Detecting Fraud, Corruption, and Collusion in International Development Contracts: The Design of a Proof-of-Concept Automated System

Emily Grace
Princeton University
Jadwin Hall
Princeton, NJ 08544
Email: emily.grace.eg@gmail.com

Ankit Rai
University of Illinois at Urbana Champaign
NCSA
Urbana, IL 61801, USA
Email: rai5@illinois.edu

Elissa Redmiles
University of Maryland
AVWilliams Suite 3400
College Park, MD 20740
Email: eredmiles@cs.umd.edu

Rayid Ghani
University of Chicago
1155 E. 60th Street
Chicago, IL 60637
Email: rayid@uchicago.edu

***Abstract*—International development banks provide low-interest loans to developing countries in an effort to stimulate social and economic development. These loans support key infrastructure projects including the building of roads, schools, and hospitals. However, despite the best efforts of development banks, these loan funds are often lost to fraud, corruption, and collusion. In an effort to sanction and deter this wrongdoing and to ensure proper use of funds, development banks conduct extensive, costly investigations that can take over a year to complete.**

This paper describes a proof-of-concept of a fully automated fraud, corruption, and collusion classification system for identifying risk in international development contracts. We developed this system in conjunction with the World Bank Group - the largest international development bank - to improve the time and cost efficiency of their investigation process. Using historical monetary award data and past investigation outcomes, our classifier assigns a "risk score" to World Bank contracts. This risk score is designed to enable World Bank investigators to identify the contracts most likely to lead to a substantiated investigation. If implemented, our automated system is predicted to successfully identify fraud, corruption, and collusion in 70% of cases.

I. Introduction

International development banks provide low-interest loans to developing countries in order to reduce poverty and encourage social and economic development [16]. The World Bank Group, the largest international development bank, awards about 20-30,000 contracts annually, which are worth over $60 billion [7]. This funding supports development in areas such as education, health, and agriculture in order to improve the quality of living in impoverished and middle income countries [4]. However, these large monetary loans and the sometimes unstable governing context in these countries come with the potential for fraud, corruption, and collusion [5], [6].

Despite the best efforts of development banks to deter and prevent wrongdoing, more than 5% of the world GDP ($2.6 trillion) is lost annually to fraud, corruption, and collusion [14]. Corruption is the biggest impediment to economic growth in more than 60 countries. As a result, the countries that most need development funding are often less likely to receive it [14]. Additionally, when money is lost through illicit flows, it often funds major crimes, such as those related to drugs and human trafficking [6].

In this paper, we present a proof-of-concept for a fully automated fraud, corruption, and collusion classification system for international development contracts. While prior work has focused primarily on detecting credit-card fraud [8], [10], [20], [15], our work is more broad, focused on identifying not only fraud but also corruption and collusion, which are often harder to detect. Our system, informed by input from experienced World Bank investigators, links together two decades of World Bank investigative and contract award data to rank the allegations of wrongdoing most likely to be substantiated and proactively identify the contracts most likely to be tainted with corrupt practices. Our system is predicted to have a 70% success rate in predicting allegations that will be substantiated, an 84% increase from the current investigation success rate. Our system is fully automated from data collection through pre-processing and modeling.

II. Current Approach

The World Bank provides loans to developing countries for infrastructure and development projects. When the World Bank approves a funding request, a particular implementing agency within the country, such as the Ministry of Finance, is designated to coordinate the bidding and contract selection process. The first step in this process is for the project-implementing agency to post a Request for Proposal (RFP) seeking suppliers with the skills necessary to complete the work. Project funds are typically split among multiple contracts [4]. For example,

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

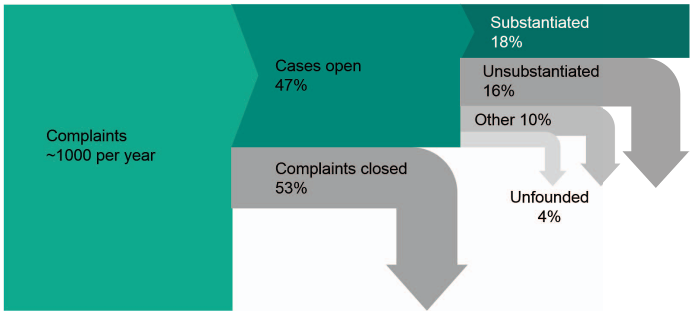

Fig. 1. Prevalence of outcomes in the World Bank investigative process. Figure courtesy of Frances Gagnon.

in the case of a road infrastructure project, there may be separate contracts for engineering work, provision of cement, and provision of construction equipment with each of of these contracts having an RFP.

The procurement process can vary among different contracts and proposals. Several different procurement methods exist, such as "International Competitive Bidding", which allows international advertising of the contract. In addition, each RFP designates a bid evaluation method outlining the criteria that will be used for selection. For example, one evaluation method relies primarily on the criteria of cost while another, "Quality- and Cost-Based Selection", has evaluation criteria based both on the technical merit and on the cost of the bid [2]. The country's implementing agency then evaluates the bids and selects the winning contract.

Fraud, corruption, and collusion can occur at any point in this process including during the bidding and contract selection processes as well as during the project implementation phase. Per the World Bank guidelines, a fraudulent practice is "any act or omission, including a misrepresentation, that knowingly or recklessly misleads, or attempts to mislead, a party to obtain a financial or other benefit or to avoid an obligation" [3]. A corrupt practice is "the offering, giving, receiving or soliciting, directly or indirectly, anything of value to influence improperly the actions of another party." A collusive practice "is an arrangement between two or more parties designed to achieve an improper purpose."

To discover fraud, corruption, and collusion, the World Bank currently uses a complaint-based investigation process. This process begins when someone who knows of or suspects wrongdoing submits a complaint to the World Bank. The World Bank's Integrity Vice Presidency (INT) then enters this complaint into the investigation database. World Bank INT investigators evaluate these complaints to determine if there is enough evidence to merit a case, a process that can take a few months from the date the complaint is received. If further attention is warranted, the World Bank INT opens a case on this complaint and thoroughly investigates the allegations. These investigations can take over a year to complete and include a full review of all documents as well as site visits by investigators to conduct interviews, view project sites, and collect evidence.

An investigation ends with the determination of an allegation outcome, which is entered into the World Bank INT investigations database. Possible allegation outcomes include an **unsubstantiated allegation**, which indicates that not enough evidence was found to sustain or refute the allegations; an **unfounded allegation**, which indicates that there was no evidence of wrongdoing and the supplier has therefore been cleared; **referral** to a different development bank or sanctioning body for investigation; or a **substantiated allegation**, which indicates that evidence of wrongdoing was found. A substantiated allegation can lead to several results, including a settlement or a sanction against the supplier such as debarment from working with the World Bank and other development banks for a period of time. Figure 1 shows the prevalence of these outcomes.

A. *Problems with Current Approach*

There are two main problems with the current approach:

1) The current process is reactive and complaint-driven.
2) Fewer than half of the investigations result in substantiated outcomes.

The current investigation process is based exclusively on complaints filed by individuals who may have knowledge of wrongdoing. Given cultural and political differences across the different geographic regions served by the World Bank, contracts in certain regions receive more complaints than those in others. Using the current approach, if no complaint is made, no investigation will take place. Thus, there is no process in place to investigate or assess the risk of projects about which there have been no complaints. In addition to being reactive, this process also leads to few or no investigations in certain regions, despite investigator suspicions that wrongdoing is present yet unreported.

III. Our Solution

Our machine learning system would make three primary improvements to the current investigative process. First, it enables the addition of proactive investigations to the current, primarily complaint-driven, process. Second, this classification system increases time and cost efficiency by providing an automated mechanism that can be used to prioritize complaints for investigators. Third, it is predicted to improve the rate of substantiated investigations by 84%.

The solution we developed to identify fraudulent, corrupt, and collusive contracts uses two primary data sources: past investigatory outcomes and a database of World Bank contract awards that meet a set of value thresholds. Using this data, which is described in more detail in Section IV, we built a classifier to assign a "risk score" to World Bank contracts. This risk score allows investigators to prioritize which contracts to investigate and avoid allocating resources to cases that are unlikely to be fruitful. Further, our classification system enables intelligent selection of contracts, especially those in countries where few complaints are received, for proactive inquiry.

Fig. 2. Contracts awarded by the World Bank per country from 2000-2014.

Fig. 3. Heat map of allegation categories over allegation outcomes for World Bank investigations from 2011-2014 [6].

This classification system consists of the following components, which are described in further detail in the remainder of the paper:

1) Data ETL
2) Feature Generation
3) Modeling
4) Evaluation

This entire pipeline is fully automated as described in detail in Section VI.

IV. Data Sources

The data sets used in this analysis include the World Bank contracts data set, the World Bank INT investigations data set, and two monetary conversion data sets.

The contracts data set contains records for ∼200,000 World Bank contracts awarded since 2000 [9]. The information in this data set includes the monetary amount of the contract, the geographic region of the project, the name of the supplier, and the sector of the project. This data is entered by World Bank field agents in each country and is routinely validated.

In this data set, funding was awarded to projects conducted in 168 countries and for contract work by suppliers who were based in 198 different countries. The geographic distribution of these contracts is shown in Figure 2.

In addition to the contracts data set, we used a data set containing records of 4,045 World Bank INT investigations conducted since 2000. The investigations data set is a confidential data set that is maintained by the World Bank INT, the group that is responsible for conducting all investigations on World Bank contracts. Fields in this data set include allegation category (e.g. fraud, corruption, collusion) and investigation outcome (e.g. substantiated, unfounded, unsubstantiated). The number of each allegation type that resulted in each investigative outcome is shown in Figure 3 for the subset of the data which was matched to relevant contracts. The allegation outcomes from this data set were used as the training labels for our predictive model while the category of the allegation was used a feature in that model.

In addition to these data sets, we used the FCRF official exchange rate data set produced by the International Monetary Fund and the purchasing power parity (PPP) conversion factor data set produced by the World Bank [13], [12]. The FCRF data set provides the official exchange rate from the contract award amounts in U.S. dollars to the value in the local currency as determined by national authorities or the legally sanctioned exchange market [13]. The PPP data set provides data regarding how much of that local currency would be required to purchase the same goods and services domestically as the U.S. dollar would buy in the United States [12].

In combination, we used these two data sets to normalize contract award amounts with regard to yearly inflation and local purchasing power. Both the monetary value of a contract and its supplier's total historical award amount from the World Bank were features included in our predictive model. Therefore, it was necessary to normalize monetary values across localities and years.

V. Classification Pipeline

The full classification pipeline used to rank World Bank contracts based on the likelihood of being associated with sanctionable practices is shown in Figure 4. Individual components of the pipeline are described in Sections V-A, V-B, and V-C. The tools used to develop this pipeline were PostgreSQL and Python. PostgreSQL was used to store and aggregate our data sets. Data pre-processing and feature generation were done in Python. Modeling was run using the Python scikit-learn module [17]. The source code for this project is publicly available on our GitHub repository [11]. The full pipeline is automated, and the process and challenges of this automation are addressed in Section VI. Finally, we developed a prototype for a dashboard that could be used to display the

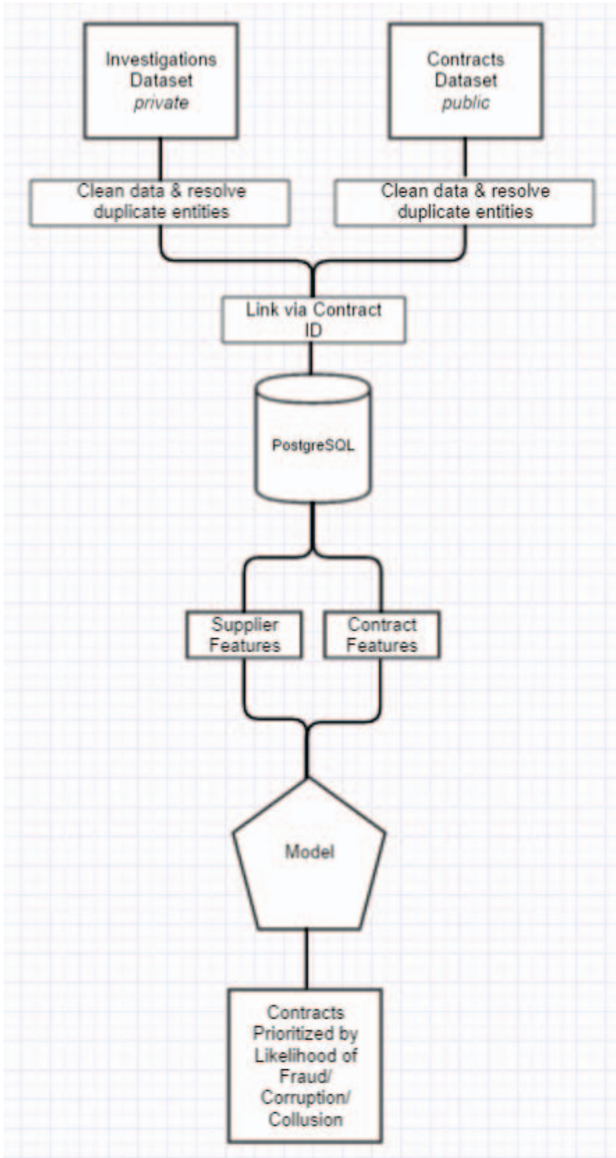

Fig. 4. The classification pipeline.

results of our modeling pipeline; our design process and output are described in Section VII.

A. Data ETL

The contracts data set and investigations data set that were used for this project came from separate systems within the World Bank. Both data sets were composed of data manually entered by World Bank professionals across the world over the past two decades. Due to discrepancies in date formats, differing notations for missing data, and non-standard naming of similar data fields, extensive transformation was required to merge the data. This transformation consisted of two steps:

1) Data standardization
2) Entity resolution

Step 1 includes standardizing date entries, de-duplication, and handling of missing values via a straightforward Python script. Step 2 is less straightforward, however, and merits further discussion.

The contracts and investigations data we extracted from the World Bank contained 73,858 unique supplier names. Many of these company names refer to the same entity (e.g. Panda Thames Central [1], PTC, PandaThames-Central, etc.). As discussed in Section V-B, features of these suppliers are important in our data analysis. Thus, it is necessary that features of PTC be shared with Panda Thames Central, as they represent the same entity and features indicating wrongdoing on the part of PTC should also indicate wrongdoing on the part of Panda Thames Central. For this reason, it is important for us resolve these suppliers to one common supplier name. To do so, we directly collaborated with researchers at the University of Cincinnati who are developing an entity resolution tool for this use case as part of a separate project at the World Bank. We used this tool to generate an entity resolution list for the entities in the data sets [18].

After cleaning and resolving entities, we linked the investigations and contracts data sets together using the World Bank contract ID, a unique identifier assigned to every contract when it is awarded. Investigations are always completed with regard to one or more contracts, and thus each investigation record should be linked to at least one contract ID. However, due to data entry discrepancies, not all investigations were tagged with a valid contract ID. After the linking process, our final labeled dataset contained 600 investigations linked to one or more contracts.

B. Feature Generation

After cleaning and linking the contracts and investigations data sets, we generated contract- and supplier-related features for use in the predictive model. All of the contract-related features are summarized in Table I. The most basic contract-related features include the country in which the contract was awarded and the type of bidding or selection process that was used in the award. To develop more complex contract- and supplier-related features, we had numerous discussions with investigators at the World Bank who investigate cases and manually determine which complaints merit investigation as full cases. These investigators shared with us a number of indicators of fraud, corruption, and collusion that they have identified through their experience. In the remainder of this section, we will present examples of features developed based on the investigators' insights.

Contractors engaged in wrongdoing may bribe or collude with other contractors in an effort to solicit an unusually high proportion of the money awarded to a given project. Given this information, we included several monetary contract-related features including the total financial amount awarded to the supplier (PPP corrected [12]), the total cost of the contract's parent development project (also PPP corrected), and the percentage this contract contributes to that total project cost.

[1]This is a fictional company name used only as an illustrative example.

Feature	Description
country	country in which contract work was completed
region	region in which contract work was completed
bidding_process	the type of selection process used
project_amount	total cost of project
contract_amount	total contract funds awarded
percent_of_project	percentage of project funds awarded to contract
date_diff	time elapsed between contract award date and contract signing date
num_supp_awards	number of additional monetary awards required to complete work

TABLE I
CONTRACT-RELATED FEATURES

Named Historical Features
Percent of previous contracts in the past year in the agricultural sector
Percent of previous contracts in the last 3 years in Brazil
Percent of all previous contracts awarded for medical equipment
Ranked Historical Features
Percent of previous contracts in last 5 years in the supplier's most common sector
Percent of previous contracts in the past year in the supplier's third most common country
Percent of all previous contracts in the supplier's fifth most common procurement type

TABLE II
SELECTED EXAMPLES OF SUPPLIER RELATED FEATURES. IN TOTAL AROUND 4000 SUCH FEATURES WERE INCLUDED IN THE MODEL.

Another potential indicator is the number of supplemental requests for funds made by a supplier after a contract has been awarded. A supplier may have submitted an artificially low bid in order to win the contract, and then made up for this low bid by continuously requesting additional funds. Thus, we included the number of supplemental awards given to the supplier as a contract-related feature in our model.

Further, the investigators suggested that when a long period of time elapses between the official awarding of a contract and the actual commencement of work by the supplier it could be indicative of illicit deals and negotiations. Consequently, we included additional contract-related temporal features such as the time elapsed between the date the contract was awarded and the date that the supplier officially began work.

Yet another potential indicator is when suppliers begin to provide services outside their usual domain. For instance, a given supplier, which has predominantly worked in the agricultural sector, may abruptly begin working on contracts in the education sector. This type of switching may be the result of a change of sectors by a government official previously bribed by the supplier. The switching may also indicate that a shell company under the control or influence of a public official is being used to win contracts.

Building upon this information, we generated a variety of supplier-related features to characterize the historical behavior of each supplier within the World Bank contracting system. Examples of these types of features can be seen in Table II. At the time that a supplier is awarded a new contract, we consider the types of contracts that supplier has previously worked on. For example, one specific feature details the percentage of a supplier's contracts in the past 3 years that were in Africa, while another quantifies the percentage of the supplier's previous contracts that were related to the transportation sector [2]. We refer to these features as historical features. In total, the following five categorical variables were considered when generating these supplier-related features:

1) Country of previous contracts
2) Region of previous contracts
3) Major Sector of previous contracts (e.g. agriculture, transportation, etc.)
4) Procurement Type of previous contracts (e.g. infrastructure, medical equipment, maintenance, etc.)
5) Procurement Category of previous contracts (e.g. goods, consultant services, or civil works)

In addition, to capture how varied the supplier's previous contracts were, without specifying exactly what type of contracts they were awarded, an additional set of aggregated features was generated. We refer to these as "ranked features." As an example, one set of these features enumerates the percentage of a supplier's previous contracts that are in the country in which it most commonly works, its second most common country, etc. This set of features was intended to capture a supplier's level of international operation and its level of work specialization, without designating the specific countries or categories in which it worked. A very specialized supplier who worked only in a single sector would have 100% of their contracts in their "first most common" or top sector while a supplier whose work was more distributed might have 50% of their previous contracts in their top sector, 30% in their second most common sector, and 20% in their "third" sector. Our reasoning for eliminating country specific information from our model is described in more detail in Section V-C2.

To capture changes in the supplier behavior over time, these types of aggregations were performed over 1, 3, and 5 year time periods as well as over the full history of the supplier. Additionally, for all of these aggregations, two full sets of features were created: one where the percentages

[2]These examples are used merely for illustration and are not necessarily indicative of wrongdoing.

were based upon total money awarded and the second based upon raw contract counts. Generating all combinations of the different categorical variables, the different aggregation time periods, the named and ranked features, and the amount/count variation, resulted in approximately 4,000 historical supplier features.

C. Modeling

The features described in Section V-B were used as input to a supervised machine learning pipeline designed to predict the likelihood of substantiating a claim of fraud, collusion, or corruption. The model was trained on past investigations data, where the outcome of the investigation was used as the training label. A substantiated allegation was considered a positive outcome, while an unsubstantiated or unfounded allegation was considered a negative outcome.

To select the best model for this use case, we evaluated the performance of different modeling techniques. The models evaluated include random forests, logistic regression, gradient boosting, and k-nearest neighbors classifiers. The parameters of these models were optimized over. The models and tuning parameters that were evaluated are summarized in Table III.

Additionally, different combinations of features were tested in the model. This was done to check for any reduction in model performance from including certain features, an indication of over-fitting. The features were split into groups with the aim of evaluating the impact of the following groups of features:

1) **Country specific features**
The World Bank wants to target corruption in all countries and regions rather than simply providing extra scrutiny to countries that are already known for widespread corruption. Therefore, models that did not take into account the specific country of a contract or company were considered.
2) **Aggregated contract count and amount feature**
Features were generated that aggregated a supplier's previous behavior over both its number of contracts and the amount of money awarded. In order to evaluate the relative important of the number of contracts vs. the amount of the contracts, different feature sets were used with only one or the other.
3) **Aggregation time period**
Aggregated supplier histories were generated for different time periods prior to each contract of interest - 1 year, 3 years, 5 years, and the full supplier history. In order to investigate which time periods contained relevant information for detecting patterns of corruption, different feature sets were used which contained each of these aggregation time periods on its own.

For each combination of a specific model method, set of model parameters, and feature set, the model performance was validated using the metrics described in Section V-C1. In order to iterate over all possible combinations of feature sets, models, model parameters, and feature sets a grid search optimization was performed as highlighted in the pipeline structure shown in Code Snippet 1. On the basis of this evaluation, we determined that the optimal classifier was a Gradient Boosting Classifier, with 500 estimators, a maximum depth of 160, a minimum sample split of 15, and a learning rate of 0.1. Further details on this model and selection are provided in Section V-C2.

```
data_splits=[2008,2009,2010,2011,2012,2013,2014]
features_sets=[[set1_cols],[set2_cols],...]
models=[RandomForestClassifier(), AdaBoostClassifier(),...]
param_sets=[[n_classify=100,max_depth=40],...]

for train_data, test_data in data_splits:
        for feature_set in feature_sets:
                for model in base_models:
                        for param_set in param_sets:

                                fit_model(training_data)
                                predict_model(testing_data)
                                evaluate_model(testing_data)
```

Code 1: Pseudocode for modeling selection and evaluation.

1) Evaluation: To evaluate the performance of the different modeling techniques described in Section V-C, we used a temporal validation strategy that simulates how our model would be deployed and used by the World Bank.

We created subsets of our data from date X to date Y, called training sets. We then trained each of our models on these training data sets and evaluated the performance of our model on a series of test sets, which are subsets of data from date Y to some future date Z with known outcomes. We then compared our model's predictions with the known outcomes of the test data. In our evaluation procedure, a total of thirteen such temporal validation splits were used, beginning with a training set containing contracts from the 2008 calendar year and a test set containing contracts from the 2009 calendar year. Each test set contains one year of data after the end of the training period. This allowed the simulation of the effect of using each model for prediction at different points in the past and the evaluation of the performance of the model on the known results of the investigations. See Table IV for each of the 13 train-test splits used to evaluate the model.

Although older data was available for training and testing, we used data from the past six years in our evaluation because it is most reflective of the current international development and fraud climate. Thus, this more recent data provides a more accurate projection of the performance of our model when applied to the World Bank use case.

Evaluation Metric: Precision. The goal of this system is to enable the World Bank INT to be proactive in their investigations as well as to use their resources efficiently. Therefore, the goal of the predictive model developed for this project is to increase the efficiency of the investigation process and focus resources on complaints or contracts that are most likely to result in a substantiated allegation. By investigating complaints or contracts that are most likely to be substantiated, the World Bank will be able to take action against more suppliers in an effort to eradicate more wrongdoing.

Since we were interested in identifying contracts or complaints that were *most* likely to be substantiated the best metric

Classifier Type	Parameter	Values
RandomForestClassifier	n_estimators	500, 1000
	max_depth	40, 80, 160, 500, 1000
	min_samples_split	2, 5, 10
LogisticRegression	C	0.1, 0.5, 1.0
AdaBoostClassifier	n_estimators	500, 1000
	learning_rate	0.1, 0.5, 0.75, 1.0
SVC	C	0.1, 0.5, 1.0
	kernel	linear, rbf
GradientBoostingClassifier	n_estimators	**500**, 1000
	max_depth	40, 80, **160**, 500
	min_samples_split	2, 5, 10, **15**
	learning_rate	**0.1**, 0.5, 1.0
KNeighborsClassifier	n_neighbors	3, 5, 7, 11, 13, 15, 17, 19

TABLE III

MODELS AND PARAMETER SPACE EXPLORED TO FIND THE MODEL WITH THE OPTIMAL PERFORMANCE. THE BEST MODEL FOR OUR USE CASE, THE GRADIENT BOOSTING CLASSIFIER, AND ITS PARAMETERS ARE HIGHLIGHTED IN BOLD.

Train End	Test Range	Train Size	Test Size
12/31/07	1/1/08 - 12/31/08	59	44
06/31/08	7/1/08 - 06/31/09	73	36
12/31/08	1/1/09 - 12/31/09	103	16
06/31/09	7/1/09 - 06/31/10	109	15
12/31/09	1/1/10 - 12/31/10	119	13
06/31/10	7/1/10 - 06/31/11	124	17
12/31/10	1/1/11 - 12/31/11	132	22
06/31/11	7/1/11 - 06/31/12	141	24
12/31/11	1/1/12 - 12/31/12	154	36
06/31/12	7/1/12 - 06/31/13	165	42
12/31/12	1/1/13 - 12/31/13	190	34
06/31/13	7/1/13 - 06/31/14	207	20
12/31/13	1/1/14 - 12/31/14	224	6

TABLE IV

MODEL EVALUATION TRAIN/TEST SPLITS

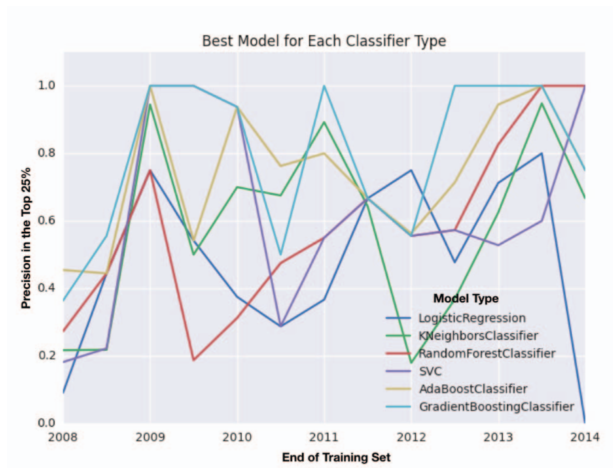

Fig. 5. The maximum precision achieved in the top 25% of allegations for different model types across the different test/train splits. The gradient boosting classifier achieved the most consistent performance.

for evaluating the performance of the model was precision. More specifically, we evaluated models based on the precision of the predictions in the upper portion of the ranked list. The purpose of this evaluation metric is to take the fixed number of investigation resources the World Bank has, and maximize the number of substantiated cases. In order to select the highest performing model from the full set described above, the metrics of precision in the top 10%, 25%, and 50% were considered.

2) Final Model: Based on precision at those levels, the best performing model across all train and test sets was the Gradient Boosting Classifier. In particular, model parameters from the scikit-learn implementation of the model of n_estimators = 500, max_depth = 160, min_sample_split = 15, and learning rate = 500 achieved the best performance by the metric of precision in the top 25%. A comparison of the best performing model of each type can be seen in Figure 5. Averaging over the simulated past performance of the best-performing model, we can expect around an 80% substantiation rate in the top 25% of the complaint ranking and a 65% substantiation rate in the top 50% of the complaint ranking. The 25% and 50% precision

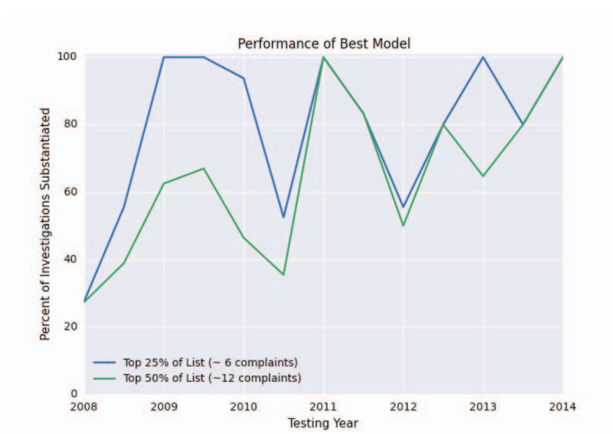

Fig. 6. The success rate of targeting substantiated cases the best performing model would have achieved in each year in the past if the top 25% and 50% of the recommendations were investigated.

levels for each training/test set can be seen in Figure 6.

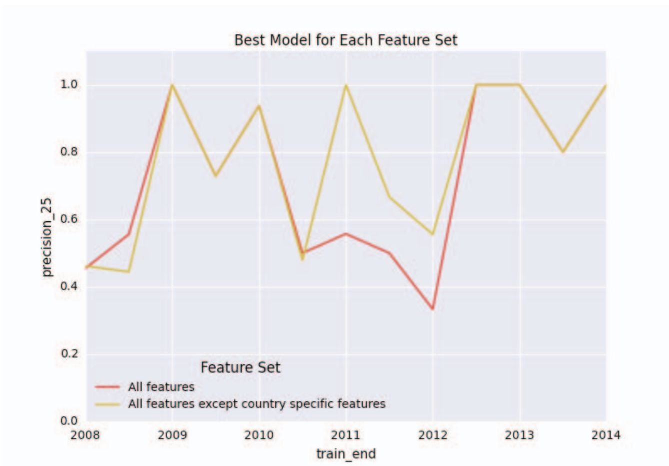

Fig. 7. A comparison of the success rate of substantiation between a model which included country specific information and one which did not. The performance of the two models was largely similar, except for a period of several years where using the country specific information made the model worse due to over-fitting.

High variability was observed in the performance of each specific model across the different training and test sets. The origin of this high variability can be attributed to the small size of the test set data splits. Improved evaluation of the model performance could be achieved with a larger set of investigation data. The average precision of the final model in the top 25% of recommendations from the ranked list over all test sets is 70%. Therefore, we predict that use of this model to select complaints to investigate by taking the top 25% of recommendations would result in a 70% rate of substantiated investigation outcome.

3) Feature Importance: Both the basic contract description features and the aggregated supplier features achieved high feature importance scores from the classification model. However, within these categories, we analyzed the relative importance of including country-specific features, the length of supplier history, and the number vs. monetary value of contracts awarded.

Country-Specific. One of the top priorities of the World Bank is to target wrongdoing across all countries. Therefore one of the evaluation criteria for model performance was the extent to which country was being relied on as a predictive factor. We found that models that included certain country-specific features produced results that were biased against specific countries. In these models, there were many countries in which almost all contracts were predicted to be involved in wrongdoing while in other countries almost no contracts were flagged. This issue was largely improved by removing country-specific information from the model - including the country of the borrower, the supplier, and the countries in which the supplier had worked previously. In addition to providing a less biased model, the non-country specific model performs as well or better than the full model in terms of precision. A

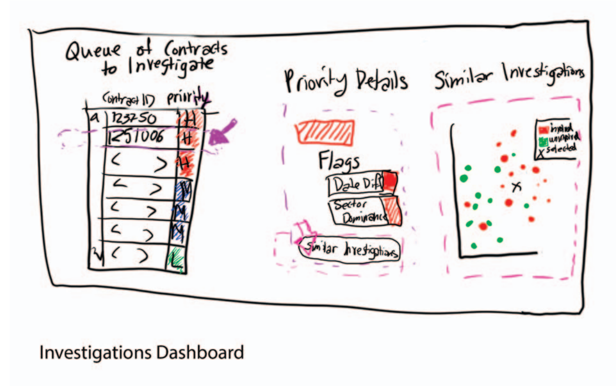

Fig. 8. Stage one sketch used in participatory design of the analytics dashboard used to display the classification results to investigators.

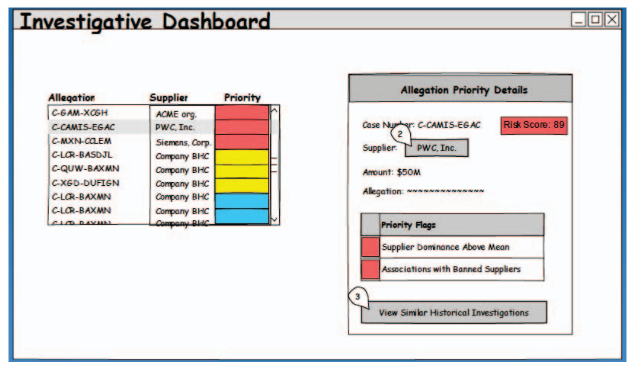

Fig. 9. Stage two clickable wireframe of the analytics dashboard used to display the classification results to investigators.

comparison of the performance of the two models can be seen in Figure 7.

Contract Counts vs. Amounts. Substituting a feature set containing aggregation by contract counts instead of by contract amounts had no impact on model performance. This suggests that the historical monetary amounts of the contracts awarded to the supplier do not add predictive information compared with the simple contract counts.

Supplier History. Similarly, varying the length of supplier history supplied to the model (1, 3 or 5 years) did not significantly impact the model performance. This may suggest that the most recent history of the supplier has the highest relevance to predicting current involvement in wrongdoing.

VI. Automation

The World Bank INT is interested in determining the probability that complaints will be substantiated, as well as determining the probability that contracts which have not been reported are tainted with fraud or corruption. To provide both of these insights, our classifier outputs multiple ranked lists. The first list ranks the fraud and corruption complaints received by the World Bank INT in order of the likelihood of substantiation. Another set of lists is produced to rank all contracts awarded by the World Bank, not just those that have received a specific complaint. Finally, separate lists are

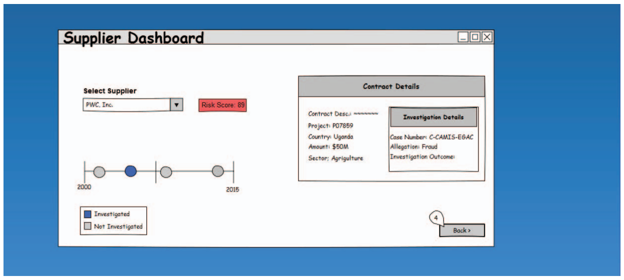

Fig. 10. Stage two clickable wireframe of the analytics dashboard used to display the classification results to investigators.

produced for each category of wrongdoing, such that one list presents a ranking of likelihood to be involved in corruption while another presents a ranking of the likelihood to be involved in fraud, etc.

We automated the generation of these lists from the beginning to the end of our classification pipeline: extracting the data, transforming and cleaning it, generating features, loading the prepared data to the database, and, finally, modeling. Using a BASH script, we first extract publicly available, up-to-date World Bank contracts data from the web and internally extract the internal investigations data set. Next, the script runs these files through a set of resilient data transformation scripts, which are designed to automatically handle variations in date format, convert all currencies to USD using the process described in Section IV, and perform other cleaning steps. One of the challenges in this automation was developing scripts that were robust enough to handle data format changes. In order to test the resilience of our scripts, we ran archived data sets from the past 10 years at the Bank through the pipeline to verify that changes in data format did not lead to fatal errors.

The transformed data is then loaded into a PostgreSQL database and another set of scripts is used to generate the features, which are loaded to the database. Finally, this database is used as an input to the modeling script that is used to output the set of ranked lists described above. If a failure occurs at any point in the process, the pipeline will post an alert in the log. Additionally, the automated pipeline can be easily run through a single click desktop application.

VII. Dashboard Design

In order to make the output of our model more actionable, and to extract meaning from the features that we created, we developed a prototype of an investigative dashboard with the World Bank. We used a participatory design process [19] to co-design this dashboard. First, we drew a rudimentary sketch (Figure 8), and asked our World Bank partners to erase and redraw portions of that sketch while walking us through their thought process. Next, we constructed a clickable wireframe using Pencil [1], see Figures 9 and 10. The final dashboard design included the following elements: displaying a ranked list of contracts (Figure 9), displaying features of a given contract that influenced its risk score (Figure 9) and showing a time-line of a given supplier, which includes in which industries they held contracts, whether allegations were lodged about those contracts, and if the contracts were investigated, the investigation outcomes (Figure 10). The goal of developing this dashboard prototype was to illustrate how the results of our machine learning model and the insights gained from linking together disparate World Bank data sets could be easily integrated into the investigative work-flow.

VIII. Future Work

While the proof-of-concept predictive model described above has shown promising results using historical test data, there remain a number of potential avenues for improved performance and evaluation of the classification system.

The entity resolution work used in our model is still in development [18]. Once a more complete and validated version of this tool is available, World Bank INT will be able to update the pipeline to use this new tool. The new tool will be able to more completely resolve the 74,000 entities in the data - thus enabling better cross-linkage of features across entities. Current entity resolution errors may result in some entities not being linked to all relevant historical features.

In our data sets, there were a large number of open-ended text data fields. These text fields contain detailed investigation reports and the text originally submitted by the individual entering an allegation. Although we performed preliminary topic modeling and clustering on this text data, we did not generate any features based on this data nor pursue this angle of analysis in-depth, given our large quantity of other features. The World Bank INT has plans to create more close-ended data fields in the future, and is also interested in pursuing a more in-depth analysis of the existing text data. This data could fuel additional productive features that would improve the precision of our classification system.

Further, we had approximately 200,000 contracts and 4,000 original investigations data points on which to train and evaluate our models. Unfortunately, when linking the investigations data set to the contracts data set, we discovered that we were only able to link 600 investigations to at least one contract. Thus, additional data from other international development banks or World Bank data with more consistent tracking of contract IDs would enable the building of an even better and more generalize-able classification pipeline for fraud, corruption, and collusion in development contracts. The World Bank INT has plans to implement several data collection improvements, including better linking of investigations entries to contract IDs.

Finally, there are rarely enough resources to investigate all of the cases that the model identifies as likely to be substantiated. In order to cull down this list even further, threat assessments, risk analyses, and additional human investigator input should be used to develop additional features that could be integrated into our existing model. By developing metrics that speak to the cost-benefit of investigating cases, we could better optimize our suggestions for the use of investigative resources, thus improving the ROI of our classifier.

IX. ORGANIZATIONAL IMPACTS

This proof-of-concept system is the first application, to our knowledge, of machine learning to identify fraud, corruption, and collusion in contracts financed by an international development organization. The development of this system has spurred conversations among staff charged with monitoring integrity in the World Bank regarding the development of new tools and dashboards to leverage machine learning in integrity processes.

X. SUMMARY

Fraud, corruption, and collusion in contracts financed by international development organizations drain needed funds from developing countries, with the greatest impacts disproportionately affecting the poor. This wrongdoing channels money intended to improve the quality of life toward illicit activities [6]. Thus, the World Bank and other international development organizations go to great lengths to investigate and eradicate fraud, corruption, and collusion. We created a an automated classification system to help the World Bank's Integrity Vice Presidency automatically identify contracts that are likely to be tainted by corrupt practices.

The proof-of-concept classification system presented in this paper is the first, to our knowledge, to use machine learning to predict wrongdoing in contracts financed by an international development organization and provides the basis for potential future big data applications. The system leverages a data set containing contracts awarded and investigations conducted by the World Bank since 2000. Over 4,000 features were generated from these data sets and used in a Gradient Boosting Classifier to create a predictive model for fraud, corruption, and collusion. The classification pipeline is fully automated from ETL through generation of ranked lists of contracts, and we propose a prototype for a dashboard that could be used to integrate these results into an investigative work-flow. If implemented, this model is predicted to have a 70% success rate identifying contracts and complaints within World Bank projects where fraud, corruption, and collusion are most likely to be substantiated if investigated. The implementation of a machine learning system such as ours would enable fraud, corruption, and collusion detection to move from retrospective modeling to prospective modeling such that agencies can act to prevent wrongdoing, rather than being limited to only post-transgression sanctioning.

XI. ACKNOWLEDGMENTS

We wish to thank Elizabeth Wiramidjaja and Alexandra Habershon from the World Bank Group, Alan Fritzler from the University of Chicago, and Kristin Rozier from the University of Cincinnati for their assistance and support on this project. We also wish to thank the Eric and Wendy Schmidt Foundation and the University of Chicago for their financial support of this work through the Data Science for Social Good Fellowship.

REFERENCES

[1] *Pencil wireframe tool*[Online]. Available: http://pencil:evolus:vn/.
[2] *Resource guide: Procurement methods*[Online]. Available: http://web.worldbank.org
[3] The World Bank, *what is fraud and corruption?* [Online]. Available: http://www:worldbank:org/en/about/unit/integrity-vice-presidency/what-is-fraud-and-corruption.
[4] *World Bank what we do*[Online]. Available: http://www:worldbank:org/en/about/what-we-do.
[5] Transparency International.*Corruption perceptions index*[Online]. Available: http://www:transparency:org/cpi2010/results,2010.
[6] The World Bank Group Integrity Vice Presidency [Online]. Available: http://siteresurces.worldbank.org
[7] World Bank annual report. 2015.
[8] J. Akhilomen, ""Data mining application for cyber credit-card fraud detection system," *In Proc. of the 13th International Conf. on Advances in Data Mining: Applications and Theoretical Aspects, ICDM'13*, Berlin, Heidelberg, 2013, Springer-Verlag, pp. 218-228.
[9] The World Bank Group *Major contract awards* [Online]. Available: https://finances:worldbank:org/api/views/kdui-wcs3/rows:csvaccessType=DOWNLOAD.
[10] P. K. Chan and S. J. Stolfo, "Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection," *In Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining*,AAAI Press, 1998. pp. 164-168.
[11] E. Grace, A. Rai, and E. Redmiles. *Fraud-Corruption-Detection-Data-Science-Pipeline-DSSG2015* [Online]. Available: https://github:com/eredmiles/Fraud-Corruption-Detection-Data-Science-Pipeline-DSSG2015, 2015.
[12] The World Bank Group. *PPP conversion factor*[Online]. Available: http://data:worldbank:org/indicator/PA:NUS:PPP.
[13] IMF. *FRCF official exchange rate*[Online]. Available: http://data:worldbank:org/indicator/PA:NUS:FCRF.
[14] O. Irisova, "The cost of corruption," World Economic Journal, August 2014.
[15] Y. Kou, C. T. Lu, S. Sirwongwattana, and Y. P. Huang, "Survey of fraud detection techniques," In Networking, *Sensing and Control, 2004 IEEE International Conf.,2004*, vol. 2, pages 749-754.
[16] R. Nelson. *Multilateral development banks: Overview and issues for congress*[Online]. Available: http://fpc:state:gov documents/organization/189143:pdf, April 2012.
[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol.12, pp.2825-2830, 2011.
[18] E. Rozier. *World bank: Development of an entity resolution methodology for the world bank group* [Online]. Available: http://ceas:uc:edu/cyberops/research:html.
[19] D. Schuler and A. Namioka, *Participatory Design: Principles and Practices* L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1993.
[20] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," *Data Mining and Knowledge Discovery*, vol. 18, no.1, pp.30-55, 2009.