National culture ‘profiling’ in machine-learning applications: The utility and ethics of applying value ascriptions in global alert models

John W. Goodell[a], Cal Muckley, Parvati Neelakantan, Darragh Ryan[1], Pei-Shan Yu[1]

[a] *The University of Akron, Akron, Ohio, USA*
[b] *University College Dublin, Belfield, Dublin, Republic of Ireland.*

Abstract

We examine the utility and ethics of incorporating national culture profiling in bank-level machine-learning informed alert models, which relate to financial malfeasance. At a globally important financial institution, we use binary classifier type alert models and establish the utility of dimensions of national culture in formulating anti-money laundering predictions. For corporate (individual) accounts, Hofstede individuality (individuality, and national-level corruption perception and financial secrecy) scores of the country in which a customer is resident, or from which a wire is sent/received, are of paramount importance. When combined with extensive account and transaction data; as well as even a proprietary institutional algorithm, national culture traits markedly enhance the models' predictive performances. We consider the ethical implications of ascribing values, against a global standard, to dimensions of national culture. We offer an ethical framework for the use of national profiling in anti-fraud alert models.

Keywords: National Culture Profiling, Machine Learning, Anti-Money Laundering
JEL Classification: C52, C55, C58

1. Introduction and Motivation

According to a 2019 survey by Forrester, as reported in the Financial Times, at least half of financial services and insurance firms use artificial intelligence, with this percentage expected to grow. Artificial intelligence is used in a wide variety of ways by these firms, including to analyze financial markets, and to process large amounts of data (e.g. individual income and spending patterns) related to identifying anomalies to detect fraud. Artificial intelligence is supported by several technologies including robotic process automation, natural language understanding, and, the focus of this paper, machine learning. Particularly relevant, to our paper, machine learning is now commonly deployed by banks to detect fraud.

Machine learning is also making in roads into business research, particularly in the areas of finance and economics. For instance, Bianchi, Büchner, and Tamoni (2020) apply machine learning to estimating bond risk premiums. Gu, Kelly, and Xiu (2018) examine the use of machine learning for asset pricing. However, in contrast to the typical practice of financial economics research, machine learning is often conducted with consideration of utility toward an application rather than in an interest of hypothesis testing. Being often solely data-driven, machine learning typically provides a forecasting, or in some cases a detection model, with evaluation of its forecast ability relative to random prediction. Therefore, the need, and in some cases the potential social awkwardness, of forming, and perhaps publicly identifying, testable hypotheses is avoided. Consequently, a demographic or national characteristic found via machine learning to be a strong predictor of, for instance, bank-fraud, may not be supported by hypotheses. Machine learning procedures present a point of departure for both business applications, as well as for business research methodology.

Issues of profiling have recently been to the fore of the public mind around the world, largely because of the sweeping reaction to the killings in the US of George Floyd and other persons of color in police hands. While, of course, there are considerable differences between racial profiling and criminal profiling, it is also arguable that profiling in general, as it is applied in various contexts, readily leads to blurring of such distinctions. Clearly, there is a need to consider, in a variety of contexts, the ethics of profiling inputs to machine learning alongside their intended uses.

We first test and establish the utility of national culture traits to inform a machine learning alert model for the detection of money laundering at a globally important financial institution. It is important to consider whether national factors can potentially function as profiling factors, with concomitant ethical concerns. Certainly, discussions regarding trade-offs between the predictive efficacy of machine learning models, and the ethics and societal implications of including national culture inputs to machine learning models would be far less pressing if it is generally found that such inputs have little predictive power.

We find country-level factors, particularly national culture as comprising strong predictors of identifying suspect bank wire transfers. Using binary classifier type alert models, together with corrections for data imbalance, our results reflect the strength of national culture dimensions in formulating anti-money laundering predictions. For corporate (individual) accounts, Hofstede Individuality (Individuality, and national-level corruption perception and financial secrecy) scores of the country in which a customer is resident, or from which a wire is sent/received, are the most important factors. National culture alone provides a high degree of predictive power. And when combined with extensive account and transaction data; as well as even proprietary institutional algorithms already in use, its inclusion greatly enhances predictive ability.

We consider our findings relevant to both the conduciveness of machine learning to incorporating national culture; as well as reflecting on the wide body of research that has ascribed values, particularly with regard to ethical practice and discernment, to cultural dimensions.

The role of national culture has been affirmed in a wide array of business research (Kirkman, Lowe, and Gibson, 2006). Karolyi (2016) notes an uneasiness with Hofstede's construction, but also notes the almost uncanny tendency for research findings to flow from what would be expected from using Hofstede's dimensions. National culture is often supported in research by strongly framed hypotheses, in contrast to what might ensue from use of racial and demographic qualities in detection and alert models. Indeed, use of national culture in machine learning can be seen as a gray area between use of ethnic and gender subsets of cultures, with obvious concomitant issues of social injustice and unfairness, and yet opening a door to such outcomes. Within global commerce and banking, there is a need to consider unfairness toward members of some nations over others. Further, national culture has been modeled on subnational levels, particularly with religion-based demographic data (Stulz and Williamson, 2003). While ascriptions of values, perhaps laudably, opens the door to considering that differing cultures evolve their own meanings of values, governance, and corruption etc. But use of national culture in machine-learning detection and alert models also invites global unfairness by assigning attributes to individuals that are assigned in aggregate in ways that possibly lack transparency.

Our paper strongly connects to business ethics research by considering the use of national culture in machine-learning. National culture has prominently been associated with qualities of ethics and discernment in business ethics research. Additionally, our study and discussion follows on from recent articles on the ethical issues inherent in the design of algorithms deployed in society. Our paper connects to the very timely issue of acknowledging that latent racial and ethnic biases may be present in any sort of functional profiling or predictive model.

Regarding the contents and transparency of algorithms, for instance, Martin (2019) notes how algorithms can influence hiring, promotion, and loan approvals. And yet such life-structuring algorithms are usually devised without transparency. Buhmann, Paßmann, and Fieseler (2019) express considerable concern regarding algorithmic accountability. This includes concern over how algorithms are established—similar to absence of hypothesis formation we note above as a underpinning of machine-learning based research. Seele et al. (2019) also highlight issues regarding the quality of transparency of algorithms, particularly with regard to dynamic pricing models.

We highlight national culture as a predictor of bank fraud. This naturally follows on from a broad stream of literature connecting national culture dimensions to the quality of ethical behavior and perception (e.g.; Armstrong, 1996; Davis and Ruhe, 2003; Getz and Volkema, 2001; Vitell, Nwachukwu, and Barnes, 1993; Volkema, 2004). For instance Vitell, Nwachukwu, and Barnes (1993) propose a theoretical application of Hofstede's cultural dimensions as a basis for understanding the effects of culture on ethical decision making. That national culture has been proposed as an important factor in understanding ethical decision making is intuitive, as culture represents the shared beliefs, values, and ideals of a society.

Vitell, Nwachukwu, and Barnes (1993) suggest that greater hierarchy, relating closely to Hofstede's Power Distance, leads to ethical cues being taken from superiors rather than peers, with formal codes of ethics concomitantly having a greater influence than informal norms. Related, Getz and Volkema (2001) conclude that both high-level public officials and members of the underclass are more susceptible to unethical behavior (bribery, extortion) in high power distance cultures. In their view, high-ranking officials exploit class privilege to obtain personal benefits from their official positions, while members of the lower classes in highly hierarchical and unequal societies justify unethical behavior as a reasonable to need to catch up their standards of living. Volkema (2004) suggests that power distance is positively associated with the use of competitive and dubious negotiation practices.

With regard to individualism, Vitell, Nwachukwu, and Barnes (1993) posit that countries high in individualism will be less likely than countries high in collectivism to take into consideration both formal codes of ethics and informal norms. Being more oriented towards self-reliance, freedom, and achievement, someone from an individualist culture might be inclined to see his or her actions as above reproach. Similarly, Volkema (2004) suggests that individualism will be directly related to the perceived appropriateness and the likelihood of using competitive and questionable negotiation behaviors.

Volkema (2004) suggests that the cultural dimension of masculinity will be directly related to the perceived appropriateness and the likelihood of using competitive and questionable negotiation behaviors. Volkema (2004) suggests that individuals from a masculine culture are more likely to exude exaggerated self-promotion and aggressive bidding for new clients. Getz and Volkema (2001) contend that masculine cultures are more likely than feminine cultures to engage in bribery and corruption because achievement in masculine cultures is measured by commercial success, with the ends thought to justify the means.

Volkema (2004) considers that uncertainty avoidance will be inversely related to the perceived appropriateness and the likelihood of using competitive and questionable negotiation behaviors. Similarly, Vitell, Nwachukwu, and Barnes (1993) suggest that business practitioners from societies that are strong on uncertainty avoidance are more likely to be relatively more intolerant of any deviations from group norms. Vitell, Nwachukwu, and Barnes (1993) suggest that business practitioners in countries that are high in uncertainty avoidance (e.g., Japan) will be more likely to consider formal professional codes of ethics when forming their own deontological norms than business practitioners in countries that are low in uncertainty avoidance (e.g., the U.S. or Canada). Vitell, Nwachukwu, and Barnes (1993) also suggest that business practitioners in countries that are high in uncertainty avoidance will be less likely to perceive ethical problems than business practitioners in countries that are low in uncertainty avoidance. Aggarwal, J. E. Goodell, and J. W. Goodell (2014) amalgamate literature associating Hofstede's dimensions with levels of ethical discernment and practice, finding that higher GMAT test takers from less ethical national cultural backgrounds score higher.

Further, accounting literature, with a natural emphasis on reporting quality, has for many years considered varying levels of firm-level of transparency in terms of national culture. In a seminal paper on cultural accounting, Gray (1988) hypothesizes that power distance is negatively associated with transparency (positive with secrecy). Gray (1988) suggests that this is because less information is needed to preserve power inequalities. Correspondingly, De Jong, Smeets, and Smits (2006) find a negative association of openness with power distance (see also: Velayutham and Perera, 2004). However, Zarzeski (1996), Jaggi and Low (2000) and Hope (2003) all find a positive association of financial disclosure and power distance. Although, other research (e.g.; J. J. Archambault and M. E. Archambault, 2003; Salter and Niswander, 1995) are inconclusive regarding the association of power distance and financial disclosure. Cohen, Pant, and Sharp (1996) find a majority of experts predicting a positive relationship between power distance and unethical behavior.

Clearly, national factors, particularly cultural factors for machine-learning generated prediction or detection presents a host of ethical questions. Are we going to deploy loan-application and insurance-claim verification models etc. that treat clients from some countries with greater scrutiny than others? This issue becomes more compelling when we consider the number of papers that have used religion as a proxy or instrument for culture. For instance, research looks at relative proportions of Catholics versus Protestants in US countries as a guide to understanding differences in dividend payouts etc (e.g. Ucar (2016)). Will US citizens have a greater or lesser chance of having loans approved if they live in one county versus another?

The use of demographic inputs, particularly country-level factors and cultural factors in machine learning models touches on a wide variety of literature that incorporate cultural and demographic variables that imply ascriptions of value to these characteristics.[1] We offer, a framework for assessing the ethics of using country-level factors in machine learning prediction and detection. We identify several characteristics of the use of country-level factors in machine-learning procedures that are central to evaluating the ethics of their respective usage. These include: 1) Do public good concerns in countering money laundering outweigh 'collective treatment' concerns in national profiling in algorithms? 2) Do those producing the alerts have permission to use the personal data? 3) Who is responsible for the design of an algorithm? 4) Are algorithms accountable? 5) Are the algorithms used for detection, or, alternatively, for prediction?—and are there subtle distinctions regarding this?; 6) Are alert models reflective of global, national or sub-national; public or private regulation?; 7) Do the algorithms in use exacerbate tangential societal biases? and 8) Can the deployment of an algorithm, due to automation, transform the workplace?

The rest of this paper is organized as follows: Section 2 provides rationale for our test

[1] See, for instance, the area of microfinance, where the notion of female borrowers being more trustworthy is a pillar of the industry (Aggarwal, J. W. Goodell, and Selleck, 2015)

of whether national culture can inform anti-money laundering alert models, and whether the inclusion of such national traits is ethical. Section 3 provides an overview of the problem of global money laundering. Section 4 describes the nature of the data used in our research, the dependent variable and the features we construct to model it. We also discuss the challenges of modelling an unbalanced data set and the solutions we use to combat this problem. Section 5 illustrates the machine learning algorithms we use to model our data set, the theory behind them and how they compare to one another. Section 6 presents the results of our models and explains the metrics used for their evaluation as well as the inferences we draw from the algorithms' outputs. Section 7 discusses a framework for evaluating the ethics of machine learning prediction and alert models. Section 8 concludes.

2. Rationale

Our study on the efficacy of machine learning models to predict money laundering with a small number of national factors, as well as account and transaction level data, raises a number of issues: 1) Is national culture a valid predictor of an individual's behavior? This question involves the debate over the veracity of national culture as part of a well-reasoned mechanism to predict an individual's behavior. 2) Can machine learning procedures provide accurate and pragmatic money laundering alerts when supplied with only a small number of national factors? How important are these national factors in the context of heavily parametrized models including account and transaction level feature data? This question involves consideration that machine-learning procedures are often seen as providing advantages, e.g. capturing non-linearities in the data, over other more traditional modelling in the particular context of 'big data' (Coulombe et al., 2020). Algorithms based on a small number of variables are presumably more transparent than heavily parameterized models; although their non-linear specifications may still leave these models as opaque. 3) Considerable regard needs to be given to whether we should endorse public or private use of machine learning algorithms based on simple national factors generally, and national culture scores in particular. Is the use of such algorithms fair? And do they give "voice to values" (Arce and Gentile, 2015)? In other words do placing into operations algorithms based on endowed national culture implicitly endorse certain culture traits as 'positive' and other traits as 'negative'?

2.1. Is national culture a valid predictor of individual behaviour?

Noting that national culture has been considered as an important factor in a broad range of contexts (Kirkman, Lowe, and Gibson, 2006; Kirkman, Lowe, and Gibson, 2017), we consider that national culture facets (i.e. societal cultural value dimensions (Peterson and Barreto, 2018)) can indicate the societal context of individuals. This entails a dual process of, at one level, acknowledging the cognition of respective individuals: personal attitudes and values together with, at another level, apprehending those societal culture facets that also inform individuals' cognition.

Consistent with argumentation in Peterson and Barreto (2018), we suggest that national culture facets can reflect contextual characteristics which more strongly shape an individual's cognition than do consciously expressed personal values. For instance, it can be comparatively insightful to understand the societal context of an individual, as opposed to self-professed attitudes and values. Doing so is not equivalent to erroneously assigning societal characteristics to individuals as a replacement of omitted information about their personal values (Tung and Verbeke, 2010; Tung and Stahl, 2018). Rather, highlighting the contextual effects of societal institutions and norms can inform 'individuals experience and, hence, what people unconsciously intuit and consciously understand (Peterson and Barreto, 2018; J. W. Goodell, 2019). This can, in turn, contribute to an individual's actions, cognition, and choices. In this way, national culture facets identify the context in which a society's members react to culture. More specifically, in this study, we consider if a banking customer's opportunity and inclination to commit financial misconduct – laundering money, is informed by his/her cultural context.

We, hence, consider that the propensity of banking service clients toward malfeasance, can vary markedly across national cultures. As individuals may not always hold unbiased beliefs and can behave irrationally (J.-B. Kim, Wang, and Zhang, 2016), the anticipated incentives and deterrents of misconduct and the anticipated likelihood of being held to account for wrongdoing, can vary substantively across national cultures (Husted, 2000). The social normative nature of national culture (J. W. Goodell, 2019), in particular, can influence misconduct exhibited by the customers of financial institutions. We therefore hypothesize the following.

H1 A limited number of national factors, or even national culture alone, can accurately forecast the likelihood of money laundering.

2.2. Explanatory power of parsimonious models of national predictors.

Literature associates national culture with a host of finance-related behaviour. For instance, the propensity of market investors to have excessive confidence (Chui et al 2010), as well as to have greater predilection to hedge (Lievenbruck and Schmid, 2014) and to prefer to contracts to relationship financing (Aggarwal and Goodell 2009); choice of firm leverage (Chui, Lloyd and Kwok, 2002), ambition to undertake investment (Shao et al. 2013) and the design of public financial policies (Aggarwal and Goodell 2013). Clearly given the breadth of studies associating national culture with behaviour in business, there are ample reasons to consider the national culture will impact the predlilection for bank fraud.

2.3. Explanatory power of parsimonious models of national predictors.

Machine learning algorithms are often highly complicated and consequently difficult to explain—or to justify (Barocas, Hood, and Ziewitz, 2013; Introna, 2016; Musiani, 2013; Seaver, 2019; Ziewitz, 2016). This complexity can act against assigning responsibility to the developer or to the user, as this assignment is deemed inefficient and even impossible. Algorithms based primarily on a small set of national factors, which are therefore comparatively transparent, can have adequate explanatory power to accurately prompt money laundering alerts. It may turn out, to some extent at least, that unaccountable and complex algorithms, inherently lacking transparency, are not needed. Or at least, if the algorithm of a machine learning procedure remains complex, the inputs used as factors need not be.

We investigate whether algorithms using just simple national factors, corresponding to the procedures we highlight in this paper, can be very effective. Therefore, we evidence that complexity of algorithms, with concomitant lack of transparency, may not be needed. While the inscrutability of algorithms has often been highlighted as both necessary to achieve predictive performance, and as a means to avoid accountability (Desai and Kroll, 2017), our results, at least in the context of this study, suggest otherwise.

3. Money Laundering

Money laundering is an issue of global importance, undermining local economies and penetrating through borders. It facilitates the generation and disbursement of illicit proceeds from criminal activities through integration into the financial system, which can be used to further finance illegal activity, compounding the problem. Although difficult to measure with any degree of confidence, estimates for the total amount of money laundered worldwide range from 2-5% of global GDP (approximately $600 billion to $1.6 trillion).

The process of money laundering can generally be regarded as a procedure for transforming *dirty money* (i.e. money generated from illegal activities) to *clean money* (i.e. doesn't raise suspicion) by integrating it into any available legitimate financial system, so it can be subsequently used without raising any suspicion. In short, the money is transacted so as to conceal or obscure its link with its criminal origins.

Efforts in combating money laundering require cooperation between the public and private sectors. However, current compliance requirements, such as transaction monitoring and suspicious activity reporting, impose significant costs on the private sector with very limited returns. Nowhere is the impotency of this endeavour highlighted better than in the recent high-profile scandals involving Danske Bank and Swedbank. The burden of anti-money laundering (AML) surveillance is shared by all financial institutions in the private sector, with failure to meet certain standards of monitoring punishable by hefty fines and penalties imposed by the financial governing bodies. AML surveillance can be painstakingly inefficient -– thousands of potentially suspicious transactions need to be vetted each day in a time-consuming and labour-intensive process, but since the failure to identify and report suspicious activity can be so costly, financial institutions tend to err on the side of caution by casting a wide net, which unfortunately leads to a greater rate of *false positives*[2].

Since 2000, the International Monetary Fund (IMF) has redoubled its work on AML and, after the tragic events of September 11th 2001, its activities have been expanded to include combating the financing of terrorism (CFT). In 2009, the IMF launched a donor-supported trust fund to finance AML/CFT capacity development in its member countries. On top of government and international body expenditure on AML programs, the private sector also has had to foot the bill in order to comply with the rules set out in the Bank Secrecy Act (BSA) in the US. According to a 2018 report, Lexis Nexis Risk Solutions estimates that AML compliance costs US financial firms approximately $25 billion annually.

By its very nature, money laundering is difficult to measure since it occurs outside the forum of normal economic activity. People rely on inference from best relevant data that are available most, if not all of the time. An example of such data is the 2002 National Money Laundering Strategy, an annual report from 1999-2003 by the US Treasury on Anti-Money Laundering (AML) efforts. According to this report, $386 million worth of assets were seized in relation to money laundering in 2001, with a corresponding figure of $ 241 million in forfeited assets. However, such sums are considered only a small fraction of the true total. Various techniques and schools of thought have been employed in order to make reliable and consistent estimation on the extent of money laundering. The macroeconomic approach holds that the demand for money laundering is related to the monetary component of the so called shadow economy, and tools such as currency-demand analysis (Tanzi, 1980) prove useful in this regard. One study, conducted by the United Nations Office on Drugs and Crime (UNODC), investigated the volume of illegal funds generated by drug trafficking and organised crime and to what extent

[2] Essentially, a false alarm, whereby seemingly innocuous activity has unnecessarily garnered attention.

these funds are laundered. Their findings estimated that, in 2009, criminal activity amounted to 3.6% of global GDP with 2.7% being laundered, amounting to $ 1.6 trillion.

In light of the recent work done on the role culture plays in corporate misconduct and bank failure (Liu, 2016; Berger et al., 2019), we explore the relevance of several country-specific cultural and institution quality indices in the context of modelling the incidence of suspicious money movement within a financial institution. Another country-specific measure incorporated in this study is the concept of secrecy jurisdiction, introduced by Cobham, Jansk, and Meinzer (2015). They suggest jurisdictions are situated across a spectrum of secrecy in terms of financial sector and global market share. This is in contrast to binary classification of Tax Haven/Offshore Finance. The aim here is to shift the narrow tax-focused narrative onto a broader sense of financial secrecy and transparency, which eventually may facilitate changes in policy and practice. We believe this measure is particularly relevant in the context of money laundering activity detection, as a jurisdiction with a higher level of secrecy in financial sector is more likely to attract higher volume of transactions initiated with the intent of concealing the illegal origin, as the regulations of such jurisdiction make it more difficult to obtain necessary information to trace the money flows. We benefit from a large proprietary dataset containing cross-border money movements via wires in a major global financial institution. A proportion of the wires are flagged as 'suspicious activities' by the institution's designated investigative team, which can be regarded as a precursor to money laundering. We further draw inspirations from an extensive literature of machine learning and its application in various finance context, such as fraud detection, default risk rating, etc. (Khandani, A. J. Kim, and Lo, 2010; Butaru et al., 2016; Kumar et al., 2019). Our data set provide us with a clearly labelled response variable (ISSUE), and hence supervised learning for a classification problem is the suitable methodological framework. We employ four different machine learning algorithms, which are derived from classification and regression trees (CART) (Breiman et al., 1984) and all popular choices among academics as well as data science practitioners. We find that the introduction of these variables complement the institution's own account and transaction-level data, since the inclusion of these predictors as an added layer of attributes enhance the performance of our models. We aim to provide practical implications for the financial services sector in terms of AML compliance strategy in this study. In additional to existing practices already in place, such as Know-Your-Customer (KYC), AML operations within the private sector could further benefit from incorporating geopolitical or regulatory information, and hence the investigative resources could be concentrated on these money laundering hot-spots.

We believe findings in this study can provide practical implications for the financial services sector in terms of AML compliance and prevention strategy. The introduction of country-specific variables complement the institution's own account- and transaction-level data, since the inclusion of these predictors as an added layer of attributes enhance the performance of the predictive models. In additional to existing practices already in place, such as Know-Your-Customer (KYC), AML operations within the private sector could further benefit from incorporating geopolitical or regulatory information, and hence the investigative resources could be concentrated on these money laundering hot spots. As indicated in an IIF (Institute of International Finance) study, the potential benefits of applying machine learning in anti-money laundering operation are inevitably faced with several challenges as well. A few key aspects highlighted in the study include AML specific challenges such as data quality, obstacles regarding data sharing and legacy/dated IT infrastructure, while machine learning-specific challenges such as was for ML talents, generalisation of trained models as well as interpretation of results, among other issues. We have first-hand experience with some, if not all, of these challenges during different stages of this study, in particular in the aspects of data quality, legacy IT systems, data sharing and protection, and the nature of the real-world data having extremely imbalanced classes. Nevertheless, we utilise the data to the best of our knowledge and obtain satisfactory results which provide insights and empirical evidence that how financial institutions can benefit from incorporating machine learning and publicly available data along with their own data to enhance AML operation.

4. Data

In this study, we use a large proprietary data set from a major international financial institution collected over a decade, from 1$^{\text{st}}$ January 2009, to 31$^{\text{st}}$ December 2018. The data pertains to alerts generated by international wire transfers both to and from customers of that institution, whereby details relating to the wire amounts and countries involved automatically flag up potentially suspicious activity on their monitoring systems. The alerts are subsequently investigated by a dedicated team of experts and those alerts deemed highly suspicious are escalated to the status of *issue case* and passed on to a higher authority for processing.

The accounts associated with the alerts number greater than 60,000 and can be broadly split into six different categories.[3] For the purposes of our study, it was helpful to focus on two categories, or account registration types; corporate-related and people-related. These account for 78.23% of the alerts and 93.77% of the issue cases and differ fundamentally in their nature of activity, and so it seems only natural to treat them as separate problems. Table 1, Panel A illustrates the incidence of alerts and subsequent issue cases by year and how these numbers break down over these two categories. By only considering these two registration types, our data set reduces from 206,751 alerts to 153,917 alerts.[4]

4.1. Sample Selection

As part of our data set, we have access to details on customers, accounts and their transaction history, in addition to the information relating to the wire transactions that triggered the alerts. As such, we will seek to build our model variables from this data (see Section 4.2) but, unfortunately, not all the information available to us is complete. Table 1, Panel B illustrates the number of alerts that can be successfully matched to each selection criteria. Due to incompleteness in our data, only about 60% of the alerts can be matched to the wire transactions that triggered them on the day. The next issue with our data incompleteness is the Customer Age information, however this has more of an ontological reason: many accounts associated with the alerts belong to corporate entities and obviously cannot be assigned a date of birth, as such. As you will see later on, we will only include this information in our models when we are looking at people-related alerts in isolation.[5]

[Please insert Table 1 about here.]

[3] See Internet Appendix A.

[4] See Panel B in Table 1.

[5] As a consequence, people-related models will always contain an additional Customer Age variable.

4.2. Feature Selection

From the data set available to us, we have grouped the candidate predictors available to us into 3 broad categories; (1) Account-level, (2) Country-level and (3) Transaction level. The details are summarize in Table 2. Note again that the Customer Age predictor can only be used with the people-related alerts.

4.2.1. Country-level Predictors

For each alert, we distinguish between the origin/destination country of the wire transactions that comprise that alert, and the residence country of the customer receiving/sending the wire. The reason for this seemingly arbitrary dichotomy (as opposed to simply differentiating between sending and receiving country) is due to the asymmetrical nature of our data: the residence country of the client is reliably documented whereas there is sometimes uncertainty about the identity of the country to/from which the client is sending/receiving the wire.[6] To translate this information into quantifiable numbers, we use several internationally recognised indices that attempt to measure the levels of corruption and financial secrecy of a country and the observed cultural measures that may be relevant to suspicious wire activity.

1. **Corruption Perception Index** Provided by Transparency International, this index ranks countries "by their perceived levels of public sector corruption, as determined by expert assessments and opinion surveys."

2. **Financial Secrecy Index** Provided by the Tax Justice Network, this index ranks countries according to their secrecy and the scale of their offshore financial activities. A politically neutral ranking, it is a tool for understanding global financial secrecy, tax havens or secrecy jurisdictions, and illicit financial flows or capital flight. (Puspitasari et al., n.d.; Houqe et al., 2015; Michalos and Hatch, 2019; Hassan and Giorgioni, 2015)

3. **Uncertainty Avoidance Index** One of Hofstede's dimensions of culture, this index measures the tolerance a society has for the unknown or ambiguous. A country high on this scale typically feels uncomfortable with uncertainty and so seeks to instill beliefs and institutions that provide certainty and conformity, avoiding unorthodox behaviours. Countries at the other end of the scale tend to be have a more relaxed attitude to legislation and take more risks, which can be a good thing (innovation) or a bad thing (bank failure).

4. **Masculinity Index** One of Hofstede's dimensions of culture, this index measures the degree to which typically male characteristics (competitiveness, heroism, assertiveness, leadership, achievement etc.) are valued in a society. At the other end of the scale, societies with more feminine values promote cooperation, modesty, duty of care to more vulnerable members of society, etc.

[6] For example, sometimes an IBAN is included, other times information in the address or instructions field needs to be used to help identify the country.

5. **Individualism Index** One of Hofstede's dimensions of culture, this index measures the degree to which a society values the role of the individual versus that of the collective. Hofstede defines this as: "a preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families." At the other end of the scale, a more collectivist society has broader criteria for the groups they identify with, which take care of their members in exchange for unquestioning loyalty.

6. **Power-Distance Index** One of Hofstede's dimensions of culture, this index measures the extent to which the authority of people in positions of power is accepted by those lower down the food chain. Typically, in a country high on this scale, the population holds relatively authoritarian views, based on more traditional, rather than secular, arguments. These societies tend to be more stratified and display more conformity. At the other end of the scale, we have a society that strives to equalise power and more readily decry injustices or abuses of power.

Thus, the customer's country of residence, and the country of wire origin/destination contribute two sets of variables, which we will distinguish by the subscripts R and W denoting "Residence country" and "Wire country" respectively.

4.2.2. Account-level Predictors

The Customer Age is defined as the age of the customer (i.e. private individual) on the date the alert is generated.[7] Similarly, the Account Age is the time elapsed between the date the account was establishes and when the alert was triggered. The Customer Net Worth is defined as the sum total of the balances on all the accounts belonging to that customer. The Alert Supplier Code relates to one of two systemic methods that the monitoring system of the major international financial institution uses when collecting the alerts.

4.2.3. Transaction-level Predictors

The variables belonging to this category come from 2 channels; (1) the wire transactions that triggered the alert (2) the transaction history of the account associated with an alert. For the wire variables, we measure the number of incoming wires, the aggregated amount of incoming wires, the standard deviation of incoming wire amounts, the number of outgoing wires, the aggregated amount of outgoing wires, the standard deviation of outgoing wire amounts. For the transaction history variables, over a 180 day period preceding an alert, we measure the number of incoming transfer-type transactions, the aggregated amount of incoming transfers, the number of outgoing transfers, the aggregated amount of outgoing transfers, the number of incoming check-type transactions, the aggregated amount of incoming checks, the number of outgoing checks, the aggregated amount of outgoing checks.

[Please insert Table 2 about here.]

[7] This predictor only applies to alerts associated with people-type accounts.

4.3. Dependent Variable

The dependent variable in this study is the outcome of an investigation − specifically, whether or not an alert is deemed to be highly suspicious, i.e. an *issue case*. As mentioned earlier, the raising of alerts is an automated process, determined by a customer's aggregated wire transactions exceeding a certain threshold, for wires involving blacklisted countries on a given day. A team of investigators is tasked with examining in detail, each case flagged up by the alert system. An alert passes through several phases of escalation before reaching the ultimate status of issue case, at which point the case is passed onto the authorities for legal processing.

Though highly inefficient,[8] this method of screening transactions for suspicious activity remains industry standard for the simple reason that financial governing bodies enforce harsh penalties for institutions judged too lax in such matters.

5. Methodologies

This section presents a discussion of the various data resampling methods used in this paper to meaningfully draw information from the data. It then presents a discussion of the machine learning methodologies employed and the performance evaluation metrics used to evaluate the models. Finally, it presents a discussion on feature importance. We discuss the data resampling techniques in subsection 5.1; machine learning methodologies in subsection 5.2; performance evaluation metrics in subsection 5.3; and feature importance in subsection 5.4.

5.1. Data Balancing

The dependent variable suffers from severe class imbalance. That is, the number of observations belonging to the positive class (issue case) is significantly exceeded by the number of observations belonging to the negative class (generated alert not an issue case). Models trained on such data prioritize the prevalent class at the expense of the minority class which leads to an overly optimistic measure of accuracy. Such models can detect a non-fraudulent transaction with high accuracy; however, they may fail to detect highly suspicious transactions. If the highly suspicious transactions go undetected, then they may pose a threat to the financial institutions' professional credibility and could also lead to regulatory sanctions on them.

In this study we avail of various data-resampling techniques to overcome the challenges posed by the imbalanced class distribution. Below, we discuss the resampling techniques employed in our study.

1. **Under-sampling:** This technique randomly discards observations from the majority class to better balance the skewed distribution. In reducing the majority class size to match the minority class, this technique, however, forgoes potentially useful information from the majority class.
2. **Hybrid-sampling:** Combination of under-sampling and over-sampling[9] methods, this technique applies under-sampling technique to the majority class and over-sampling technique to the minority class to balance the class distribution.

[8] The proportion of false alarms typically exceeds 99%. To avoid confusion, we reserve the use of the term "false positives", for reference to the model results.

[9] Over-sampling: This technique randomly duplicates observations from the minority class to match the majority class size. We refrain from employing this technique as it can be computationally expensive (in cases of severe class imbalance it may almost double the size of the dataset) and it often leads to overfitting the model.

3. **Synthetic-sampling:** This technique works like over-sampling, however, instead of randomly duplicating observations from the minority class, it introduces artificial noise to perturb its predictor values to avoid over-fitting. In our study, we use ROSE (Random Over-sampling Examples) synthetic-sampling method. This method utilizes the hybrid-sampling technique in addition to synthetic-sampling to overcome the computational challenges of a much larger data set.

5.2. Machine Learning Frameworks

In this subsection we discuss the machine learning algorithms, namely logistic regression, random forests, support vector machines, and gradient boosted machines, employed in our study to detect money-laundering at the financial institution.

5.2.1. Logistic Regression

Logistic regression (LR) models the probability of an observation belonging to a particular class. It employs the logistic function,

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}} \tag{1}$$

to model the probability of the categorical response variable, Y. In the above logistic function $X_1, X_2, \ldots, X_p$ are the p features. Simple manipulation of the above logistic function gives us,

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p} \tag{2}$$

and

$$ln(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{3}$$

which shows that the logit, $ln(p(X)/(1-p(X)))$, is a linear function of the features$X_1, X_2, \ldots, X_p$. We estimate the coefficients using the Maximum likelihood method. After the coefficient estimation, we select a suitable probability threshold to classify observations to the two distinct classes. Logistic regression is easy to implement and does not require making assumptions about the class distributions in the feature space. However, since it assumes a linear relationship between the logit and the features, this algorithm fails to more complex non-linear behavior - unless such a relationship is explicitly accounted for.[10]

5.2.2. Random Forest

A tree-based machine learning algorithm that in generating multiple decorrelated trees, Random Forest combines their corresponding predictions to arrive at a single prediction. The rationale for this algorithm consists in improving the prediction accuracy vis-à-vis the Decision Tree algorithm. When predictions of several decorrelated decision trees are combined, the resulting machine learning method in registering lower variance leads to better prediction accuracy. Although, the Random Forest achieves higher prediction accuracy than a single decision tree model, it does so at the expense of lower model interpretability. Below, we briefly discuss the decision tree algorithm and then examine the random forest algorithm.

[10] That is, our variables must be transformed accordingly in order capture the behavior we wish to model, e.g., a quadratic or logarithmic function.

Decision Trees

Decision Trees involve stratifying the feature space into non-overlapping regions. For a test observation that falls in a particular region R_i, the Decision Tree predicts the response value for the test observation to be the mean or mode (depending on whether the response variable is quantitative or qualitative) of the response values of the training observations in the region R_i.

Thus, the recursive binary splitting approach is adopted in constructing the non-overlapping regions of the feature space. This approach popularly known as the top-down greedy approach begins at the top and splits the feature space successively. A feature that results in the highest reduction in the residual sum of squares / classification error rate is considered for a split, at a given step in the tree building process. Each split creates two additional non-overlapping regions. To split one or both the resulting regions, the algorithm chooses the features that minimize residual sum of squares / classification error rate within the regions. This process of splitting ceases when the stopping criterion is met. This approach is called 'greedy' since the feature that minimizes the residual sum of squares / classification error rate the most at a given point privileges a readily available split candidate rather than opting for a feature that could result in a better decision tree in the long-term.

Once the decision tree is developed, for any given test observation, the algorithm first identifies the region to which the test observation belongs. It then assigns the mean/mode of the response values of the training observations belonging to the same region as the response value for the test observation. Although the Decision Tree algorithm is intuitive, unbiased (when grown sufficiently deep), and offers highly interpretable results, being prone to high variance its predictions are often unreliable. The Random Forest algorithm, an ensemble of particularly constructed decision trees, effectively overcomes this challenge. And we will focus on the Random Forest algorithm.

Random Forest

Random Forest (Breiman, 2001) algorithm in generating multiple decorrelated decision trees averages their predictions to yield a single prediction. Relying on the premise that averaging a set of independent observations having equal variances, this algorithm decreases the variance of the mean of the observations. The algorithm first generates a large number, say 'B,' bootstrapped samples from the training dataset. It then fits and trains the Decision Tree model on each of these B bootstrapped samples. The algorithm fits the decision trees on to the bootstrapped samples such that a random sample of 'm' features are considered as split candidates every time a split is made, rather than the entire set of features. Anytime a split is made, a fresh sample of random 'm' features are chosen for split consideration. Generally, 'm' is the square root of the total number of features. By drawing a fresh sample of 'm' features, the algorithm allows every feature to be considered for a split. This in turn produces uncorrelated decision trees which result in uncorrelated predictions. Averaging these uncorrelated predictions leads in a reduction of the variance of the ensemble method.

More rigorously, if $\hat{f}^1(x), \hat{f}^2(x), ..., \hat{f}^B(x)$, are the predictions that go with the B distinct decorrelated decision trees for the test observation x, then the Random Forest offers the prediction,

$$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \tag{4}$$

Note that the B decorrelated decision trees are grown deep and, therefore, register high variance and low bias. However, by averaging these decorrelated trees, the resulting Random Forest model achieves lower variance which improves its prediction accuracy.

5.2.3. Support Vector Machines

Support Vector Machine (SVM) is a machine learning algorithm predominantly applied to binary classification problems. Its approach builds on the Maximal Margin Classifier algorithm applied in classifying linearly separable observations. Since most datasets cannot separate the observations by a linear boundary, the Maximal Margin Classifier has limited applications. By introducing Soft Margin and Kernel concepts to the Maximal Margin Classifier, SVM can classify observations with non-linear decision boundaries. Soft Margin is a boundary that basically classifies the observations into two different classes, though it cannot be said to do this perfectly. It misclassifies a few observations for the sake of improving its classification for a majority of training observations and achieving better robustness to individual observations. Further, to account for non-linear decision boundaries, SVM enlarges the feature space efficiently using specific functions called Kernels that quantify the level of similarity between the two observations. In adopting appropriate Soft Margin and Kernel, the resulting SVM model achieves lower variance and accounts for non-linear decision boundaries.

Maximal Margin Classifier relies on the existence of a hyperplane.[11] If a hyperplane exists, then this could act as a classifier such that an observation belonging to one side of the hyperplane is classified as class 1; if the observation belongs to the other side, then it is classified as class 2. Thus, an observation X belongs to class 1 if, say, for example,

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p > 0 \tag{5}$$

And it belongs to class 2 if,

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p < 0 \tag{6}$$

Additionally, the magnitude $f(X)$ acts as a measure of confidence in the class assignment. If $f(X)$ is far from zero, then we can be confident about the class assignment. Whereas if f(X) is close to zero, then the class assignment may not be reliable.

Once we establish the existence of a hyperplane, then the Maximal Margin Classifier qualifies as the optimal hyperplane. Thus, it is the hyperplane that has the largest minimum distance from the training observations. We expect the optimal hyperplane to have the largest minimum distance from the training observations such that it can restore confidence in the class assignment of the observations. Once the Maximal Margin Classifier is located, the algorithm assigns a test observation to a class depending upon which side of Classifier it lies.

It so happens that the Maximal Margin Classifier depends only on a few training observations called the support vectors. Shifting a support vector or introducing a new observation that lies within the Margin of the optimal hyperplane could result in a new optimal hyperplane.

[11] A hyperplane is a linear boundary that separates a dataset's observations into two different classes. For instance, consider a two-dimensional feature space such that its observations could be separated by a linear boundary. In this case, the linear boundary, a hyperplane, is a line that divides the two-dimensional feature space into halves. Formally, $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$ is a hyperplane in a two-dimensional scenario where β_0,β_1, and β_2 are the parameters. The idea behind this notation could be extended to any arbitrary p-dimension feature space, where a hyperplane is defined as an affine subspace of dimension $p-1$. In other words, a hyperplane can be thought of as a flat subspace of dimension $p-1$ that divides the feature space into halves and follows the definition $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0$.

This suggests that the algorithm is prone to overfitting the training dataset. To sidestep overfitting, we misclassify a few training observations for achieving better robustness to individual observations and assigning most of the training observations to the correct classes. More tolerant to a few misclassifications, the new classifier is called the Soft Margin Classifier. The number of misclassified observations violating the optimal hyperplane is governed by a tuning parameter. Much like the Maximal Margin Classifier, the Soft Margin also depends solely on the support vectors. The optimization problem for the Soft Margin Classifier could be modified by including additional functions to its features so that it could classify observations that could only be separated by a non-linear boundary. However, including additional functions could render the algorithm computationally expensive. Therefore, to obtain a computationally feasible non-linear decision boundary, SVM introduces Kernels to the Soft Margin Classifier.

5.2.4. Gradient Boosted Models

A recently developed 'black-box' machine learning algorithm, Gradient Boosting Machine (GBM) has gained popularity for its high predictive accuracy. Being highly flexible, it could also be applied to a wide range of problems. GBM is an ensemble of weak predictive models where a weak model is defined as one whose prediction accuracy is only marginally better than random guessing. Any model can be a candidate for a weak model, however, for classification problems, such as ours, Classification Decision Trees are predominantly used (Kuhn, K. Johnson, et al., 2013). Our chosen weak predictive models, the Classification Decision Trees, are generally grown shallow with the number of splits ranging from 1-6. For our dataset, we note that GBM with 4 splits yield optimum results.

GBM was inspired by another boosting algorithm called AdaBoost, developed by Freund, Schapire, et al. (1996). In AdaBoost, a weak predictive model is fit to the weighted residuals of the ensemble created at the previous step so that the new weak predictive model could improve upon the errors made by the previous ensemble. In other words, a weak model is fit, in iteration, $i + 1$, to the residuals of the ensemble created in iteration i, such that the residuals corresponding to the incorrectly predicted observations by the ensemble are assigned higher weights compared to those predicted correctly. Assigning higher weights to the observations whose response values are difficult to predict, allows the new weak model to focus on improving the prediction accuracy for these observations, hence improving the overall prediction accuracy for the whole ensemble.

Much like AdaBoost, GBM algorithm consists in fitting weak predictive models sequentially to the ensemble such that their inclusion improves the predictive performance of the whole ensemble. The weak predictive models are constructed such that these models and the negative gradient of the loss function associated with the whole ensemble are maximally correlated (Friedman, 2001). Below, we outline the GBM methodology.

Consider a training dataset $(\mathbf{x}_i, y_i)_{i=1}^N$ where $\mathbf{x}$ denotes the explanatory variables and y denotes the response variable such that the true relationship between $\mathbf{x}$ and y is given by f. We estimate a model $\hat{f}(\mathbf{x})$ such that it minimizes the expected value of the loss function[12] $L(y, f(\mathbf{x}))$,

[12] Since our response variable is binary, we consider the binomial loss function.

$$\hat{f}(\mathbf{x}) = y \tag{7}$$

$$\hat{f}(\mathbf{x}) = argmin_{f(\mathbf{x})} E_{\mathbf{x}}[E_y(L(y, f(\mathbf{x})))|\mathbf{x}]$$

In restricting the search for the estimated model to the family of parametric functions, we consider the following "additive" expansion for the true function (in the equation below, M is the number of iterations),

$$f(\mathbf{x}; \{\beta_m, \mathbf{a}_m\}_1^M) = \sum_{m=1}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{8}$$

In the above function, $h(\mathbf{x}; \mathbf{a})$ is a parameterized function of the explanatory variables $\mathbf{x}$, characterized by the parameters $\mathbf{a} = \{a_1, a_2, \dots\}$. In our case, $h(\mathbf{x}; \mathbf{a}_m)$ is a shallow classification tree and therefore the parameters $\mathbf{a}_m$ are the split variables, split locations, and the modes of the terminal node for the individual trees.

By choosing a parameterized model $f(\mathbf{x}; \mathbf{P})$, where $\mathbf{P} = \{P_1, P_2, ...\}$ is a finite set of parameters, the function optimization problem changes to the following parameter optimization problem,

$$\mathbf{P}^* = argmin_{\mathbf{P}} \Phi(\mathbf{P}) \tag{9}$$

where

$$\Phi(\mathbf{P}) = E_{y,\mathbf{x}} L(y, f(\mathbf{x}; \mathbf{P})) \tag{10}$$

We, therefore, get

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}; \mathbf{P}^*) \tag{11}$$

Applying numerical optimization methods to solve for $\mathbf{P}^*$ imposes the solution for the parameters as $\mathbf{P}^* = \sum_{m=0}^{M} \mathbf{p}_m$. In this solution for $\mathbf{P}^*$, $\mathbf{p}_0$ and $\{\mathbf{p}_m\}_1^M$ are the initial guess and the successive increments ("boosts"), respectively. Each "boost" depends on the sequence of preceding "boosts" and to solve the optimization problem, the algorithm chooses Steepest-descent numerical minimization method. In defining the increments $\{\mathbf{p}_m\}_1^M$, first the gradient, $\mathbf{g}_m$, is computed,

$$\mathbf{g}_m = \{g_{jm}\} = \{[\frac{\partial \Phi(\mathbf{P})}{\partial P_j}]_{\mathbf{P}=\mathbf{P}_{m-1}}\} \tag{12}$$

where $\mathbf{P}_{m-1} = \sum_{i=0}^{m-1} \mathbf{p}_i$. The increment is then defined as $\mathbf{p}_m = -\rho_m \mathbf{g}_m$, where,

$$\rho_m = argmin_{\rho} \Phi(\mathbf{P}_{m-1} - \rho \mathbf{g}_m) \tag{13}$$

In the above notation, $-\mathbf{g}_m$ is the direction of "steepest-descent" and ρ_m is the "line search" along this direction.

Contrarily, we can also apply numerical optimization in the function space. In other words, we treat $f(x)$ as a parameter and minimize $\Phi(f) = E_{y,\mathbf{x}} L(y, f(\mathbf{x})) = E_{\mathbf{x}}[E_y(L(y, f(\mathbf{x})))\mathbf{x}]$. We consider the solution to have the following functional form,

$$\hat{f}(\mathbf{x}) = \sum_{m=0}^{M} f_m^*(\mathbf{x}) \tag{14}$$

where $f_0^*(\mathbf{x})$ and $\{f_m^*(\mathbf{x})\}_1^M$are the initial guess and increment functions ("boosts") defined by the optimization, respectively. Each "boost" is updated as follows,

$$f_m^*(\mathbf{x}) = -\rho_m g_m(\mathbf{x}) \tag{15}$$

where

$$g_m(\mathbf{x}) = [\frac{\partial \phi(f(\mathbf{x}))}{\partial f(\mathbf{x})}]_{f(\mathbf{x})=f_{m-1}(\mathbf{x})} = [\frac{\partial E_y[L(y, f(\mathbf{x}))|\mathbf{x}]}{\partial f(\mathbf{x})}]_{f(\mathbf{x})=f_{m-1}(\mathbf{x})} \tag{16}$$

is the gradient[13] and

$$\rho_m = argmin_\rho E_{y,\mathbf{x}} L(y, f_{m-1}(\mathbf{x}) - \rho g_m(\mathbf{x})) \tag{17}$$

is the "line search" along the direction of $-g_m$. This non-parametric approach can no longer be applied when the joint distribution of $(\mathbf{x}, y)$ is estimated by the finite sample $(\mathbf{x}_i, y_i)_{i=1}^N$. To sidestep this, we can consider a parameterized form, as assumed in case of the parametric method discussed, thereby converting the optimization problem to a parametric optimization problem,

$$(\beta_m, \mathbf{a}_m)_1^M = argmin_{\{\beta_m^{'} \mathbf{a}_m^{'}\}_1^M} \sum_{i=1}^N L(y_i, \sum_{m=1}^M \beta_m^{'} h(\mathbf{x}_i; \mathbf{a}_m^{'})) \tag{18}$$

If the given approach also fails, then the "greedy stagewise" can be adopted as follows,

$$(\beta_m, \mathbf{a}_m) = argmin_{\beta,\mathbf{a}} \sum_{i=1}^N L(y_i, f_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a})) \qquad For \quad m = 1, 2, \ldots, M \tag{19}$$

And the ensemble is updated as follows,

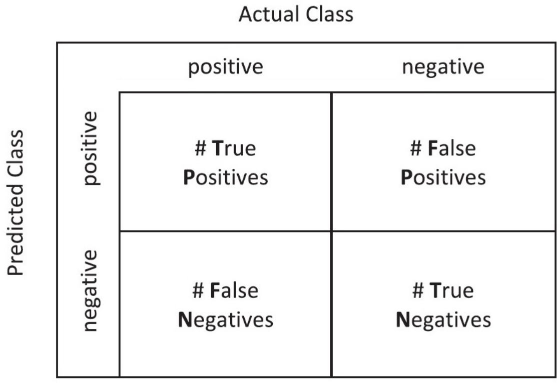
$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{20}$$

Thus, the choice of the loss function and weak predictive models determine the model properties of GBM. However, these choices in providing the algorithm with high flexibility render their applicability to a wide range of problems.

5.3. Model Evaluation

We now discuss the performance metrics used to evaluate our models. For evaluating the out-of-sample predictions, the data sample is split into training and test samples. The models are trained on the training sample and its predictive performance is estimated on the test sample via its confusion matrix (Figure 1). A confusion matrix tabulates a model's class predictions against the actual class assignment of the observations. We label the entries of the confusion matrix as follows:

- **TP:** the number of **true positives**, i.e. positive class observations that the model has correctly classified.
- **TN:** the number of **true negatives**, i.e. negative class observations that the model has correctly classified.
- **FN:** the number of **false negatives**, i.e. positive class observations that the model has incorrectly classified.
- **FP:** the number of **false positives**, i.e. negative class observations that the model has incorrectly classified.



		Actual Class	
		positive	negative
Predicted Class	positive	# **T**rue **P**ositives	# **F**alse **P**ositives
	negative	# **F**alse **N**egatives	# **T**rue **N**egatives

Figure 1: Confusion Matrix

Note that the total number of observations (N) in our sample must be the sum of these four quantities, i.e.,

$$N = TP + TN + FN + FP \tag{21}$$

We now define our metrics, true positive rate (TPR) and false positive rate (FPR), with reference to the confusion matrix.

TPR, also referred as sensitivity and recall, measures the proportion of positive observations correctly classified by a model:

$$TPR = \frac{TP}{(TP + FN)} \tag{22}$$

FPR, or fall-out, measures the proportion of negative observations misclassified by a model:

$$FPR = \frac{FP}{(FP + TN)} \tag{23}$$

Both TPR and FPR lie between 0 and 1. Typically, we want TPR to be as high as possible and FPR to be as low as possible. Unfortunately, these two metrics do not vary independently of each other, unless we are dealing with a perfect model . To achieve high TPR we require a more sensitive model which comes at the cost of higher false positives, i.e., higher FPR. This trade-off is a general feature of any classification model.

Most ML classification algorithms estimate the probability of an observation belonging to the positive class. Typically, a value of 0.5 is used as the probability threshold, i.e., an observation whose estimated probability is greater than the threshold is assigned the positive class; whereas, if the estimated probability is less than the threshold, it is assigned the negative class. Lowering the threshold increases the number of true positives, however, it also increases the number of false positives; whereas raising the threshold lowers the number of false positives, however, at the expense of reducing number of true positives. Therefore, to measure the overall performance of a model, we plot the receiver operator characteristic (ROC) curve. ROC curve is the graphical representation of the relationship between the true positive rate and false positive rate when the probability threshold is varied.

[13] $\phi(f(\mathbf{x})) = E_y[L(y, f(\mathbf{x}))|\mathbf{x}]$ and $f_{m-1}(\mathbf{x}) = \sum_{i=0}^{m-1} f_i^*(\mathbf{x})$

Figure 2 shows a typical ROC curve for a classification model. Each point on the curve provides the TPR (y-coordinate) and FPR (x-coordinate) corresponding to a probability threshold. Ideally, a model with TPR equal to 1 and FPR equal to 0 yields the best predictive capacity. However, in practice, we choose a model that hugs the top left corner of the ROC curve. Additionally, to measure the model's out-of-sample predictive performance we compute the area under the ROC curve (AUC). AUC lies between 0 and 1. A model with AUC of 0.5 is no better than randomly guessing (random classifier) the class for an observation; a model with AUC less than 0.5 performs worse than the random classifier; and a model with AUC greater than 0.5 demonstrates predictive capacity.

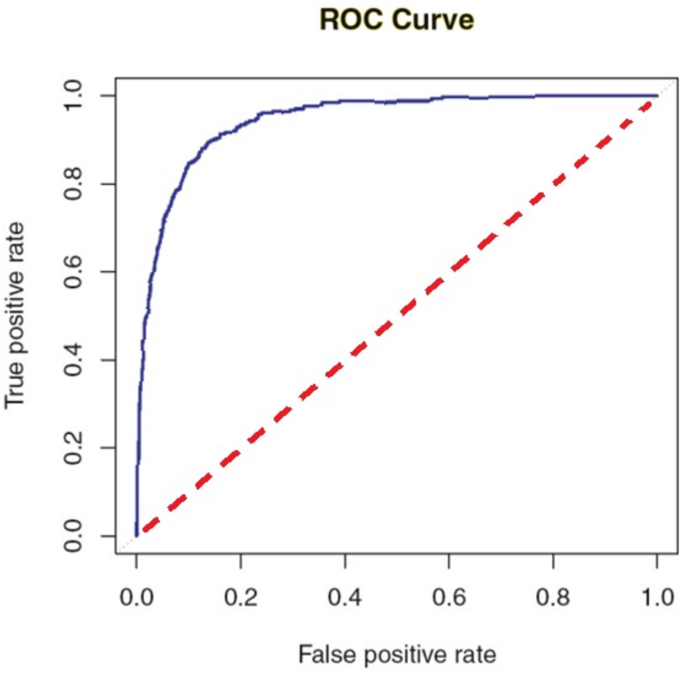


Figure 2: Confusion Matrix

5.4. Predictor Importance

Finally, we investigate the relative importance of features in determining whether a transaction is fraudulent. In case of logistic regression, we use the statistical significance and the magnitude of coefficient estimates to infer the relative importance of features. For random forests and gradient boosted machines, we estimate the total decrease in node purity corresponding to each predictor. The SVM algorithm does not naturally extend itself towards estimating feature contribution, however, a heuristic can be constructed. This method, unfortunately, does not provide consistent and reliable estimates. Therefore, we do not compute feature importance for the SVM model. We choose the models estimated on hybrid-sampled dataset to compute feature importance since these models outperform the models fitted on datasets resampled by other techniques employed in this study.

6. Results

This section presents our baseline empirical results. It then presents results of robustness tests. We discuss the baseline results in subsection 6.1. The results of the robustness tests are discussed in subsections 6.2 and 6.3.

6.1. Model Performance and Interpretation

We first determine whether the various country-level features employed in our study can detect money-laundering at the financial institution. To meaningfully gauge the predictive capacity of these features, we first decompose our dataset into transactions involving private customers (people-related) and corporate clients (corporate-related).[14] We then train our models on the country-level attributes of the people-related, corporate-related, and combined dataset to estimate the out-of-sample performance of our models. We train 48 models; 4 machines learning algorithms trained on 3 datasets (people-related, corporate-related and the combined dataset) that are balanced by 4 balancing techniques. A randomized 50:50 split is performed on the datasets to create training and test datasets.[15] We further perform cross-validation to test the validity of our models and estimate relative importance of various country-level features.

6.1.1. Predictive capacity of Country-level features

Table 3 shows the TPR, FPR, and AUC results of the models trained on the country-level features of the three datasets. Our features include the Hofstede country-specific culture dimensions and two institution quality indices for the customer's country of residence and origin/destination country of the wire.[16] These features are: CPI_R, CPI_W, FSI_R, FSI_W, IDV_R, IDV_W, MAS_R, MAS_W, PDI_R, PDI_W, UAI_R, and UAI_W. For models trained on the combined dataset, we note that the AUCs are in the 0.70-0.80 range. This demonstrates that our models can discern between suspicious and legitimate transactions. We find that our models can discern better for the corporate-related dataset with AUCs as high as 0.88. We further find evidence for predictive capacity for the country-level features for the people-related data, however, compared to the combined and corporate-related data, these results are modest with AUCs in the 0.65-0.72 range.[17] We further note that all the models trained on datasets balanced by the hybrid-sampling technique consistently provide significant out-of-sample performance. Additionally, we find that the RF and GBM models have the best out-of-sample performance for all the three datasets balanced by the under- and hybrid-sampling techniques.

[Please insert Table 3 about here.]

[14] See Table 1 and Table A (Internet Appendices).

[15] Except in the case of cross validation, where 80:20 and 90:10 splits are performed.

[16] Please see Table 2 for concise definitions.

[17] We further train our models on the Hofstede country-specific culture dimensions, excluding the institution quality indices. We report the models including only the national culture dimensions are comparable to models including the institution quality indices as well. Please see Tables B1-B3 of the Internet Appendix B. These may reflect that national culture and national governance are endogenously related. However, our objective is to determine whether national culture is an effective predictor and whether there are important concomitant ethical concerns. We do not aim to identify a causal relationship between national culture traits and malfeasance in banks.

6.1.2. Determining the validity of our models using cross-validation techniques

To determine the validity of our models we perform K-fold cross-validations. K-fold cross-validation estimates how well a model generalizes to an independent dataset by dividing the dataset into K equal parts, using one part as a hold-out test set, and training the model on the remaining K-1 parts. This is then repeated K times, such that each of the K equal parts is considered for a test dataset. The out-of-sample model performance is then computed as the average of the K results. We perform 5- and 10-fold cross-validations, that is for each of the K instances, we train our models on 80% and 90% of the datasets, balanced by the hybrid-sampling method, respectively.[18] We then estimate the out-of-sample performance on the remaining 20% and 10% of the datasets. In Table 4 we report the AUC metric, estimated by the cross-validation technique, to measure performance for all the models. The results demonstrate that the predictive capacity of country-level variables remain similar to that reported in Table 3. The low standard deviation (σ) further attests to the reliability of our models.

[Please insert Table 4 about here.]

6.1.3. Investigating the relative importance of Country-level features in detecting money-laundering

Table 5 presents the relative importance of our country-level features for the models trained on the three datasets.[19] We find that for both corporate-related and combined alerts, the individuality rating of both the customer's residence country (IDV_R) and country of wire origination/destination (IDV_W) are of paramount importance. This is followed by the corruption perception score of the country of wire origination/destination (CPI_W) and the customer's residence country (CPI_R) for the corporate-related alerts; and (CPI_W) and the financial secrecy score of the customer's resident country (FSI_R) for the combined alerts. For people-related alerts, the corruption perception score for the country of wire origination/destination (CPI_W) and the financial secrecy score of the resident country (FSI_R) are the two most important features, followed by the CPI_R and IDV_R.

[Please insert Table 5 about here.]

6.2. Can we improve the predictive capacity of our models by enlarging the feature space?

In this section, we extend our feature space to include account- and transaction-level variables. We further include the proprietorial risk score (PROP) in our enlarged feature space to assess the predictive capacity of our models.[20]

[18] We train our models on the three datasets balanced by the hybrid-sampling method since this method results in models with high predictive accuracy across all the three datasets.

[19] We do not report feature importance results for the SVM model since there does not exist a reliable model-specific feature importance method for SVM algorithm.

[20] All features are defined in Table 2.

6.2.1. Predictive capacity of country-, account-, and transaction-level features in detecting money-laundering

We further extend our feature space to include customers' account- and transaction-level information to determine whether we could improve the predictive capacity of our models. This extends our feature space to include 24 features with 12 country-level features, 4 account-level features, and 8 transaction-level features.[21]

Table 6 presents the TPR, FPR, and AUC scores for models trained on the enlarged feature space. These models enhance the predictive capacity across all the models reported in Table 3, with AUCs ranging between 0.72-0.91, 0.83-0.94, and 0.60-0.85, on the combined, corporate-related, and people-related datasets, respectively. We further note that the models trained on the datasets balanced by the hybrid technique are better able to discern between a fraudulent and non-fraudulent transaction with AUC scores between 0.75-0.91, 0.85-0.94, and 0.71-0.85 for the combined, corporate-, and people-related datasets, respectively. We report a significant increase in the predictive capacity of our models across all the three datasets. We again find that the RF and GBM models with under- and hybrid-sampling are the optimal models.

[Please insert Table 6 about here.]

6.2.2. Predictive capacity of country-, account-, and transaction-level features along with the proprietorial risk score in detecting money-laundering

Finally, we include the proprietorial risk score, PROP Score, to our enlarged feature space to determine whether its inclusion markedly enhances the predictive capacity of the models reported in Table 6. We report the out-of-sample performance of these models in Table 7. Interestingly, we find only a slight improvement, of approximately 1-2% on average, in performance. This indicates that models with the country-, account- and transaction-level information provide useful predictive power.

[Please insert Table 7 about here.]

6.3. Does national culture traits remain useful in the extended dataset?

In this section, we investigate whether the country-specific culture and institution quality indices pertaining to customer's residence country and the country of wire origination/destination remain useful in detecting money-laundering in the enlarged feature space.

6.3.1. Does national culture traits remain useful in comparison with account-level and transaction-level features?

We estimate feature importance for models reported in Table 6 to determine whether country-level features of the customers provide useful predictive capacity in detecting fraudulent

[21] We include an additional feature, Customer Age, in the people-related models which extends the feature set to include 25 predictors.

wire transactions in the enlarged feature space. We present these results in Table 8. We note that for corporate-related alerts, the county-level features that rank among the top five features are the individuality rating of the customer's country of residence (IDV_R), individuality rating of the country of wire origination/destination (IDV_W), and the uncertainty avoidance cultural trait of the customer's residence country (UAI_R). We further find that the power-distance index score of the customer's residence country (PDI_R) informs the customer's predilections for committing financial misconduct. For people-related alerts, the individuality score of the customer's residence country (IDV_R), corruption perception score of the country of wire origination/destination (CPI_W), and financial secrecy score of the customer's country of residence (FSI_R) are the most important county-level features that rank among the top ten features. These results provide evidence of the usefulness of culture traits of customers for detecting both corporate and individual malfeasance, however, the country-level features are more pronounced in detecting corporate malfeasance than individual malfeasance.[22] For the combined alerts, we note that (IDV_R), (IDV_W), and (FSI_R) rank among the top ten features; further providing evidence of the usefulness of the country-level features in detecting malfeasance.

[Please insert Table 8 about here.]

6.3.2. Does national culture traits remain useful in comparison with a proprietorial risk score along with account- and transaction-level features?

Table 9 reports the feature importance for models reported in Table 7. For corporate-related alerts, we again find that the individuality scores of both the country of the wire origination/destination (IDV_W) and customer's resident country (IDV_R) are important country-level features. These features also rank among the top five features influencing a customer's predilections for committing financial misconduct.We further note that the corruption perception score of the country of wire origination/destination (CPI_W) and power-distance index score of the customer's residence country (PDI_R) are among the top ten features. Interestingly, we find that IDV_W, IDV_R, and CPI_W have higher predictive capacity than the proprietorial risk score. However, in case of people-related alerts, the PROP score is the most important feature. This suggests that the proprietary algorithm, used by the financial institution, is more effective in detecting fraudulent transactions pertaining to individual accounts than for corporate accounts. Further, in case of people-related alerts, the financial secrecy score of the customer's residence country (FSI_R), corruption perception score of the customer's residence country (CPI_R), and corruption perception score of the country of wire origination/destination (CPI_W) rank among the top ten features in detecting money-laundering in our models. For the combined alerts, the features that influence the models in decreasing order are IDV_W, $PROPscore$, IDV_R, and FSI_R. These features also rank among the top ten features. In addition to results reported in Table 8, these results further underline the usefulness of adopting country-specific features to complement current account and transaction variables for AML monitoring.

[22] In the Internet Appendices, Tables C1-C3, we also report results using the Schwarz cultural variables in lieu of the Hofstede variables in our models. The Schwarz cultural model uses 3 dimensions to measure culture, see Internet Appendix C for further details.

[Please insert Table 9 about here.]

7. Discussion and Ethical Framework

At least since Donaldson and Dunfee (1994), scholars have acknowledged that business ethics research while informed by empirical ideas, can also be informed by normative concepts, by prescriptive ideas which, although not necessarily approachable by empirical analysis, suggest what societies should do. Indeed, they indicate that empirical analysis is often not the appropriate tool to determine what societies "ought" to do (see also Sorley (1885)). As a result, the attainment of money laundering out-of-sample predictive accuracy alone, as established in this paper, is an inadequate reason for the deployment of a machine learning alert model.

The potential for unethical repercussions related to AI applications, especially those which inform decisions to impact people, is immense. Examples include recruitment, promotion, flight risk and cessation of employment algorithms as well as credit extension, insurance risk scoring and dynamic pricing algorithms, among many others. Fraud detection, informed with machine learning, arguably falls on the lower end of the spectrum of potentially unethical AI – after all its aim is to mitigate financial malfeasance. Nevertheless, it is critically important to consider the societal ramifications of using national background as a prompt for further scrutiny of individuals.

With a view, hence, to "giving voice to values" (Arce and Gentile, 2015), we seek to identify the ethical considerations of incorporating profiling, whether intentional or not, within machine learning algorithms. We discuss ethical issues pertinent to the use of national culture in machine learning in general and money-laundering alert models in particular.

We frame our discussion around a number of ethical discussion: 1) Do public good concerns in countering money laundering outweigh 'collective treatment' concerns in national profiling in algorithms? 2) Do those producing the alerts have permission to use the personal data? 3) Who is responsible for the design of an algorithm? 4) Are algorithms accountable? 5) Are the algorithms used for detection, or, alternatively, for prediction?—and are there subtle distinctions regarding this?; 6) Are alert models reflective of global, national or sub-national; public or private regulation?; 7) Do the algorithms in use exacerbate tangential societal biases? and 8) Can the deployment of an algorithm, due to automation, transform the workplace?

7.1. Do public good concerns in countering money laundering outweigh 'collective treatment' concerns in national profiling in algorithms?

Alter and Darley (2009) define 'collective treatment' as the act of behaving toward more than one individual uniformly. Collective treatment is distinguished from individualized treatment, in which individuals are treated differently from one another according to relevant criteria. An example is punishing a gang for being offenders as opposed to prosecuting individuals according to their relative contributions to a crime. As noted by Alter and Darley (2009), collective treatment relies on people who share salient features being treated as interchangeable members of a group defined by those features. Of course, as noted Brewer and Harasty (1996); Campbell (1958); Dasgupta, Banaji, and Abelson (1999) and many others, such salient features can include race, ethnicity, socioeconomic status, religion, physical appearance, relative income, and whether and individual has a disability. Clearly, we can add to these ways of grouping individuals' national culture, especially as pertaining to alert models designed to detect fraud. However, while one of the prominent dangers of collective treatment is that it can be administered by individuals in authority to reward, punish, or restrict the rights of a group within a population. For instance, a judge who sentences a gang of criminals rather than individuals etc. One advantage of machine-learning-based alert models is that it avoids the situation of individuals choosing to impose or not to impose collective treatment in arbitrary circumstances.

7.2. Do those producing the alerts have permission to use the personal data?

Simply because it is legal to gather and mine certain data does not make it ethical. 'Ethics' regards sets of moral codes beyond legally required minimums. With regard to the mining of data, inherent with machine learning procedures, questions of an ethical nature invariably arise.

Whether the institution conducting machine learning is allowed to use the data incorporated into its algorithms is both an important legal issue, as well as an ethical issue. There may be legal barriers to using particular data. But there are ethical issues that extend beyond simple legality. In many cases, machine learning can employ data that they might not have proper permissions to use. As noted by Adomavicius and Tuzhilin (2001), data mining with regard to individuals has been seen as either 'factual', who the customer is, or 'transactional', what the customer has done or is doing (see also Cook, 2008). Adomavicius and Tuzhilin (2001) suggest the latter is more commonly used for criminal identification; as well as more commonly objected to as an intrusion into individual privacy. However, in the money laundering alert model highlighted in this paper, we suggest that simply using primarily data about who the customer is, i.e., the customer's home country, is sufficient to generate area-under-the-curve predictions that are almost 90 percent. So, one potential advantage of using national culture as a predictor is to avoid more intrusive gathering of customer behavior. Use of national culture avoids issues of using personal data. The sweeping aggregate generality of national culture avoiding invasive use, likely without permission, of individual characteristics. In this regard, with respect to including national culture in machine-learning models, a relevant question is, if not national factors, then what level of factors? And what would be the alternative set of implications? Overall, with regard to permissions to use data, national culture, while arguable a rough profiling of people from respective nations, avoids the use more individual and likely personal data. In summary though, an ethical minefield is established between use of sweeping generalities, and the potential invasion of privacy by collecting personal, often transactional, data.

7.3. Who is responsible for algorithmic design?

In this vein (Martin, 2019), while making a strong case for developers of algorithms having responsibility for how they are used, lists a number of examples from news sources highlighting disturbing ways algorithms can be used. These include predicting whether you are a terrorist, what you will pay for an online product that is presented with bespoke pricing to individual online users, whether you will receive a loan, or if an incarcerated inmate will receive parole, among many examples. We also discuss and highlight further in this paper, a number of contexts in which 'profiling' has been identified as being unfair and detrimental to the social fabric.

It is worth considering the question of whether authors of academic studies are also responsible for the use of presented ideas and findings. This seems too much of an extension, to consider academic researchers responsible for algorithms subsequently developed in part because of their findings— and a point of view that would certainly inhibit scholarly investigation. However, we as authors are concerned that the evidence presented in this paper, that national factors, particularly national culture, may be particularly efficacious in money laundering alert models, and may have various moral consequences. As noted by Martin (2019), algorithms are inherently value laden and need to be constructed to preserve stakeholder's "rights and dignity."

7.4. Are algorithms accountable?

Another concern with use of machine learning is algorithmic accountability, as noted previously (e.g., Buhmann, Paßmann, and Fieseler, 2019; Martin, 2019; Seele et al., 2019) . This includes concern over how algorithms are established, in terms of what hypotheses, either explicitly or implicitly are formed. However, from an alternative perspective, algorithm transparency also provides criminals with insight into the factors used in respective algorithms, facilitating subsequent avenues of evasion. However, with respect to national culture, how well could knowledge of the inclusion of national culture in an alert algorithm be gamed by would-be illicit actors? Would this encourage actors to channel banking transactions through other countries with differing identified cultural characteristics? Do clever money launderers already know that culture is being used to help establish money-laundering alerts? Overall, it may be that being transparent about the use of national culture in machine-learning algorithms is less inherently exploitable than transparency about other details of algorithms.

7.5. Are the algorithms used for detection, or, alternatively, for prediction?—and are there subtle distinctions regarding this?

Another fundamental distinction, that touches on the ethics of algorithms in the use of machine-learning, is whether alert procedures are to be used in detection or, alternatively, in prediction. The also involves broader issues of the implications of ex ante or ex post investigation. Depending on the context, prediction can lead, or not lead, to particular consequential actions. For instance, an algorithm to predict personal loan default might lead a person being denied financing (Fuster et al., 2018). On the other hand, a prediction algorithm to identify possible bank fraud might lead to time and resources being devoted to simply closer scrutiny. In this regard, using machine learning to detection money-laundering, as in the example of this paper, could be viewed, as simply reducing the costs of detection, rather than establishing an unfair barrier to banking inclusion.

In contrast, using machine learning for the purpose of prediction of what might take place creates identifiable issues of fairness. For instance, there is the highly controversial practice in the US of using ethnic and racial profiling in prediction of whether prison inmates under consideration for parole will recidivate (Hartney, 2009). This has even been extended to machine learning algorithms (Berk, 2017; Lee, 2018). Examples such as this display an obvious unfairness and social injustice. This is inherently different than the context of identifying whether money-laundering has already taken place. Or is it so different? Certainly, there is the possibility of organizations transferring usage of algorithms from detection to pre-emption. In which case, factors included in detection are now used to unfairly exclude. Further, identifying persons from particular countries in the context of global regulation appears at least somewhat differently than law enforcement in a particular country or sub-national component of a country targeting certain citizens based on demographic characteristics for additional scrutiny. Or is a case of global regulation being that different? Certainly, this issue beckons much further reflection.

Of additional concern, the distinction between detection and prediction becomes blurred for situations where 'everyone' is doing a particular illegal action. For instance, it is not uncommon on many of the interstate highways in the US, where speed cameras are generally not used as widely as in other countries, for the great majority of drivers to be driving over the speed limit. However, it is generally regarded that in the US African American drivers are much more likely to be pulled over by the police (Harris, 1996). This could be due to racial prejudice of course. But a commonly put forth reasoning for this is that speeding African American drivers are also greater opportunities for law enforcement to discover other violations because of a higher percentage of African Americans having criminal records. This obviously fuels a self-fulfilling prophecy. Another example is the controversial practice of the US Internal Revenue Service (IRS) in US paying closer scrutiny to vocal anti-tax groups when these groups apply for tax-exempt status. On the one hand, this practice could be viewed as the targeting political opposition to the government. On the other hand, is it not reasonable to consider vocal anti-tax groups as more likely to evade taxes? Pulling over African Americans for traffic violations has obvious aspects of intimidation and social repression. Greater scrutiny of the taxes of groups with a particular political orientation presents similar concerns.

Global regulation that focuses more on some countries than others seems is arguably a different context than inflection of legal authority unevenly within a particular country or sub-national jurisdiction. Perhaps because global monitoring, as with money-laundering alerts, is often in the realm of the private sector. Consequently, much of the social unfairness of the public sector isolating groups within a society is avoided. On the other hand, if we consider the world as an increasingly globalized society then such distinctions lessen. It is arguable that private global firms do have a governance role and that such a role needs to be evenly administered. An interesting parallel is the openly disclosed pillar of the microfinance industry to focus on women borrowers. In other words, to be less inclined to grant loans to men (Aggarwal, J. W. Goodell, and Selleck, 2015). This is put forward as micro-finance to women, as opposed to men, will offer greatly societal outreach benefits, and that women are more reliable to pay back micro-finance obligations. Perhaps identifying a particular gender as more likely to repay a loan is not fundamentally different from identifying people of a particular national cultures as more likely to repay loans—or, the subject of this study, more or less likely to conduct money laundering.

Issues of social fairness become much more glaring however, when we consider the wide variety of research that seeks to model national at a sub-national level with demographic, particularly religion data. For instance, Baxamusa and Jalal (2014) model religion as indicating levels of what national culture would describe as uncertainty avoidance. A significant problem with using national culture in algorithms is that there is the potential for biases about particular national cultures to diffuse to sub-national levels.

7.6. Do the algorithms in use exacerbate tangential societal biases?

Another concern is whether the respective machine-learning algorithm is incorporating factors that tangentially engender implicit biases? An example of this is provided by Williams, Nathanson, and Paulhus (2010), who highlight a case of low verbal skills being identified for greater likelihood of scholastic cheating. Clearly, this identification can foster attitudes of bias against immigrant communities or others with generally lower skills in the given language of instruction. Or can establish bias against those with speech impediments for instance. Do algorithms that incorporate national culture foster prejudice? The study we highlight in this paper for instance suggests people from more individualist countries are more likely to engage in money laundering. Consider the roles of profiling by national culture in other contexts. For instance, profiling potential CEOs or board members as to whether they would be advocates of CSR, or whether they have optimal demographic characteristics (S. G. Johnson, Schnatterly, and Hill, 2013) Would certain CEO candidates be disfavored or disqualified because of their country of origin or ethnic background?

7.7. Can the deployment of an algorithm, due to automation, transform the workplace?

????

Clearly, the use of national culture in bank alert models provides some predictive accuracy, while it at the same time raises a number of important social and ethical issues. We hope this paper invites further analysis and discussion of this important issue.

8. Conclusions

In light of the recent scandals involving major international banks such as Danske Bank and Swedbank, we see greater focus placed upon AML policy and a need for innovation in the current protocols than monitor and detect suspicious activity that may be an indicator of money laundering or fraud. A growing field of research looks at the cultural and behavioral aspects that govern decisions at the institutional level especially in the corporate and financial domain. Our paper uses this research to inform and improve current practice in AML policy for financial institutions.

Using our data set of over 200,000 international wire transactions collected over a ten year period, we build machine learning models that reference the levels of corruption and financial secrecy in a country, as well as the cultural measures of individualism, masculinity, power-distance and uncertainty avoidance. We find that, on top of the industry standard account- and transaction-level variables, these country-level variables greatly improve our models predictive power, particularly in the category of corporate accounts. Using the machine learning algorithms to estimate the relative importance of the predictors in the most successful models, we find that individualism scores for the customer's resident country, as well as the individualism score for the wire's country of origin/destination, are by far the most important of the country-level variables and indeed all the variables outright, for the models involving corporate accounts. As for the personal account models, the corruption perception score for the wire's country of origin/destination and the financial secrecy score for the customer's resident country prove to be the most important country-level variables, with other predictors outside of country-level variables proving important overall in predicting the incidence of suspicious wire activity. Overall, however, our results suggest that country-level data, particularly national culture scores of either the sender or receiver of wire transfers, either alone, or in combination with measures of control of corruption and financial secrecy, provide highly effective prediction modelings. Given the societal implications long identified regarding 'collective treatment,' or results provoke considerable reflection on the ethical concerns of using country-level variables by financial institutions to form money laundering alert models.

Our findings indicate the importance of cultural and behavioral measures when considering the potential for money laundering and fraud in international money movement, especially when it comes to corporate activity and provide strongly predictive models for capturing such behaviour. Furthermore, the models applied to the segregated data sample (corporate account vs. individual account) demonstrate rather distinct differences in terms of predictive performance as well as feature importance. Practitioners can benefit from making careful configurations regarding sample segmentation as well as feature selection. Applying a more contextual brush to current AML surveillance practices may prove a valuable resource in the fight against money laundering and fraud worldwide.

In light of our findings, we examine the ethical issues of incorporating country factors, especially national culture, in machine learning applications, illustrating the potency of such factors to inform bank alert models. The use of machine-learning algorithms is expanding rapidly across the globe in a host of assessment and prediction contexts. Catalyzed by recent events, the ethical implications of profiling are now of great interest. The ethical implications of ascribing values, against a global standard, to national culture qualities, long done in the literature, needs further consideration, especially now, given their likely future inclusion in machine learning applications. Examining detection of money laundering at a globally important financial institution, we avail of binary classifier type alert models, together with corrections for data imbalance, to show the surprising utility of national culture in formulating anti-money laundering predictions. For corporate (individual) accounts, Hofstede Individuality (Individuality, and national-level corruption perception and financial secrecy) scores of the country in which a customer is resident, or from which a wire is sent/received, are the most important factors. National culture alone provides a high degree of predictive power. And when combined with extensive account and transaction data; as well as even proprietary institutional algorithms already in use, its inclusion greatly enhances predictive ability. While discussing these results, we offer a framework for considering the ethical implications of incorporating profiling information into predictive machine learning models. We frame our discussion around a number of important distinctions: 1) Do those conducting alerts have permission to use data? 2) Are algorithms transparent? 3) Are the algorithms used for detection or alternatively for prediction?—and are there subtle distinctions regarding this?; 4) Are alert models reflective of global, national or subnational; public or private regulation? And 5) Do the algorithms in use encourage tangential societal biases?

We conclude that inclusion of national culture in machine-learning algorithms both avoids some common ethical shortcomings; as well as invites ethical concerns. In our discussion we consider that context matters, offering an outline of when collective treatment by financial institutions may result in greater ethical costs. The use of national culture in machine learning algorithms can serve a global public good by nature of its efficacy, but there is also a cost to the global public good because of a broad set of ethical concerns regarding collective treatments.

Tables

Table 1: Data Cross-section and Sample Selection

Panel A: Alerts and Issue Cases by Year

	Combined		**Corporate**		**People**	
Year	**#Alerts**	**#Issues**	**#Alerts**	**#Issues**	**#Alerts**	**#Issues**
2009	22,183	878	5,752	448	16,431	430
2010	23,154	485	6,643	215	16,511	270
2011	20,335	216	6,193	68	14,142	148
2012	18,572	143	5,298	29	13,274	114
2013	21,088	205	5,984	71	15,104	134
2014	11,098	87	2,617	41	8,481	46
2015	11,468	34	2,937	7	8,531	27
2016	11,779	71	2,841	14	8,938	57
2017	9,885	76	2,771	19	7,114	57
2018	4,355	11	1,236	2	3,119	9
Total	**153,917**	**2,206**	**42,272**	**914**	**111,645**	**1,292**

Panel B: Sample Selection

	Combined		**Corporate**		**People**	
Selection Criteria	**#Alerts**	**#Issues**	**#Alerts**	**#Issues**	**#Alerts**	**#Issues**
All Alerts	206,751	2,440	42,272	914	111,645	1,292
Corp & Ppl Accounts	153,917	2,206	42,272	914	111,645	1,292
Country-level Variables	74,832	1,183	30,303	524	44,529	659
Account/Transaction-level Variables	74,246	1,172	30,292	524	43,954	648

Notes: The table reports the cross-section of our data (Panel A) and the sample selection (Panel B). An alert is raised when a customer's wire activity raises certain flags and an Issue case indicates that the subsequent investigation has deemed the activity to be highly suspicious. The sample selection shows the number of alerts available our data set according to each criterion, applied in sequence. A more detailed description of our variables is available in Table 2.

Table 2: Predictor Details

Predictor	Details	Abbreviation
COUNTRY-LEVEL		
Corruption Perception Index	Score of customer's residence country (R) and country of origin/destination of wire (W) according to Transparency International's Corruption Perception Index.	$\mathrm{CPI_R}$ / $\mathrm{CPI_W}$
Financial Secrecy Index	Score of customer's residence country (R) and country of origin/destination of wire (W) according to Transparency International's Financial Secrecy Index.	$\mathrm{FSI_R}$ / $\mathrm{FSI_W}$
Individualism Index	Score of customer's residence country (R) and country of origin/destination of wire (W) based on Hofstede's "Individualism" dimension of culture.	$\mathrm{IDV_R}$ / $\mathrm{IDV_W}$
Masculinity Index	Score of customer's residence country (R) and country of origin/destination of wire (W) based on Hofstede's "Masculinity" dimension of culture.	$\mathrm{MAS_R}$ / $\mathrm{MAS_W}$
Power-Distance Index	Score of customer's residence country (R) and country of origin/destination of wire (W) based on Hofstede's "Power-Distance" dimension of culture.	$\mathrm{PDI_R}$ / $\mathrm{PDI_W}$
Uncertainty Avoidance Index	Score of customer's residence country (R) and country of origin/destination of wire (W) based on Hofstede's "Uncertainty Avoidance" dimension of culture.	$\mathrm{UAI_R}$ / $\mathrm{UAI_W}$
ACCOUNT-LEVEL		
Customer Age	Age of customer associated with alert, at time of alert.	CUS_AGE
Account Age	Age of account associated with alert, at time of alert.	ACC_AGE
Customer Net Worth	Net Worth of customer associated with alert	NET_WRTH
Alert Supplier Code	Code denoting source of alert, whether alert is generated by Business or Retail transactions.	SUPP_CO
TRANSACTION-LEVEL		
Amount Transfers In	Aggregate amount of incoming wire and electronic transfers over 180 days before alert.	$\Sigma\mathrm{TFI}_{180}$
No. Transfers In	Number of incoming wire and electronic transfers over 180 days before alert.	$\#\mathrm{TFI}_{180}$
Amount Transfers Out	Aggregate amount of outgoing wire and electronic transfers over 180 days before alert.	$\Sigma\mathrm{TFO}_{180}$
No. Transfers Out	Number of outgoing wire and electronic transfers over 180 days before alert.	$\#\mathrm{TFO}_{180}$
Amount Checks In	Aggregate amount of incoming checks over 180 days before alert.	$\Sigma\mathrm{CKI}_{180}$
No. Checks In	Number of incoming checks over 180 days before alert.	$\#\mathrm{CKI}_{180}$
Amount Checks Out	Aggregate amount of outgoing checks over 180 days before alert.	$\Sigma\mathrm{CKO}_{180}$
No. Checks Out	Number of outgoing checks over 180 days before alert.	$\#\mathrm{CKO}_{180}$
Proprietary		
PROP Score	Risk score based on proprietary alert algorithm of financial institution.	PROP

Notes: The table reports the complete set of predictors used in our models along with their definitions and abbreviations for reference. The "Wire" variables refer only to the wire transactions on the day of an alert whereas the "Transfer" and "Check" variables refer to all relevant transactions appearing on accounts associated with an alert in the 180 day period preceding that alert.

Table 3: Country-level Models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**	0.70	0.43	0.722	0.89	0.50	0.845	0.59	0.42	0.670
	Under-sampling	0.76	0.49	0.727	0.90	0.51	0.850	0.61	0.43	0.664
	Hybrid-sampling	0.76	0.50	0.726	0.90	0.47	0.851	0.58	0.40	0.670
	Synthetic-sampling	0.71	0.43	0.723	0.92	0.53	0.861	0.60	0.41	0.659
RF	**No Balancing**	1.00	1.00	0.543	1.00	1.00	0.674	1.00	1.00	0.505
	Under-sampling	0.71	0.40	0.741	0.89	0.41	0.875	0.66	0.41	0.702
	Hybrid-sampling	0.65	0.31	0.726	0.88	0.34	0.878	0.66	0.41	0.695
	Synthetic-sampling	1.00	1.00	0.696	1.00	1.00	0.859	1.00	1.00	0.641
SVM	**No Balancing**	0.53	0.47	0.521	0.42	0.42	0.504	0.62	0.56	0.516
	Under-sampling	0.78	0.60	0.704	0.88	0.59	0.805	0.66	0.44	0.660
	Hybrid-sampling	0.68	0.50	0.662	0.88	0.59	0.807	0.68	0.47	0.636
	Synthetic-sampling	0.77	0.51	0.645	0.89	0.60	0.816	0.59	0.41	0.610
GBM	**No Balancing**	0.87	0.60	0.768	0.91	0.49	0.878	0.84	0.56	0.719
	Under-sampling	0.87	0.59	0.770	0.85	0.41	0.870	0.83	0.57	0.708
	Hybrid-sampling	0.74	0.40	0.771	0.90	0.43	0.881	0.83	0.55	0.716
	Synthetic-sampling	0.68	0.41	0.724	0.94	0.60	0.868	0.73	0.57	0.660

Notes: The table reports the performance of our Country-level model using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 74,724 alerts (30,292 corporate-related and 43,954 people-related) with 1,183 Issue cases (524 corporate-related and 648 people-related). The model has 12 predictors.

Table 4: Cross-validation for Country-level Models with Hybrid-sampling.

Panel A: 5-Fold Cross-validation on AUC scores

	Combined				**Corporate**				**People**			
Round	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**
1	0.717	0.722	0.656	0.765	0.758	0.774	0.785	0.813	0.658	0.662	0.594	0.683
2	0.726	0.726	0.705	0.777	0.856	0.862	0.811	0.896	0.674	0.706	0.663	0.724
3	0.737	0.729	0.705	0.766	0.861	0.873	0.829	0.902	0.671	0.675	0.598	0.709
4	0.743	0.739	0.678	0.789	0.852	0.872	0.842	0.887	0.660	0.681	0.606	0.723
5	0.762	0.768	0.725	0.797	0.820	0.848	0.833	0.874	0.677	0.733	0.688	0.746
μ	0.737	0.737	0.694	0.779	0.829	0.846	0.820	0.874	0.668	0.691	0.630	0.717
σ	0.017	0.019	0.027	0.014	0.043	0.041	0.023	0.036	0.009	0.028	0.043	0.023

Panel B: 10-Fold Cross-validation on AUC scores

	Combined				**Corporate**				**People**			
Round	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**
1	0.769	0.757	0.721	0.799	0.870	0.902	0.865	0.915	0.655	0.692	0.571	0.708
2	0.777	0.770	0.760	0.814	0.798	0.811	0.780	0.864	0.664	0.697	0.570	0.718
3	0.719	0.722	0.674	0.761	0.823	0.818	0.804	0.870	0.651	0.677	0.595	0.702
4	0.746	0.754	0.720	0.812	0.820	0.824	0.823	0.844	0.689	0.672	0.622	0.685
5	0.710	0.693	0.692	0.762	0.810	0.819	0.810	0.867	0.670	0.657	0.570	0.746
6	0.754	0.726	0.714	0.763	0.870	0.867	0.818	0.907	0.686	0.761	0.642	0.774
7	0.743	0.736	0.696	0.769	0.866	0.888	0.847	0.917	0.691	0.735	0.683	0.766
8	0.726	0.740	0.713	0.786	0.807	0.850	0.819	0.869	0.653	0.651	0.550	0.669
9	0.724	0.736	0.738	0.769	0.847	0.865	0.804	0.877	0.648	0.660	0.572	0.670
10	0.729	0.728	0.702	0.775	0.807	0.854	0.788	0.894	0.700	0.738	0.663	0.750
μ	0.740	0.736	0.713	0.781	0.832	0.850	0.816	0.882	0.671	0.694	0.604	0.719
σ	0.022	0.021	0.024	0.021	0.029	0.031	0.025	0.025	0.019	0.038	0.046	0.039

Notes: The table reports the AUCs for 5-fold and 10-fold cross-validation for the hybrid-sampled Country-level model with logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM). The data sample comprises of 82,964 alerts (30,303 corporate-related and 44,529 people-related) with 1,240 Issue cases (524 corporate-related and 659 people-related). The model has 12 predictors.

Table 5: Country-level Predictor Importance for Country-level Model with Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
$\mathbf{CPI_R}$	*	5	5	5	***	5	4	4	***	3	3	3
$\mathbf{FSI_R}$	***	6	3	4	·	8	6	8	***	1	2	2
$\mathbf{IDV_R}$	***	1	1	1	***	1	1	1	***	2	4	4
$\mathbf{MAS_R}$	***	9	6	8		9	11	9	***	9	8	8
$\mathbf{PDI_R}$	***	4	7	6	***	4	7	5	***	6	6	5
$\mathbf{UAI_R}$	***	8	10	9	***	7	5	6	***	8	7	7
$\mathbf{CPI_W}$	***	3	4	3	***	3	3	3	***	4	1	1
$\mathbf{FSI_W}$		12	8	11	***	10	12	12	*	12	11	12
$\mathbf{IDV_W}$		2	2	2	***	2	2	2	***	7	12	11
$\mathbf{MAS_W}$	***	10	11	10		11	10	10		5	10	9
$\mathbf{PDI_W}$	***	7	9	7	*	6	8	7	***	11	9	10
$\mathbf{UAI_W}$	***	11	12	12	***	12	9	11	***	10	5	6

Notes: The table reports the importance of the Country-level predictors by ranking for the Hybrid-sampled Country-level model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 82,964 alerts (30,303 corporate-related and 44,529 people-related) with 1,240 Issue cases (524 corporate-related and 659 people-related). The model has 12 predictors.

Table 6: Country, Account & Transaction-level Models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**	0.72	0.44	0.740	0.84	0.41	0.836	0.73	0.46	0.714
	Under-sampling	0.73	0.41	0.747	0.90	0.46	0.849	0.74	0.48	0.706
	Hybrid-sampling	0.72	0.42	0.747	0.90	0.54	0.846	0.69	0.43	0.712
	Synthetic-sampling	0.71	0.42	0.740	0.87	0.52	0.831	0.74	0.49	0.685
RF	**No Balancing**	0.96	0.53	0.908	0.92	0.28	0.930	1.00	1.00	0.842
	Under-sampling	0.93	0.48	0.895	0.97	0.55	0.932	0.91	0.59	0.835
	Hybrid-sampling	0.94	0.47	0.911	0.96	0.41	0.938	0.88	0.49	0.848
	Synthetic-sampling	1.00	1.00	0.772	0.89	0.42	0.877	1.00	1.00	0.638
SVM	**No Balancing**	0.88	0.51	0.835	0.91	0.59	0.881	0.73	0.48	0.737
	Under-sampling	0.89	0.51	0.801	0.95	0.54	0.880	0.79	0.57	0.739
	Hybrid-sampling	0.86	0.43	0.845	0.90	0.53	0.886	0.72	0.53	0.722
	Synthetic-sampling	0.65	0.41	0.723	0.86	0.52	0.847	0.55	0.40	0.603
GBM	**No Balancing**	0.88	0.47	0.842	0.93	0.59	0.880	0.84	0.57	0.769
	Under-sampling	0.93	0.53	0.853	0.95	0.40	0.916	0.82	0.40	0.799
	Hybrid-sampling	0.87	0.41	0.863	0.93	0.40	0.921	0.83	0.47	0.799
	Synthetic-sampling	0.67	0.40	0.717	0.88	0.40	0.846	0.76	0.55	0.652

Notes: The table reports the performance of our Country, Account & Transaction-level model using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,182 Issue cases (524 corporate-related and 648 people-related). The model has 24 predictors.

Table 7: Country, Account & Transaction-level Models with PROP Score

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**	0.77	0.48	0.754	0.82	0.40	0.840	0.79	0.47	0.733
	Under-sampling	0.78	0.45	0.763	0.91	0.49	0.848	0.79	0.47	0.734
	Hybrid-sampling	0.77	0.44	0.764	0.90	0.58	0.851	0.76	0.46	0.733
	Synthetic-sampling	0.79	0.47	0.756	0.86	0.47	0.834	0.83	0.56	0.723
RF	**No Balancing**	0.89	0.40	0.894	0.95	0.33	0.946	1.00	1.00	0.845
	Under-sampling	0.87	0.43	0.878	0.95	0.40	0.943	0.91	0.57	0.846
	Hybrid-sampling	0.92	0.44	0.901	0.97	0.44	0.952	0.83	0.41	0.855
	Synthetic-sampling	0.90	0.58	0.790	0.88	0.43	0.873	1.00	1.00	0.690
SVM	**No Balancing**	0.87	0.50	0.828	0.89	0.54	0.883	0.78	0.59	0.742
	Under-sampling	0.88	0.57	0.789	0.94	0.51	0.896	0.81	0.57	0.753
	Hybrid-sampling	0.88	0.53	0.833	0.91	0.57	0.896	0.75	0.54	0.731
	Synthetic-sampling	0.78	0.56	0.745	0.83	0.41	0.848	0.65	0.49	0.667
GBM	**No Balancing**	0.89	0.56	0.829	0.92	0.58	0.886	0.91	0.57	0.818
	Under-sampling	0.87	0.43	0.850	0.94	0.40	0.922	0.91	0.51	0.828
	Hybrid-sampling	0.87	0.42	0.855	0.96	0.55	0.926	0.83	0.40	0.828
	Synthetic-sampling	0.74	0.48	0.709	0.88	0.48	0.840	0.80	0.58	0.689

Notes: The table reports the performance of our Country, Account & Transaction-level model, with the PROP score variable included, using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 74,724 alerts (30,292 corporate-related and 43,954 people-related) with 1,182 Issue cases (524 corporate-related and 648 people-related). The model has 25 predictors.

Table 8: Absolute Country-level Predictor Importance for Country, Account & Transaction-level Model with Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
$\mathbf{CPI_R}$		12	18	17	***	12	18	12	***	14	13	13
$\mathbf{FSI_R}$	***	8	6	7		14	12	14	***	9	9	9
$\mathbf{IDV_R}$		5	1	1	***	2	1	1	***	7	1	2
$\mathbf{MAS_R}$	***	15	11	13	***	15	19	16	***	12	17	16
$\mathbf{PDI_R}$	***	11	12	11	***	9	10	10	·	16	19	17
$\mathbf{UAI_R}$	***	13	16	15	***	5	4	4	·	15	14	15
$\mathbf{CPI_W}$	***	17	7	10	***	10	14	11	***	17	4	7
$\mathbf{FSI_W}$	***	21	17	20		16	20	20	**	23	22	24
$\mathbf{IDV_W}$	***	10	4	5		4	2	3	***	20	24	22
$\mathbf{MAS_W}$	*	18	21	19		18	17	19		18	23	21
$\mathbf{PDI_W}$	***	22	20	21		13	13	13	***	24	21	23
$\mathbf{UAI_W}$	***	23	23	23	***	21	23	21		21	15	18

Notes: The table reports the importance of the Country-level predictors by absolute ranking for the Hybrid-sampled Country, Account & Transaction-level model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,182 Issue cases (524 corporate-related and 648 people-related). The model has 24 predictors.

Table 9: Absolute Country-level Predictor Importance for Country, Account & Transaction-level Model with Hybrid-sampling and PROP Score included

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
PROP	***	5	3	3	***	7	10	9	***	4	1	1
CPI$_{\mathbf{R}}$		12	11	12	***	14	12	13	***	10	6	8
FSI$_{\mathbf{R}}$	***	9	6	8		15	14	15	***	8	4	7
IDV$_{\mathbf{R}}$		8	4	5	***	3	2	2	***	13	12	13
MAS$_{\mathbf{R}}$	***	19	17	17	***	17	18	19	***	15	13	16
PDI$_{\mathbf{R}}$	***	14	20	18	***	11	8	10	·	14	16	15
UAI$_{\mathbf{R}}$	***	13	14	13	***	12	11	11	·	17	17	17
CPI$_{\mathbf{W}}$	***	18	10	11	***	13	6	6	***	18	5	10
FSI$_{\mathbf{W}}$	***	24	19	21		16	17	17	**	24	20	24
IDV$_{\mathbf{W}}$	***	11	1	2		2	1	1	***	21	24	21
MAS$_{\mathbf{W}}$	*	16	22	20		18	16	16		20	25	20
PDI$_{\mathbf{W}}$	***	21	21	22		8	23	14	***	25	19	25
UAI$_{\mathbf{W}}$	***	23	23	23	***	22	20	22		23	21	22

Notes: The table reports the importance of the Country-level predictors by absolute ranking for the Hybrid-sampled Country, Account & Transaction-level model applied to the full sample (combined) and its partitions (Corporate & People accounts) with the PROP score variable included. Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 74,724 alerts (30,292 corporate-related and 43,954 people-related) with 1,182 Issue cases (524 corporate-related and 648 people-related). the model has 25 predictors.

References

Adomavicius, Gediminas and Alexander Tuzhilin (2001). “Using data mining methods to build customer profiles”. In: *Computer* 34(2), pp. 74–82.

Aggarwal, Raj, Joanne E Goodell, and John W Goodell (2014). “Culture, gender, and GMAT scores: Implications for corporate ethics”. In: *Journal of Business Ethics* 123(1), pp. 125–143.

Aggarwal, Raj, John W Goodell, and Lauren J Selleck (2015). “Lending to women in microfinance: Role of social trust”. In: *International Business Review* 24(1), pp. 55–65.

Alter, Adam L and John M Darley (2009). “When the association between appearance and outcome contaminates social judgment: A bidirectional model linking group homogeneity and collective treatment.” In: *Journal of Personality and Social Psychology* 97(5), p. 776.

Arce, Daniel G and Mary C Gentile (2015). “Giving voice to values as a leverage point in business ethics education”. In: *Journal of Business Ethics* 131(3), pp. 535–542.

Archambault, Jeffrey J and Marie E Archambault (2003). “A multinational test of determinants of corporate disclosure”. In: *The International Journal of Accounting* 38(2), pp. 173–194.

Armstrong, Robert W (1996). “The relationship between culture and perception of ethical problems in international marketing”. In: *Journal of Business Ethics* 15(11), pp. 1199–1208.

Barocas, Solon, Sophie Hood, and Malte Ziewitz (2013). “Governing algorithms: A provocation piece”. In: *Available at SSRN 2245322.*

Baxamusa, Mufaddal and Abu Jalal (2014). “Does religion affect capital structure?” In: *Research in International Business and Finance* 31, pp. 112–131.

Berger, Allen N et al. (2019). “The effects of cultural values on bank failures around the world”. In: *Journal of Financial and Quantitative Analysis (JFQA), Forthcoming.*

Berk, Richard (2017). “An impact assessment of machine learning risk forecasts on parole board decisions and recidivism”. In: *Journal of Experimental Criminology* 13(2), pp. 193–216.

Bianchi, Daniele, Matthias Büchner, and Andrea Tamoni (2020). “Bond risk premiums with machine learning”. In: *The Review of Financial Studies.*

Breiman, Leo et al. (1984). *Classification and regression trees.* CRC press.

Brewer, Marilynn B and Amy S Harasty (1996). “Seeing groups as entities: The role of perceiver motivation.” In.

Buhmann, Alexander, Johannes Paßmann, and Christian Fieseler (2019). “Managing algorithmic accountability: Balancing reputational concerns, engagement strategies, and the potential of rational discourse”. In: *Journal of Business Ethics*, pp. 1–16.

Butaru, Florentin et al. (2016). “Risk and risk management in the credit card industry”. In: *Journal of Banking & Finance* 72, pp. 218–239.

Campbell, Donald T (1958). “Common fate, similarity, and other indices of the status of aggregates of persons as social entities”. In: *Behavioral Science* 3(1), p. 14.

Cobham, Alex, Petr Jansk, and Markus Meinzer (2015). “The financial secrecy index: Shedding new light on the geography of secrecy”. In: *Economic Geography* 91(3), pp. 281–303.

Cohen, Jeffrey R, Laurie W Pant, and David J Sharp (1996). “A methodological note on cross-cultural accounting ethics research”. In: *The International Journal of Accounting* 31(1), pp. 55–66.

Cook, Jack (2008). “Ethics of data mining”. In: *Information Security and Ethics: Concepts, Methodologies, Tools, and Applications.* IGI Global, pp. 211–217.

Coulombe, Philippe Goulet et al. (2020). “How is machine learning useful for macroeconomic forecasting?” In: *arXiv preprint arXiv:2008.12477.*

Dasgupta, Nilanjana, Mahzarin R Banaji, and Robert P Abelson (1999). "Group entitativity and group perception: Associations between physical features and psychological judgment." In: *Journal of Personality and Social Psychology* 77(5), p. 991.

Davis, James H and John A Ruhe (2003). "Perceptions of country corruption: Antecedents and outcomes". In: *Journal of Business Ethics* 43(4), pp. 275–288.

De Jong, Eelke, Roger Smeets, and Jeroen Smits (2006). "Culture and openness". In: *Social Indicators Research* 78(1), pp. 111–136.

Desai, Deven R and Joshua A Kroll (2017). "Trust but verify: A guide to algorithms and the law". In: *Harv. JL & Tech.* 31, p. 1.

Donaldson, Thomas and Thomas W Dunfee (1994). "Toward a unified conception of business ethics: Integrative social contracts theory". In: *Academy of management review* 19(2), pp. 252–284.

Freund, Yoav, Robert E Schapire, et al. (1996). "Experiments with a new boosting algorithm". In: *icml*. Vol. 96. Citeseer, pp. 148–156.

Friedman, Jerome H (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*, pp. 1189–1232.

Fuster, Andreas et al. (2018). "Predictably unequal? the effects of machine learning on credit markets". In: *The Effects of Machine Learning on Credit Markets (November 6, 2018).*

Getz, Kathleen A and Roger J Volkema (2001). "Culture, perceived corruption, and economics: A model of predictors and outcomes". In: *Business & Society* 40(1), pp. 7–30.

Goodell, John W (2019). "Comparing normative institutionalism with intended rationality in cultural-finance research". In: *International Review of Financial Analysis* 62, pp. 124–134.

Gray, Sidney J (1988). "Towards a theory of cultural influence on the development of accounting systems internationally". In: *Abacus* 24(1), pp. 1–15.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2018). *Empirical asset pricing via machine learning.* Tech. rep. National Bureau of Economic Research.

Harris, David A (1996). "Driving while black and all other traffic offenses: The Supreme Court and pretextual traffic stops". In: *J. Crim. L. & Criminology* 87, p. 544.

Hartney, Christopher (2009). *Created equal: Racial and ethnic disparities in the US criminal justice system.* National Council on Crime and Delinquency.

Hassan, Omaima AG and Gianluigi Giorgioni (2015). "Analyst coverage, corruption and financial secrecy: a multi-country study". In: *Corruption and Financial Secrecy: A Multi-Country Study (February 18, 2015).*

Hope, Ole-Kristian (2003). "Firm-level disclosures and the relative roles of culture and legal origin". In: *Journal of International Financial Management & Accounting* 14(3), pp. 218–248.

Houqe, Noor et al. (2015). "Secrecy and mandatory IFRS adoption on earnings quality". In.

Husted, Bryan W (2000). "The impact of national culture on software piracy". In: *Journal of Business Ethics* 26(3), pp. 197–211.

Introna, Lucas D (2016). "Algorithms, governance, and governmentality: On governing academic writing". In: *Science, Technology, & Human Values* 41(1), pp. 17–49.

Jaggi, Bikki and Pek Yee Low (2000). "Impact of culture, market forces, and legal system on financial disclosures". In: *The International Journal of Accounting* 35(4), pp. 495–519.

Johnson, Scott G, Karen Schnatterly, and Aaron D Hill (2013). "Board composition beyond independence: Social capital, human capital, and demographics". In: *Journal of Management* 39(1), pp. 232–262.

Karolyi, G Andrew (2016). "The gravity of culture for finance". In: *Journal of Corporate Finance* 41, pp. 610–625.

Khandani, Amir E, Adlar J Kim, and Andrew W Lo (2010). “Consumer credit-risk models via machine-learning algorithms”. In: *Journal of Banking & Finance* 34(11), pp. 2767–2787.

Kim, Jeong-Bon, Zheng Wang, and Liandong Zhang (2016). “CEO overconfidence and stock price crash risk”. In: *Contemporary Accounting Research* 33(4), pp. 1720–1749.

Kirkman, Bradley L, Kevin B Lowe, and Cristina B Gibson (2006). “A quarter century of culture’s consequences: A review of empirical research incorporating Hofstede’s cultural values framework”. In: *Journal of International Business Studies* 37(3), pp. 285–320.

Kirkman, Bradley L, Kevin B Lowe, and Cristina B Gibson (2017). “A retrospective on Culture’s Consequences: The 35-year journey”. In: *Journal of International Business Studies* 48(1), pp. 12–29.

Kuhn, Max, Kjell Johnson, et al. (2013). *Applied predictive modeling.* Vol. 26. Springer.

Kumar, Gaurav et al. (2019). “Can alert models for fraud protect the elderly clients of a financial institution?” In: *The European Journal of Finance* 25(17), pp. 1683–1707.

Lee, Nicol Turner (2018). “Detecting racial bias in algorithms and machine learning”. In: *Journal of Information, Communication and Ethics in Society.*

Liu, Xiaoding (2016). “Corruption culture and corporate misconduct”. In: *Journal of Financial Economics* 122(2), pp. 307–327.

Martin, Kirsten (2019). “Ethical implications and accountability of algorithms”. In: *Journal of Business Ethics* 160(4), pp. 835–850.

Michalos, Alex C and P Maurine Hatch (2019). “Good Societies, Financial Inequality and Secrecy, and a Good Life: from Aristotle to Piketty”. In: *Applied Research in Quality of Life*, pp. 1–50.

Musiani, Francesca (2013). “Governance by algorithms”. In: *Internet Policy Review* 2(3), pp. 1–8.

Peterson, Mark F and Tais S Barreto (2018). “Interpreting societal culture value dimensions”. In: *Journal of International Business Studies* 49(9), pp. 1190–1207.

Puspitasari, Evita et al. (n.d.). “The Effect of Financial Secrecy and IFRS Adoption on Earnings Quality: A Comparative Study between Indonesia, Malaysia and Singapore”. In: ().

Salter, Stephen B and Frederick Niswander (1995). “Cultural influence on the development of accounting systems internationally: A test of Gray’s [1988] theory”. In: *Journal of International Business Studies* 26(2), pp. 379–397.

Seaver, Nick (2019). “Knowing algorithms”. In: *Digital STS*, pp. 412–422.

Seele, Peter et al. (2019). “Mapping the Ethicality of Algorithmic Pricing: A Review of Dynamic and Personalized Pricing”. In: *Journal of Business Ethics*, pp. 1–23.

Sorley, William Ritchie (1885). *On the Ethics of Naturalism.* W. Blackwood and Sons.

Stulz, Rene M and Rohan Williamson (2003). “Culture, openness, and finance”. In: *Journal of financial Economics* 70(3), pp. 313–349.

Tanzi, Vito (1980). “Inflationary expectations, economic activity, taxes, and interest rates”. In: *The American Economic Review* 70(1), pp. 12–21.

Tung, Rosalie L and Günter K Stahl (2018). “The tortuous evolution of the role of culture in IB research: What we know, what we don’t know, and where we are headed”. In: *Journal of International Business Studies* 49(9), pp. 1167–1189.

Tung, Rosalie L and Alain Verbeke (2010). *Beyond Hofstede and GLOBE: Improving the quality of cross-cultural research.*

Ucar, Erdem (2016). “Local culture and dividends”. In: *Financial Management* 45(1), pp. 105–140.

Velayutham, Sivakumar and MHB Perera (2004). "The influence of emotions and culture on accountability and governance". In: *Corporate Governance: The International Journal of Business in Society.*

Vitell, Scott J, Saviour L Nwachukwu, and James H Barnes (1993). "The effects of culture on ethical decision-making: An application of Hofstede's typology". In: *Journal of Business Ethics* 12(10), pp. 753–760.

Volkema, Roger J (2004). "Demographic, cultural, and economic predictors of perceived ethicality of negotiation behavior: A nine-country analysis". In: *Journal of Business Research* 57(1), pp. 69–78.

Williams, Kevin M, Craig Nathanson, and Delroy L Paulhus (2010). "Identifying and profiling scholastic cheaters: Their personality, cognitive ability, and motivation." In: *Journal of Experimental Psychology: Applied* 16(3), p. 293.

Zarzeski, Marilyn Taylor (1996). "Spontaneous harmonization effects of culture and market forces on accounting disclosure practices". In: *Accounting Horizons* 10(1), p. 18.

Ziewitz, Malte (2016). "Governing algorithms: Myth, mess, and methods". In: *Science, Technology, & Human Values* 41(1), pp. 3–16.

article [utf8]inputenc graphicx caption ltablex enumitem pdflscape tabularx threeparttable upgreek changepage array varwidth [title]appendix indentfirst graphicx subfloat xcolor [a4paper, total=5.7in, 8in]geometry [font=footnotesize]caption

Internet Appendices

A. Account Registration Types

Table A: Registration Type Profile

Reg Type	# Alerts	Alert Share	# Issues	Issue Share	Issue Rate
Corporate	44,159	21.08 %	936	38.13 %	2.12 %
Education	3,169	1.51 %	10	0.41 %	0.32 %
Estate-like	670	0.32 %	0	0.00 %	0.00 %
IRA	19,745	9.43 %	14	0.57 %	0.07 %
People	119,717	57.15 %	1,366	55.64 %	1.14 %
Trust	22,024	10.51 %	129	5.25 %	0.59 %

Notes: The table reports the cross-section of Alerts and Issues over the different reg types that comprise the accounts which trigger the alerts. The categories of Corporate and People together compromise 78.23% of the alerts and 93.77% of the Issue cases in total and so, for the purposes of our study, we only consider these two reg types.

B. Hofstede Indices

Table B1: Hofstede Indices Models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**	0.70	0.43	0.695	0.87	0.53	0.813	0.58	0.40	0.651
	Under-sampling	0.70	0.44	0.711	0.87	0.49	0.818	0.67	0.51	0.659
	Hybrid-sampling	0.70	0.42	0.711	0.87	0.50	0.818	0.65	0.48	0.663
	Synthetic-sampling	0.78	0.53	0.711	0.81	0.44	0.812	0.71	0.54	0.658
RF	**No Balancing**	1.00	1.00	0.573	1.00	1.00	0.664	1.00	1.00	0.514
	Under-sampling	0.71	0.42	0.747	0.86	0.42	0.848	0.75	0.49	0.718
	Hybrid-sampling	0.71	0.41	0.730	0.82	0.32	0.848	0.78	0.49	0.707
	Synthetic-sampling	1.00	1.00	0.713	1.00	1.00	0.835	1.00	1.00	0.654
SVM	**No Balancing**	0.67	0.53	0.545	0.51	0.45	0.532	0.72	0.53	0.625
	Under-sampling	0.74	0.56	0.670	0.92	0.52	0.830	0.62	0.41	0.648
	Hybrid-sampling	0.73	0.54	0.686	0.89	0.56	0.829	0.79	0.59	0.641
	Synthetic-sampling	0.60	0.59	0.531	0.62	0.41	0.719	0.66	0.58	0.586
GBM	**No Balancing**	0.82	0.52	0.765	0.89	0.57	0.867	0.82	0.56	0.727
	Under-sampling	0.81	0.52	0.767	0.84	0.41	0.861	0.82	0.56	0.726
	Hybrid-sampling	0.83	0.54	0.771	0.84	0.41	0.866	0.82	0.55	0.739
	Synthetic-sampling	0.68	0.41	0.716	0.85	0.57	0.811	0.71	0.51	0.662

Notes: The table reports the performance of our Hofstede Indices model with Cultural Distance using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 81,858 alerts (32,482 corporate-related and 49,376 people-related) with 1,273 Issue cases (537 corporate-related and 736 people-related). The model has 8 predictors.

Table B2: Predictor Importance for Hofstede Indices Model with Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
IDV$_{\mathbf{S}}$	***	1	1	1	***	1	1	1	***	1	2	1
MAS$_{\mathbf{S}}$	***	7	5	5		7	6	6	***	2	3	3
PDI$_{\mathbf{S}}$	.	2	3	3	***	2	3	3	***	4	5	5
UAI$_{\mathbf{S}}$	***	4	4	4	***	5	5	5	***	5	4	4
IDV$_{\mathbf{R}}$	***	3	2	2		3	4	4	*	6	8	7
MAS$_{\mathbf{R}}$	***	5	6	6	***	6	8	8	***	3	1	2
PDI$_{\mathbf{R}}$	.	6	7	7	***	4	2	2	**	7	7	8
UAI$_{\mathbf{R}}$	*	8	8	8	***	8	7	7	***	8	6	6

Notes: The table reports the importance of the predictors for the Hybrid-sampled Hofstede Indices model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 81,858 alerts (32,482 corporate-related and 49,376 people-related) with 1,273 Issue cases (537 corporate-related and 736 people-related). The model has 8 predictors.

Table B3: Cross-validation for Hofstede Indices Model with Hybrid-sampling.

Panel A: 5-Fold Cross-validation on AUC scores

	Combined				**Corporate**				**People**			
Round	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**
1	0.701	0.745	0.736	0.772	0.855	0.877	0.848	0.873	0.647	0.691	0.629	0.717
2	0.734	0.761	0.721	0.796	0.814	0.859	0.794	0.876	0.648	0.664	0.618	0.714
3	0.710	0.741	0.728	0.778	0.805	0.837	0.801	0.863	0.644	0.684	0.653	0.685
4	0.717	0.744	0.695	0.773	0.787	0.842	0.821	0.884	0.640	0.731	0.680	0.756
5	0.692	0.731	0.700	0.768	0.774	0.825	0.780	0.842	0.694	0.703	0.680	0.721
μ	0.711	0.744	0.716	0.777	0.807	0.848	0.809	0.868	0.655	0.695	0.652	0.719
σ	0.016	0.011	0.018	0.011	0.031	0.020	0.026	0.016	0.022	0.025	0.029	0.025

Panel B: 10-Fold Cross-validation on AUC scores

	Combined				**Corporate**				**People**			
Round	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**
1	0.757	0.764	0.689	0.785	0.760	0.810	0.751	0.851	0.658	0.668	0.618	0.699
2	0.745	0.771	0.714	0.822	0.795	0.836	0.799	0.875	0.716	0.690	0.634	0.704
3	0.706	0.735	0.703	0.766	0.772	0.789	0.781	0.825	0.618	0.691	0.612	0.710
4	0.682	0.745	0.729	0.774	0.773	0.826	0.795	0.884	0.664	0.735	0.650	0.740
5	0.762	0.775	0.736	0.817	0.706	0.770	0.756	0.821	0.654	0.687	0.626	0.720
6	0.690	0.723	0.701	0.733	0.891	0.903	0.867	0.931	0.653	0.734	0.655	0.747
7	0.729	0.764	0.716	0.779	0.865	0.907	0.869	0.898	0.665	0.777	0.636	0.767
8	0.697	0.728	0.694	0.772	0.828	0.889	0.872	0.896	0.644	0.670	0.639	0.688
9	0.703	0.757	0.720	0.781	0.814	0.864	0.844	0.898	0.640	0.735	0.667	0.723
10	0.648	0.684	0.655	0.744	0.850	0.903	0.840	0.900	0.616	0.683	0.638	0.703
μ	0.712	0.745	0.706	0.777	0.805	0.850	0.817	0.878	0.653	0.707	0.637	0.720
σ	0.036	0.028	0.023	0.028	0.055	0.051	0.047	0.035	0.028	0.036	0.017	0.025

Notes: The table reports the AUCs for 5-fold and 10-fold cross-validation for the hybrid-sampled Hofstede Indices model with logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM). The data sample comprises of 81,858 alerts (32,482 corporate-related and 49,376 people-related) with 1,273 Issue cases (537 corporate-related and 736 people-related). The model has 8 predictors.

C. Schwarz Indices

Table C1: Country, Account & Transaction-level Models (Schwartz Indices)

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**	0.72	0.41	0.736	0.78	0.40	0.821	0.80	0.49	0.753
	Under-sampling	0.76	0.47	0.746	0.86	0.40	0.827	0.78	0.46	0.754
	Hybrid-sampling	0.77	0.45	0.750	0.96	0.48	0.856	0.80	0.50	0.756
	Synthetic-sampling	0.81	0.54	0.736	0.91	0.50	0.855	0.82	0.57	0.733
RF	**No Balancing**	1.00	1.00	0.888	1.00	1.00	0.923	1.00	1.00	0.873
	Under-sampling	0.90	0.49	0.872	0.96	0.56	0.910	0.93	0.60	0.870
	Hybrid-sampling	0.94	0.54	0.899	0.95	0.43	0.928	0.93	0.47	0.892
	Synthetic-sampling	0.83	0.46	0.748	0.89	0.58	0.831	0.76	0.45	0.747
SVM	**No Balancing**	0.84	0.50	0.813	0.91	0.40	0.908	0.75	0.43	0.766
	Under-sampling	0.84	0.43	0.785	0.90	0.40	0.842	0.85	0.58	0.787
	Hybrid-sampling	0.83	0.43	0.827	0.89	0.53	0.869	0.81	0.56	0.760
	Synthetic-sampling	0.86	0.57	0.756	0.94	0.45	0.897	0.72	0.48	0.709
GBM	**No Balancing**	0.94	0.57	0.842	0.90	0.59	0.867	0.89	0.570	0.818
	Under-sampling	0.85	0.40	0.839	0.92	0.44	0.876	0.93	0.54	0.822
	Hybrid-sampling	0.96	0.57	0.860	0.96	0.52	0.905	0.93	0.60	0.841
	Synthetic-sampling	0.69	0.40	0.704	0.90	0.43	0.815	0.70	0.41	0.715

Notes: The table reports the performance of our Country, Account & Transaction-level model using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The model uses the Schwarz Indices for the Country-level predictors. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 53,956 alerts (22,125 corporate-related and 31,441 people-related) with 731 Issue cases (265 corporate-related and 563 people-related). The model has 22 predictors.

Table C2: Absolute Country-level Predictor Importance for Country, Account & Transaction-level Model with Hybrid-sampling (Schwartz Indices)

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
CPI$_{\mathbf{S}}$	***	1	1	1	***	2	3	3	***	2	2	2
FSI$_{\mathbf{S}}$	***	2	2	2	***	4	2	2	***	1	1	1
EBD$_{\mathbf{S}}$	***	4	4	4	***	5	4	5	***	4	5	4
EGA$_{\mathbf{S}}$	***	7	7	7	***	7	5	6	**	5	6	5
HIE$_{\mathbf{S}}$	***	5	5	5	***	8	10	8	***	3	3	3
CPI$_{\mathbf{R}}$	***	3	3	3	***	3	7	4	***	6	4	6
FSI$_{\mathbf{R}}$	***	8	8	8	*	6	6	7	***	9	10	10
EBD$_{\mathbf{R}}$	***	6	6	6		1	1	1	*	10	7	7
EGA$_{\mathbf{R}}$	***	10	10	10		10	8	10		7	9	9
HIE$_{\mathbf{R}}$		9	9	9	***	9	9	9		8	8	8

Notes: The table reports the importance of the Country-level predictors by absolute ranking for the Hybrid-sampled Country, Account & Transaction-level model applied to the full sample (combined) and its partitions (Corporate & People accounts). The model uses the Schwarz Indices for the Country-level predictors. Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 53,956 alerts (22,125 corporate-related and 31,441 people-related) with 731 Issue cases (265 corporate-related and 463 people-related). The model has 22 predictors.

Table C3: Cross-validation for Country, Account & Transaction-level Model with Hybrid-sampling (Schwartz Indices)

Panel A: 5-Fold Cross-validation on AUC scores

	Combined				**Corporate**				**People**			
Round	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**
1	0.745	0.915	0.857	0.882	0.863	0.951	0.902	0.927	0.753	0.875	0.818	0.820
2	0.768	0.903	0.850	0.855	0.825	0.959	0.890	0.941	0.743	0.889	0.829	0.816
3	0.782	0.931	0.860	0.866	0.845	0.952	0.915	0.925	0.784	0.935	0.869	0.900
4	0.750	0.888	0.836	0.846	0.808	0.946	0.903	0.924	0.722	0.863	0.735	0.820
5	0.758	0.934	0.895	0.875	0.855	0.940	0.898	0.938	0.791	0.925	0.806	0.873
μ	0.761	0.914	0.860	0.865	0.839	0.950	0.902	0.931	0.759	0.897	0.811	0.846
σ	0.015	0.019	0.022	0.015	0.022	0.007	0.009	0.008	0.029	0.031	0.049	0.038

Panel B: 10-Fold Cross-validation on AUC scores

	Combined				**Corporate**				**People**			
Round	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**	**LR**	**RF**	**SVM**	**GBM**
1	0.787	0.921	0.855	0.895	0.814	0.928	0.928	0.929	0.731	0.888	0.821	0.851
2	0.709	0.922	0.827	0.822	0.848	0.933	0.807	0.915	0.758	0.940	0.911	0.884
3	0.753	0.902	0.856	0.870	0.898	0.974	0.942	0.966	0.808	0.896	0.778	0.853
4	0.807	0.922	0.888	0.866	0.869	0.992	0.954	0.974	0.750	0.872	0.798	0.843
5	0.785	0.935	0.895	0.895	0.800	0.983	0.959	0.955	0.804	0.940	0.905	0.903
6	0.768	0.957	0.901	0.914	0.850	0.978	0.960	0.964	0.788	0.915	0.846	0.851
7	0.764	0.928	0.844	0.858	0.885	0.917	0.919	0.955	0.771	0.944	0.841	0.879
8	0.735	0.916	0.866	0.847	0.862	0.974	0.905	0.945	0.764	0.905	0.847	0.848
9	0.789	0.962	0.908	0.889	0.780	0.930	0.819	0.877	0.742	0.893	0.801	0.839
10	0.740	0.901	0.837	0.860	0.774	0.897	0.844	0.893	0.661	0.873	0.864	0.781
μ	0.764	0.927	0.868	0.872	0.838	0.951	0.904	0.937	0.758	0.907	0.841	0.853
σ	0.030	0.020	0.029	0.027	0.044	0.033	0.059	0.033	0.042	0.027	0.044	0.033

Notes: The table reports the AUCs for 5-fold and 10-fold cross-validation for the hybrid-sampled Country-, Account- & Transaction-level model with logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM). The model uses the Schwarz Indices for the Country-level predictors. The data sample comprises of 53,956 alerts (22,125 corporate-related and 31,441 people-related) with 731 Issue cases (265 corporate-related and 463 people-related). The model has 22 predictors.

D. Miscellaneous

Table D1: Country & Account-level models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**	0.79	0.50	0.728	0.83	0.37	0.824	0.80	0.60	0.681
	Under-sampling	0.89	0.67	0.735	0.91	0.51	0.812	0.91	0.77	0.688
	Hybrid-sampling	0.83	0.56	0.739	0.79	0.27	0.829	0.85	0.67	0.677
	Synthetic-sampling	0.88	0.64	0.727	0.92	0.53	0.814	0.84	0.65	0.671
RF	**No Balancing**	0.81	0.30	0.855	0.89	0.22	0.915	0.80	0.32	0.826
	Under-sampling	0.89	0.47	0.837	0.88	0.20	0.929	0.91	0.62	0.793
	Hybrid-sampling	0.86	0.36	0.862	0.90	0.23	0.942	0.81	0.38	0.812
	Synthetic-sampling	0.88	0.66	0.723	0.83	0.32	0.863	0.82	0.67	0.643
SVM	**No Balancing**	0.86	0.57	0.742	0.91	0.51	0.862	0.81	0.49	0.756
	Under-sampling	0.83	0.53	0.760	0.79	0.21	0.865	0.89	0.74	0.696
	Hybrid-sampling	0.90	0.53	0.807	0.86	0.27	0.904	0.80	0.51	0.732
	Synthetic-sampling	0.84	0.74	0.625	0.84	0.37	0.825	0.85	0.76	0.614
GBM	**No Balancing**	0.81	0.38	0.805	0.79	0.15	0.908	0.91	0.67	0.749
	Under-sampling	0.84	0.43	0.818	0.89	0.24	0.911	0.85	0.57	0.752
	Hybrid-sampling	0.83	0.36	0.830	0.85	0.17	0.926	0.91	0.67	0.760
	Synthetic-sampling	0.92	0.72	0.735	0.86	0.36	0.832	0.86	0.69	0.679

Notes: The table reports the performance of our Country & Account-level model using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). The model has 16 predictors.

Table D2: Account-level Models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**			0.645			0.661			0.620
	Under-sampling			0.646			0.660			0.620
	Hybrid-sampling			0.645			0.659			0.622
	Synthetic-sampling			0.645			0.659			0.622
RF	**No Balancing**			0.775			0.832			0.782
	Under-sampling			0.724			0.805			0.791
	Hybrid-sampling			0.738			0.822			0.786
	Synthetic-sampling			0.646			0.652			0.542
SVM	**No Balancing**			0.621			0.681			0.665
	Under-sampling			0.694			0.740			0.679
	Hybrid-sampling			0.700			0.754			0.719
	Synthetic-sampling			0.644			0.652			0.589
GBM	**No Balancing**			0.759			0.856			0.711
	Under-sampling			0.745			0.829			0.699
	Hybrid-sampling			0.752			0.859			0.720
	Synthetic-sampling			0.655			0.682			0.611

Notes: The table reports the performance of our Account-level model using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 151,985 alerts (42,193 corporate-related and 109,792 people-related) with 2,179 Issue cases (914 corporate-related and 1,265 people-related). The model has 4 predictors.

Table D3: Transaction-level Models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**			0.540			0.654			0.671
	Under-sampling			0.532			0.645			0.676
	Hybrid-sampling			0.546			0.635			0.676
	Synthetic-sampling			0.556			0.640			0.675
RF	**No Balancing**			0.638			0.630			0.650
	Under-sampling			0.754			0.784			0.747
	Hybrid-sampling			0.757			0.779			0.733
	Synthetic-sampling			0.573			0.622			0.600
SVM	**No Balancing**			0.644			0.673			0.562
	Under-sampling			0.682			0.709			0.683
	Hybrid-sampling			0.706			0.739			0.785
	Synthetic-sampling			0.548			0.546			0.626
GBM	**No Balancing**			0.729			0.768			0.719
	Under-sampling			0.723			0.763			0.732
	Hybrid-sampling			0.740			0.775			0.733
	Synthetic-sampling			0.535			0.596			0.611

Notes: The table reports the performance of our Transaction-level model using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 153,913 alerts (42,271 corporate-related and 111,642 people-related) with 2,206 Issue cases (914 corporate-related and 1,292 people-related). The model has 8 predictors.

Table D4: Account & Transaction-level Models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**			0.657			0.700			0.674
	Under-sampling			0.651			0.696			0.687
	Hybrid-sampling			0.656			0.699			0.692
	Synthetic-sampling			0.655			0.691			0.675
RF	**No Balancing**			0.846			0.900			0.821
	Under-sampling			0.847			0.895			0.808
	Hybrid-sampling			0.851			0.908			0.815
	Synthetic-sampling			0.654			0.638			0.625
SVM	**No Balancing**			0.677			0.827			0.698
	Under-sampling			0.757			0.831			0.719
	Hybrid-sampling			0.795			0.852			0.732
	Synthetic-sampling			0.653			0.672			0.613
GBM	**No Balancing**			0.799			0.854			0.775
	Under-sampling			0.795			0.852			0.775
	Hybrid-sampling			0.805			0.861			0.791
	Synthetic-sampling			0.644			0.564			0.626

Notes: The table reports the performance of our Account & Transaction-level model using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 151,985 alerts (42,193 corporate-related and 109,792 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). The model has 12 predictors.

Table D5: Country, Account, Transaction-level & Cultural Distance Models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**			0.750			0.861			0.742
	Under-sampling			0.762			0.867			0.755
	Hybrid-sampling			0.764			0.872			0.753
	Synthetic-sampling			0.742			0.850			0.725
RF	**No Balancing**			0.881			0.939			0.861
	Under-sampling			0.875			0.935			0.833
	Hybrid-sampling			0.888			0.938			0.858
	Synthetic-sampling			0.774			0.876			0.669
SVM	**No Balancing**			0.822			0.866			0.744
	Under-sampling			0.793			0.887			0.750
	Hybrid-sampling			0.829			0.881			0.740
	Synthetic-sampling			0.738			0.865			0.653
GBM	**No Balancing**			0.838			0.904			0.804
	Under-sampling			0.853			0.922			0.806
	Hybrid-sampling			0.859			0.936			0.830
	Synthetic-sampling			0.703			0.842			0.676

Notes: The table reports the performance of our Country, Account, Transaction-level model with Cultural Distance using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). The model has 25 predictors.

Table D6: Cultural Distance, Account & Transaction-level Models

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**			0.762			0.858			0.728
	Under-sampling			0.766			0.867			0.727
	Hybrid-sampling			0.769			0.860			0.730
	Synthetic-sampling			0.755			0.844			0.717
RF	**No Balancing**			0.901			0.957			0.849
	Under-sampling			0.891			0.952			0.837
	Hybrid-sampling			0.903			0.961			0.851
	Synthetic-sampling			0.803			0.748			0.696
SVM	**No Balancing**			0.809			0.921			0.731
	Under-sampling			0.803			0.904			0.750
	Hybrid-sampling			0.827			0.928			0.728
	Synthetic-sampling			0.765			0.919			0.700
GBM	**No Balancing**			0.842			0.910			0.798
	Under-sampling			0.853			0.927			0.802
	Hybrid-sampling			0.864			0.939			0.817
	Synthetic-sampling			0.759			0.834			0.689

Notes: The table reports the performance of our Cultural Distance, Account & Transaction-level model using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). The model has 17 predictors.

Table D7: Absolute Country-level Predictor Importance for Country & Account-level Model with Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
$\mathbf{CPI_S}$		4	5	5	***	8	8	8	***	9	9	9
$\mathbf{FSI_S}$		7	7	7	***	10	10	10	***	5	6	5
$\mathbf{IDV_S}$	***	3	1	2	***	3	3	3	***	4	3	4
$\mathbf{MAS_S}$	***	12	9	10	*	11	15	13	***	9	10	10
$\mathbf{PDI_S}$	***	5	10	8	***	6	7	7	***	10	14	11
$\mathbf{UAI_S}$	***	9	8	9	***	7	6	6	***	6	8	8
$\mathbf{CPI_R}$	***	8	6	6	***	4	2	4	***	11	5	6
$\mathbf{FSI_R}$	**	15	12	14	***	13	13	12		16	13	16
$\mathbf{IDV_R}$		6	4	4	***	5	5	5	***	13	15	14
$\mathbf{MAS_R}$	***	11	14	13		12	9	11	*	12	16	15
$\mathbf{PDI_R}$	***	13	11	11		9	14	9	***	15	11	12
$\mathbf{UAI_R}$	***	14	15	15	***	14	12	14	***	14	12	13

Notes: The table reports the importance of the Country-level predictors by absolute ranking for the Hybrid-sampled Country & Account-level model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 151,985 alerts (42,193 corporate-related and 109,792 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). The model has 16 predictors.

Table D8: Absolute Country-level Predictor Importance for Country, Account & Transaction-level Model with Cultural Distance and Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
CD	***	11	7	11	***	11	2	2	***	13	13	13
$\mathbf{CPI_S}$	***	8	6	5	***	16	10	12	***	12	6	9
$\mathbf{FSI_S}$	***	14	10	12	***	15	12	13	***	14	11	14
$\mathbf{IDV_S}$	***	5	1	1	***	5	14	10	*	9	10	11
$\mathbf{MAS_S}$	***	16	14	16	***	17	20	19	***	19	19	18
$\mathbf{PDI_S}$	***	9	15	14		12	19	16		16	17	17
$\mathbf{UAI_S}$	***	19	17	18	***	13	17	15	***	17	15	15
$\mathbf{CPI_R}$	***	17	5	10	***	6	11	9	***	15	2	5
$\mathbf{FSI_R}$	*	23	19	23	***	14	21	17	*	24	23	25
$\mathbf{IDV_R}$		13	4	6	***	7	1	1		21	25	22
$\mathbf{MAS_R}$	***	21	22	21	***	22	15	20	*	20	24	21
$\mathbf{PDI_R}$	***	20	21	20	*	10	16	11	***	25	20	25
$\mathbf{UAI_R}$	***	24	24	24	***	20	23	22	***	23	22	24

Notes: The table reports the importance of the Country, Account and Transaction-level predictors with Cultural Distance by absolute ranking for the Hybrid-sampled Cultural Distance, Account & Transaction-level model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). the model has 25 predictors.

Table D9: Absolute Country-level Predictor Importance for Cultural Distance, Account & Transaction-level Model with Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
CD	***	6	2	4	***	3	2	2	***	9	11	10
$\mathbf{CPI_S}$	*	4	6	6	·	4	6	5	***	7	6	7
$\mathbf{FSI_S}$	***	9	7	7	***	11	11	11	***	8	5	6
$\mathbf{CPI_R}$	***	7	1	1	***	1	1	1	***	11	2	5
$\mathbf{FSI_R}$	***	11	16	13	***	5	13	10	**	17	15	16

Notes: The table reports the importance of the Country-level predictors and Cultural Distance by absolute ranking fo r the Hybrid-sampled Cultural Distance, Account & Transaction-level model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). the model has 17 predictors.

Table D10: Relative Country-level Predictor Importance for Country & Account-level Model with Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
$\mathbf{CPI_S}$		2	3	3	***	6	6	6	***	4	5	5
$\mathbf{FSI_S}$		5	5	5	***	8	8	8	***	2	3	2
$\mathbf{IDV_S}$	***	1	1	1	***	1	2	1	***	1	1	1
$\mathbf{MAS_S}$	***	9	7	8	*	9	12	11	***	5	6	6
$\mathbf{PDI_S}$	***	3	8	6	***	4	5	5	***	6	10	7
$\mathbf{UAI_S}$	***	7	6	7	***	5	4	4	***	3	4	4
$\mathbf{CPI_R}$	***	6	4	4	***	2	1	2	***	7	2	3
$\mathbf{FSI_R}$	**	12	10	11	***	11	10	10		12	9	12
$\mathbf{IDV_R}$		4	2	2	***	3	3	3	***	9	11	10
$\mathbf{MAS_R}$	***	8	11	10		10	7	9	*	8	12	11
$\mathbf{PDI_R}$	***	10	9	9		7	11	7	***	11	7	8
$\mathbf{UAI_R}$	***	11	12	12	***	12	9	12	***	10	8	9

Notes: The table reports the importance of the Country-level predictors by relative ranking for the Hybrid-sampled Country & Account-level model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity.

Table D11: Relative Country-level Predictor Importance for Country, Account & Transaction-level Model with Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
$\mathbf{CPI_S}$		5	9	8	***	6	9	6	***	4	4	4
$\mathbf{FSI_S}$	***	2	3	3		8	5	8	***	2	3	3
$\mathbf{IDV_S}$		1	1	1	***	1	1	1	***	1	1	1
$\mathbf{MAS_S}$	***	7	5	6	***	9	10	9	***	3	7	6
$\mathbf{PDI_S}$	***	4	6	5	***	4	4	4	·	6	8	7
$\mathbf{UAI_S}$	***	6	7	7	***	3	3	3	·	5	5	5
$\mathbf{CPI_R}$	***	8	4	4	***	5	7	5	***	7	2	2
$\mathbf{FSI_R}$	***	10	8	10		10	11	11	**	11	10	12
$\mathbf{IDV_R}$	***	3	2	2		2	2	2	***	9	12	10
$\mathbf{MAS_R}$	*	9	11	9		11	8	10		8	11	9
$\mathbf{PDI_R}$	***	11	10	11		7	6	7	***	12	9	11
$\mathbf{UAI_R}$	***	12	12	12	***	12	12	12		10	6	8

Notes: The table reports the importance of the Country-level predictors by relative ranking for the Hybrid-sampled Country, Account & Transaction-level model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related).

Table D12: Cultural Distance Logistic Regression Coefficients

		Model 1		Model 2	
Model	**Balancing**	**Coefficient**	**Sig.**	**Coefficient**	**Sig.**
Combined	**No Balancing**	0.015	***	0.014	***
	Under-sampling	0.016	***	0.023	***
	Hybrid-sampling	0.014	***	0.020	***
	Synthetic-sampling	0.008	***	0.006	***
Corporate	**No Balancing**	0.021	***	0.020	***
	Under-sampling	0.016	***	0.032	***
	Hybrid-sampling	0.018	***	0.027	***
	Synthetic-sampling	0.010	***	0.009	***
People	**No Balancing**	0.008	***	0.005	·
	Under-sampling	0.010	***	0.023	**
	Hybrid-sampling	0.008	***	0.010	***
	Synthetic-sampling	0.005	***	0.004	***

Notes: The table reports the values of the coefficients in the Logistic Regressions for the Cultural Distance, Account & Transaction-level models (Model 1) and the Cultural Distance, Country, Account & Transaction-level models (Model 2). The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. Model 1 has 18 predictors and Model 2 has 25 predictors.

Table D13: Cultural Distance Logistic Regression Coefficients

		Model 1		Model 2	
Model	**Balancing**	**Coefficient**	**Sig.**	**Coefficient**	**Sig.**
Combined	**No Balancing**	0.015	***	0.014	***
	Under-sampling	0.016	***	0.023	***
	Hybrid-sampling	0.014	***	0.020	***
	Synthetic-sampling	0.008	***	0.006	***
Corporate	**No Balancing**	0.021	***	0.020	***
	Under-sampling	0.016	***	0.032	***
	Hybrid-sampling	0.018	***	0.027	***
	Synthetic-sampling	0.010	***	0.009	***
People	**No Balancing**	0.008	***	0.005	·
	Under-sampling	0.010	***	0.023	**
	Hybrid-sampling	0.008	***	0.010	***
	Synthetic-sampling	0.005	***	0.004	***

Notes: The table reports the values of the coefficients in the Logistic Regressions for the Cultural Distance, Account & Transaction-level models (Model 1) and the Cultural Distance, Country, Account & Transaction-level models (Model 2). The data sample comprises of 74,246 alerts (30,292 corporate-related and 43,954 people-related) with 1,172 Issue cases (524 corporate-related and 648 people-related). ***,**,* and · denote 0.1%, 1%, 5% and 10% levels of significance. Model 1 has 18 predictors and Model 2 has 25 predictors.

Table D14: Account & Transaction-level Models with CPI &FSI

		Combined			Corporate			People		
Model	**Balancing**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**	**TPR**	**FPR**	**AUC**
LR	**No Balancing**			0.705			0.793			0.706
	Under-sampling			0.709			0.797			0.728
	Hybrid-sampling			0.708			0.801			0.732
	Synthetic-sampling			0.702			0.784			0.707
RF	**No Balancing**			0.861			0.959			0.925
	Under-sampling			0.857			0.951			0.903
	Hybrid-sampling			0.861			0.967			0.912
	Synthetic-sampling			0.702			0.761			0.658
SVM	**No Balancing**			0.697			0.863			0.678
	Under-sampling			0.779			0.848			0.774
	Hybrid-sampling			0.801			0.893			0.821
	Synthetic-sampling			0.678			0.771			0.631
GBM	**No Balancing**			0.806			0.863			0.803
	Under-sampling			0.820			0.864			0.828
	Hybrid-sampling			0.822			0.879			0.823
	Synthetic-sampling			0.694			0.766			0.639

Notes: The table reports the performance of our Account & Transaction-level model with CPI & FSI using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 146,084 alerts (37,562 corporate-related and 107,024 people-related) with 2,081 Issue cases (832 corporate-related and 1,229 people-related). The model has 14 predictors.

Table D15: Schwarz Indices Models

		Combined			Corporate			People		
Model	Balancing	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
LR	**No Balancing**			0.638			0.722			0.653
	Under-sampling			0.631			0.750			0.645
	Hybrid-sampling			0.627			0.749			0.653
	Synthetic-sampling			0.633			0.731			0.651
RF	**No Balancing**			0.519			0.518			0.531
	Under-sampling			0.735			0.815			0.712
	Hybrid-sampling			0.721			0.800			0.703
	Synthetic-sampling			0.680			0.745			0.674
SVM	**No Balancing**			0.557			0.528			0.576
	Under-sampling			0.710			0.785			0.711
	Hybrid-sampling			0.667			0.770			0.651
	Synthetic-sampling			0.567			0.663			0.549
GBM	**No Balancing**			0.744			0.805			0.741
	Under-sampling			0.745			0.818			0.734
	Hybrid-sampling			0.755			0.823			0.752
	Synthetic-sampling			0.669			0.745			0.669

Notes: The table reports the performance of our Schwarz Indices model with Cultural Distance using logistic regression (LR), random forest (RF), support vector machine (SVM) and gradient boosting (GBM) in combination with no balancing, under-sampling, hybrid-sampling and synthetic-sampling, respectively. The performance is measured using True Positive Rate (TP Rate), False Positive Rate (FP Rate) and Area under the ROC Curve (AUC). The data sample comprises of 58,447 alerts (23,449 corporate-related and 34,998 people-related) with 783 Issue cases (270 corporate-related and 513 people-related). The model has 6 predictors.

Table D16: Predictor Importance for Schwarz Indices Model with Hybrid-sampling

	Combined				Corporate				People			
Predictor	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**	**LR**	**RF**	**GBM**	**Ave.**
$\mathbf{EBD_S}$	***	2	2	2	***	2	2	2	***	2	2	2
$\mathbf{EGA_S}$		5	5	5	***	6	4	4	**	3	3	3
$\mathbf{HIE_S}$	***	1	3	3	***	4	3	3	***	1	1	1
$\mathbf{EBD_R}$	***	3	1	1	***	1	1	1	***	6	6	6
$\mathbf{EGA_R}$	.	6	6	6	***	5	6	6	***	5	5	5
$\mathbf{HIE_R}$	***	4	4	4	***	3	5	5	***	4	4	4

Notes: The table reports the importance of the predictors for the Hybrid-sampled Schwarz Indices model applied to the full sample (combined) and its partitions (Corporate & People accounts). Estimates of importance are obtained from the logistic regression (LR), random forest (RF), gradient boosted model (GBM) algorithms. A weighted average of RF and GBM (Ave.) is included. For LR, ***,**,* and • denote 0.1%, 1%, 5% and 10% levels of significance. RF and GBM are both tree-based algorithms and so their estimates are based on the mean decrease in the Gini index of each node across all trees. The Gini index measures node impurity. The data sample comprises of 58,447 alerts (23,449 corporate-related and 34,998 people-related) with 783 Issue cases (270 corporate-related and 513 people-related). The model has 6 predictors.