
|   e1 = 2
|   |   k21 = 0
|   |   |   k3bc = ?
|   |   |   |   e11 = 0: 1 [1=3, 0=0]
|   |   |   |   e11 = 1: 0 [1=0, 0=2]
|   |   |   k3bc > 75
|   |   |   |   l1 > 5.500: 0 [1=3, 0=10]
|   |   |   |   l1 ≤ 5.500: 1 [1=2, 0=0]
|   |   |   k3bc ≤ 75
|   |   |   |   k3e > 60: 0 [1=0, 0=2]
|   |   |   |   k3e ≤ 60: 1 [1=204, 0=34]
|   |   k21 = 1
|   |   |   l1 > 240: 0 [1=0, 0=16]
|   |   |   l1 ≤ 240
|   |   |   |   j6a = 0: 0 [1=44, 0=87]
|   |   |   |   j6a = 1: 1 [1=20, 0=4]
|   |   |   |   j6a = ?: 0 [1=0, 0=6]
|   |   k21 = 2: 0 [1=0, 0=4]
|   |   k21 = ?: 1 [1=8, 0=1]
|   e1 = 3
|   |   l2 = ?: 1 [1=3, 0=0]
|   |   l2 > 5.500
|   |   |   k21 = 0
|   |   |   |   j3 = 0: 0 [1=2, 0=3]
|   |   |   |   j3 = 1: 1 [1=2, 0=0]
|   |   |   k21 = 1: 0 [1=0, 0=22]
|   |   l2 ≤ 5.500: 1 [1=2, 0=0]
|   e1 = ?
|   |   b2a = ?: 0 [1=0, 0=6]
|   |   b2a > 25.500: 0 [1=18, 0=157]
|   |   b2a ≤ 25.500: 1 [1=3, 0=1]
d2 > 550: 0 [1=4632, 0=11523]
d2 ≤ 550: 1 [1=209, 0=1]
```

(a) Criterion = Gain Ratio

```
e1 = 1
|   j4 = ?
|   |   a1 = 1
|   |   |   j3 = 0
|   |   |   |   d2 = ?: 1 [1=308, 0=15]
|   |   |   |   d2 > 485000: 1 [1=736, 0=293]
|   |   |   |   d2 ≤ 485000: 0 [1=4, 0=32]
|   |   |   j3 = 1: 1 [1=983, 0=14]
|   |   |   j3 = ?: 1 [1=27, 0=7]
|   |   a1 = 2
|   |   |   b7 = ?: 1 [1=25, 0=12]
|   |   |   b7 > 17.500: 1 [1=154, 0=48]
|   |   |   b7 ≤ 17.500
|   |   |   |   j3 = 0: 0 [1=55, 0=169]
|   |   |   |   j3 = 1: 1 [1=32, 0=2]
|   |   |   |   j3 = ?: 0 [1=1, 0=2]
|   |   a1 = 3
|   |   |   j3 = 0: 0 [1=85, 0=238]
|   |   |   j3 = 1: 1 [1=62, 0=5]
|   |   |   j3 = ?: 1 [1=3, 0=1]
|   |   a1 = 4: 1 [1=637, 0=64]
|   j4 > 0.500: 0 [1=0, 0=7]
|   j4 ≤ 0.500: 0 [1=255, 0=1424]
e1 = 2: 0 [1=2034, 0=5672]
e1 = 3: 0 [1=120, 0=683]
e1 = ?
|   d2 = ?
|   |   d3a = ?: 1 [1=6, 0=0]
|   |   d3a > 95: 0 [1=15, 0=153]
|   |   d3a ≤ 95: 0 [1=0, 0=11]
|   d2 > 550: 0 [1=316, 0=3086]
|   d2 ≤ 550: 1 [1=206, 0=1]
```

(b) Criterion = Information Gain

```
e1 = 1
|   j4 = ?
|   |   a1 = 1
|   |   |   j3 = 0
|   |   |   |   k3bc = ?: 1 [1=25, 0=5]
|   |   |   |   k3bc > 15.500: 0 [1=161, 0=166]
|   |   |   |   k3bc ≤ 15.500: 1 [1=862, 0=169]
|   |   |   j3 = 1: 1 [1=983, 0=14]
|   |   |   j3 = ?
|   |   |   |   d2 = ?: 1 [1=8, 0=0]
|   |   |   |   d2 > 1750000: 1 [1=19, 0=5]
|   |   |   |   d2 ≤ 1750000: 0 [1=0, 0=2]
|   |   a1 = 2
|   |   |   b7 = ?: 1 [1=25, 0=12]
|   |   |   b7 > 17.500: 1 [1=154, 0=48]
|   |   |   b7 ≤ 17.500
|   |   |   |   j3 = 0: 0 [1=55, 0=169]
|   |   |   |   j3 = 1: 1 [1=32, 0=2]
|   |   |   |   j3 = ?: 0 [1=1, 0=2]
|   |   a1 = 3
|   |   |   j3 = 0: 0 [1=85, 0=238]
|   |   |   j3 = 1
|   |   |   |   k3bc = ?: 1 [1=2, 0=0]
|   |   |   |   k3bc > 72.500: 0 [1=0, 0=2]
|   |   |   |   k3bc ≤ 72.500: 1 [1=60, 0=3]
|   |   |   j3 = ?: 1 [1=3, 0=1]
|   |   a1 = 4: 1 [1=637, 0=64]
|   j4 > 0.500: 0 [1=0, 0=7]
|   j4 ≤ 0.500: 0 [1=255, 0=1424]
e1 = 2: 0 [1=2034, 0=5672]
e1 = 3: 0 [1=120, 0=683]
e1 = ?
|   d2 = ?
|   |   d3a = ?: 1 [1=6, 0=0]
|   |   d3a > 95: 0 [1=15, 0=153]
|   |   d3a ≤ 95: 0 [1=0, 0=11]
|   d2 > 550: 0 [1=316, 0=3086]
|   d2 ≤ 550: 1 [1=206, 0=1]
```

(c) Criterion = Gini Index

Compared to the original DTs, the new DTs are more diverse after removing attributes related to age compared to those before removing them, with different most frequently used attributes. DT with Gain Ratio uses the following attributes: annual sales in last year (d2), audit in last year (k21), major market in last year (e1) and number of permanent, full-time employees at the end of 3 fiscal years ago. DTs with Information Gain and with Gini Index are similar

in structure and employ similar attributes: major market in last year (e1), frequency of inspections by tax officials (j4), country (a1), inspection by tax officials over last 12 months (j3), and annual sales in last year (d2).

After removing attributes related to firm age, high and low attributes' weights for SVM change as summarized in Table 13. Nine attributes have changed. The following are added into the low weight attribute list: top manager's experience (b7), country attributes (a1 = 1(India), a1 = 4(Brazil)), percentage of working capital financed from internal funds/retained earnings (k3a) and informal firm (e11 = 0). The following are added into the high weight attribute list: country = Russia (a1 = 2), percentage of working capital borrowed from banks (k3bc), major market (e1 = 2) and percentage owned by the largest owner(s). The following attributes are positively related to corruption: unknown major market (e1 = ?, 0.43), external audit conducted in the last year (k21 = 1, 0.33), Russia and China (0.31 and 0.22 respectively), and informal establishment (e11 = 1, 0.17). The following attributes are negatively related to corruption: no external audit conducted in the last year (k21 = 0, -0.52), major market is local (e1 = 1, -0.47), secured government contract (j6a = 1, -0.41), top manager's experience (b7, -0.35), major market is local (e1 = 1, -0.27) and India (a1 = 1, -21).

Table 12. Weights for Attributes of SVM without Firm Age

	Low Weight Attributes		High Weight Attributes	
	attribute	weight	attribute	Weight
1	k21 = 0	-0.52	e1 = ?	0.43
2	e1 = 1	-0.47	k21 = 1	0.33
3	j6a = 1	-0.41	a1 = 3	0.31
4	b7	-0.35	a1 = 2	0.22
5	j30f = 1	-0.34	e11 = 1	0.17
6	a1 = 1	-0.21	l1	0.15
7	j30f = ?	-0.18	k3bc	0.11
8	j6a = 2	-0.18	k3hd	0.11
9	b6a = ?	-0.18	k21 = 2	0.11
10	e11 = ?	-0.16	j6a = 0	0.10
11	l2	-0.14	e1 = 2	0.10
12	a7 = ?	-0.13	j3 = 0	0.06
13	k3a	-0.11	e1 = 3	0.06
14	e11 = 0	-0.10	j3 = 2	0.06
15	a1 = 4	-0.10	b3	0.04

6.2 Corruption Prediction by Individual Country

This research uses data from four emergent market countries - Brazil, Russia, India, and China - to obtain a larger dataset, instead of using each individual country's data. As countries may have different business contexts, they may exhibit different relations between corruption and its determinants. This question can be addressed by separating the dataset into individual countries and testing the algorithms for each country.

Performance results of DT with Information Gain, SVM and ANN by individual country without the attributes related to firm age are summarized in Table 13. The impact of splitting the dataset into individual country data varies among countries. The model constructed with the

data sets from India and Brazil achieves a better performance compared to the model constructed with the entire dataset. However, the model constructed with the dataset of Russia and China performs worse than the model constructed with the whole dataset. In particular, these two countries show very low recall and thus a very low f-measure. We are not uncertain about the cause of the difference in the changes, but can speculate several possible causes. For example, India and Brazil may provide better quality data sets when compared to Russia and China. Political systems may impact performance changes, since Russia and China boast a communist history, while India and Brazil does not. In contrast, the training dataset size may not be a cause since India provides the largest dataset while Brazil provides the smallest dataset, but they both see improvements.

Table 13. Performance Results by Individual Country without Firm Age

		TP	FP	FN	TN	accu-racy (%)	recall (%)	false positive rate(%)	true negative rate(%)	precision (%)	f-measure (%)
Deci-sion Tree (IG)	India	2043	243	1531	5464	80.9	57.2	4.3	95.7	89.4	69.7
	Russia	0	0	874	3346	79.3	0.0	0.0	100.0	N/A	N/A
	China	82	7	462	2149	82.6	15.1	0.3	99.7	92.1	25.9
	Brazil	888	165	184	565	80.6	82.8	22.6	77.4	84.3	83.6
	All	3161	473	2903	11466	81.2	52.1	4.0	96.0	87.0	65.2
SVM	India	2594	736	980	4971	81.5	72.6	12.9	87.1	77.9	75.1
	Russia	47	29	827	3317	79.7	5.4	0.9	99.1	61.8	9.9
	China	77	18	467	2138	82.0	14.2	0.8	99.2	81.1	24.1
	Brazil	856	116	216	614	81.6	79.9	15.9	84.1	88.1	83.8
	All	3938	1146	2126	10793	81.8	64.9	9.6	90.4	77.5	70.6
ANN	India	2511	827	1063	4880	79.6	70.3	14.5	85.5	75.2	72.7
	Russia	314	388	560	2958	77.5	35.9	11.6	88.4	44.7	39.8
	China	245	195	299	1961	81.7	45.0	9.0	91.0	55.7	49.8
	Brazil	888	199	184	531	78.7	82.8	27.3	72.7	81.7	82.3
	All	3860	1334	2204	10605	80.3	63.7	11.2	88.8	74.3	68.6

Ⅶ. Conclusions and Future Research

In this research, we apply machine learning algorithms to the prediction of corruption of businesses in emerging markets. In terms of accuracy and precision, certain predictive algorithms, such as DR, DT, SVM and ANN, seem to achieve our research objectives. However, these models exhibit a low performance in terms of recall relative to k-NN algorithm. While attributes related to the firm age are compelling attributes that impact corruption prediction, further analysis show that several algorithms work well even without these attributes. It is not clear whether learning predictive models by individual country can improve the performance of the prediction.

While the methodology of this study is meaningful in that it demonstrates an application of machine learning on social phenomena, the results are meaningful to practitioners in that it outlines the important identifiers of corporate corruption. In particular, decision trees and decision rules show that the duration of a company plays an important role: the older the

company, the higher the possibility of corporate corruption. When excluding the duration of a company, other factors such as the company's sales, target market, top manager's experience, source of capital as well as audit on the financial statements of firm are shown to be significant. While the relative importance of factors vary by model, the fact remains that they are still all influential determinants of corruption. Therefore, policy makers and practitioners should consider these factors in policies or regulations to prevent corruption of firms.

Despite our successful application of predictive algorithms in corruption prediction, we still need to expand upon this research. Firstly, it will be necessary to improve the algorithms in order to obtain a better prediction performance. Even though DR, DT, SVM and ANN achieve good accuracy and precision, they cannot achieve high recall. One possible improvement is to consider multiple algorithms together, which is known as ensemble learning techniques. Secondly, it is necessary to apply more data sets in order to generalize our findings. As this study focuses only on BRIC, it may not be applicable in other locales.

References

Ades, A. and R. Di Tella (1999), "Rents, competition, and corruption", *The American Economic Review*, 89, 982-993.

Arikan, G. G. (2004), "Fiscal decentralization: A remedy for corruption?", *International Tax and Public Finance*, 11, 175-195.

Bhattacharyya, S. and R. Holder (2010), "Natural resources, democracy and corruption", European Economic Review, 54, 608-621.

Biggs, D., B. De Ville and E. Suen (1991), "A method of choosing multiway partitions for classification and decision trees", *Journal of Applied Statistics*, 18, 49-62.

Brasoveanu, I. and L. O. Brasoveanu (2009), "Correlation between Corruption and Tax Revenues in EU 27", *Economic Computation and Economic Cybernetics Studies and Research*, 43, 133-142.

Breiman, L., J. H. Friedman, R. A. Olshen and C. J. Stone (1984), Classification and regression trees, London: Chapman and Hall/CRC.

Brunetti, A. and B. Weder (2003), "A free press is bad news for corruption", *Journal of Public Economics*, 87, 1801-1824.

Chowdhury, S. K. (2004), "The effect of democracy and press freedom on corruption: an empirical test", *Economics Letters*, 85, 93-101.

Churchill, R. Q., W. Agbodohu and P. Arhenful (2013), "Determining Factors Affecting Corruption: A Cross Country Analysis", *International Journal of Economics*, Business and Finance, 1, 275-285.

Cohen, W. W. (1995), "Fast effective rule induction", The 12th International Conference on Machine Learning, Tahoe-city, USA.

Cuervo-Cazurra, A. (2006), "Who cares about corruption?", *Journal of International Business Studies*, 37, 807-822.

De Graaf, G. (2007), "Causes of corruption: towards a contextual theory of corruption", Public Administration Quarterly, 39-86.

Delavallade, C. (2006), "Corruption and distribution of public spending in developing countries", *Journal of Economics and Finance*, 30, 222-239.

Dollar, D., R. Fisman and R. Gatti (2001), "Are women really the "fairer" sex? Corruption and women in government", *Journal of Economic Behavior & Organization*, 46, 423-429.

Egger, P. and H. Winner (2005), "Evidence on corruption as an incentive for foreign direct investment", *European Journal of Political Economy*, 21, 932-952.

Eiras, A. I. (2003), Ethics, corruption and economic freedom, Washington, D.C.: Heritage Foundation.
Elliott, K. A. (1997), Corruption and the global economy, Washington, D.C.: Peterson Institute.
Fatic, A. (2000), "Stability and Corruption in South-Eastern Europe." SEER-South-East Europe Review for Labour and Social Affairs, 04, 61-72.
Fredriksson, P. G. and J. Svensson (2003), "Political instability, corruption and policy formation: the case of environmental policy", *Journal of Public Economics*, 87, 1383-1405.
Freille, S., M. E. Haque and R. Kneller (2007), "A contribution to the empirics of press freedom and corruption", *European Journal of Political Economy*, 23, 838-862.
Graeff, P. and G. Mehlkop (2003), "The impact of economic freedom on corruption: different patterns for rich and poor countries", European Journal of Political Economy, 19, 605-620.
Gupta, S., H. R. Davoodi and E. Tiongson (2000), Corruption and the provision of health care and education services (Working Paper), Washington, D.C.: International Monetary Fund, 1-33. Available from https://www.imf.org/external/pubs/ft/wp/2000/wp00116.pdf
Gupta, S., H. R. Davoodi and R. Alonso-Terme (2002), "Does corruption affect income inequality and poverty?", *Economics of governance*, 3, 23-45.
Habib, M. and L. Zurawicki (2002), "Corruption and foreign direct investment", *Journal of International Business Studies*, 33, 291-307.
Imam, P. A. and D. Jacobs (2014), "Effect of corruption on tax revenues in the Middle East. Review of Middle East Economics and Finance Rev.", Middle East Econ. Fin., 10, 1-24.
Jain, A. K. (2001), "Corruption: A review", *Journal of Economic Surveys*, 15, 71-121.
Kart, L., G. Herschel, A. Linden and J. Hare (2016), Magic Quadrant for Advanced Analytics Platforms. Gartner. Available from https://www.gartner.com/en/documents/3204117/magic-quadrant-for-advanced-analytics-platforms (accessed July 19, 2016)
Kaufmann, D. and S. J. Wei (1999), Does "grease money" speed up the wheels of commerce? (Working Paper), Cambridge: National Bureau of Economic Research (NBER), 1-29. Available from https://www.nber.org/system/files/working_papers/w7093/w7093.pdf
Kubat, M., R. C. Holte and S. Matwin (1998), "Machine learning for the detection of oil spills in satellite radar images", Machine Learning, 30, 195-215.
Lambsdorff, J. G. (1999), Corruption in empirical research: A review (Working Paper), Berlin: Transparency International, 1-17. Available from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.463.2378&rep=rep1&type=pdf.
Lambsdorff, J. G. (2006), "Causes and consequences of corruption: what do we know from a cross-section of countries". In S. Rose-Ackerman (Eds.), International handbook on the economics of corruption, 3-51.
Leff, N. (1964), "Economic development through bureaucratic corruption", American Behavioral Scientist, 8-14.
Leite, C. A. and J. Weidmann (1999), Does mother nature corrupt Natural resources, corruption, and economic growth (Working Paper), Washington, D.C.: International Monetary Fund, 1-34. Available from https://www.imf.org/external/pubs/ft/wp/1999/wp9985.pdf
Lewis, D. D. and W. A. Gale (1994), "A sequential algorithm for training text classifiers", The 17th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland.
Majeed, M. T. (2014), "Corruption and Trade", Journal of Economic Integration, 759-782.
Mauro, P. (1995), "Corruption, country risk and growth", *Quarterly Journal of Economics*, 3, 681-712.
Mierswa, I., M. Wurst, R. Klinkenberg, M. Scholz and T. Euler (2006), "Yale: Rapid prototyping for complex data mining tasks", The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA.
Mierswa, I., M. Wurst, R. Klinkenberg, M. Scholz and T. Euler (2006), "Yale: Rapid prototyping for complex

data mining tasks", The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA.

Montalvo, J. G and M. Reynal-Querol. "Ethnic Polarization, Potential Conflict, and Civil Wars." *American Economic Review*, vol. 95, no. 3, 2005, pp. 796-816.

Myrdal, G. (1968), Asian Drama: An Inquiry into the Poverty of Nations, New York, Pantheon.

North, C. M., W. H. Orman and C. R. Gwin (2013), "Religion, corruption, and the rule of law", *Journal of Money*, Credit and Banking, 45, 757-779.

Nur-Tegin, K. and H. J. Czap (2012), "Corruption: Democracy, Autocracy, and Political Stability", *Economic Analysis and Policy*, 42, 51-66.

Pellegrini, L. (2008), "Causes of corruption: a survey of cross-country analyses and extended results", *Economics of Governance*, 9, 245–263.

Pieroni, L. and G. D' agostino (2013), "Corruption and the effects of economic freedom", *European Journal of Political Economy*, 29, 54-72.

Provost, F. J., T. Fawcett and R. Kohavi (1998), "The case against accuracy estimation for comparing induction algorithms", The 15th International Conference on Machine Learning, Madison, Wisconsin, USA.

Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proceedings of the Fifteenth International Conference on Machine Learning. Morgan Kaufmann (1998) 445–453

Quinlan, J. R. (1992), C4.5: Progams for Machine Learning, San Mateo, CA: Morgan Kaufmann Publisher.

Rose-Ackerman, S. (1999), Corruption and Government: Causes, Consequences, and Reform, Cambridge: Cambridge University Press.

Rumelhart, D. E., G. E. Hinton and R. J. Williams (1985), "Learning internal representations by error propagation". in A. Collins and E. E. Smith (Eds.), Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence, CA: Morgan Kaufmann Publisher, 399-421

Sanchez, J. I., C. Gomez and G. Wated (2008), "A value-based framework for understanding managerial tolerance of bribery in Latin America", *Journal of Business Ethics*, 83, 341-352.

Shen, C. and J. B. Williamson (2005), "Corruption, democracy, economic freedom, and state strength a cross-national analysis", International Journal of Comparative Sociology, 46, 327-345.

Stendahl, L. (2016), Fighting Corruption: A Cross-National Study on the Effect of Reserved Legislative Seats for Ethnic Groups on Corruption (Unpublished Bachelor's Thesis), Uppsala: Uppsala University, 1-42

Swaleheen, M. U. and D. Stansel (2007), "Economic freedom, corruption, and growth", Cato J., 27, 343.

Swamy, A., S. Knack, Y. Lee and O. Azfar (2001), "Gender and corruption", *Journal of Development Economics*, 64, 25-55.

Tanzi, V. (1998), Corruption around the world: Causes, consequences, scope, and cures (Working Paper), Washington, D.C.: International Monetary Fund, 1-39. Available from https://www.imf.org/en/Publications/WP/Issues/2016/12/30/Corruption-Around-the-World-Causes-Consequences-Scope-and-Cures-2583

Tanzi, V. and H. R. Davoodi (1997), Corruption, public investment, and growth (Working Paper), Washington, D.C.: International Monetary Fund, 1-23. Available from https://www.imf.org/external/pubs/ft/wp/wp97139.pdf

Tanzi, V. and H. R. Davoodi (2000), Corruption, growth, and public finances (Working Paper), Washington, D.C.: International Monetary Fund, 1-27. Available from https://www.imf.org/en/Publications/WP/Issues/2016/12/30/Corruption-Growth-and-Public-Finances-3854

Transparency International (2016a), How Do You Define Corruption? Available from http://www.transparency.org/what-is-corruption/#define (accessed July 19, 2016).

Transparency International (2016b), What are the Costs of Corruption? Available from

http://www.transparency.org/what-is-corruption/#define (accessed July 19, 2016).
Treisman, D. (2000), "The causes of corruption: a cross-national study", Journal of Public Economics, 76, 399-457.
Wei, S. J. (2000), "How taxing is corruption on international investors?", Review of Economics and Statistics, 82, 1-11.
Yehoue, E. B. (2007), Ethnic diversity, democracy, and corruption (Working Paper), Washington, D.C.: International Monetary Fund, 1-22. Available from https://www.elibrary.imf.org/view/journals/001/2007/218/article-A001-en.xml
Zhang, G. P. (2000), "Neural networks for classification: a survey", IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30, 451-462.