World Bank: Detecting Collusion, Corruption, and Fraud

Emily Grace Ankit Rai Elissa Redmiles Kristin Yvonne Rozier

August 27, 2015

Contents

1 Executive Summary

During the World Bank Group contracting process, companies may engage in fraudulent, collusive or corrupt behavior. The World Bank Group Integrity Vice Presidency (WB INT) currently relies on the complaints of whistleblowers and the domain expertise of contract investigators to identify wrong-doing. Fraud, corruption, and collusion drain $900 billion from the developing world annually[1]. Monetary losses from wrongdoing by companies in World Bank contracts contributes significantly to this drain. Thus, the World Bank Integrity Vice Presidency (World Bank INT) investigators and legal teams work tirelessly to investigate fraud, corruption, and collusion in World Bank contracts.

The Data Science for Social Good team spent this summer working with the World Bank INT to develop data-driven models and analyses that World Bank investigators can leverage in fighting fraud, corruption, and collusion across the globe. We estimate that as a result of our work, the World Bank will be able to increase their success rate in finding fraud from 38% to 70% while increasing their investigation efficiency by 50%.

The WB INT Group, chiefly Elisabeth Wiramidjaja and Alexandra Habershon, engaged a team of data scientists from the University of Chicago, through the Data Science for Social Good Program, to help the WB INT group move toward a more data-informed strategy for identifying and investigating wrongdoing in World Bank contracts. More specifically, the data science team was tasked with analyzing WB datasets containing historical contract, project and investigation information and developing a repository of code designed to generate a ranked list of allegations most likely to be substantiated if investigated. The Data Science for Social Good team has now developed a repeatable code pipeline which outputs the following assets:

- The latest contracts data set with resolved company names[1], investigation results[2], and new aggregated data fields generated by the University of Chicago team
- A list of new allegations with scores showing the probability of this allegation to become substantiated
- A list of contracts which are likely to be involved in corruption, fraud, or collusion even if they have not received complaints

We hope that these outputs and the code pipeline we built will be used by the World Bank INT group to combat fraud, corruption and collusion in World Bank contracts. Further, the future plan is for the World Bank Business Intelligence team to build out a custom investigator dashboard within the BI Portal that investigators can use to leverage the work of the Data Science for Social Good team. A wireframe design of this dashboard and further technical details are described in the remainder of this document.

2 World Bank Background Information

The World Bank is organized into a number of different groups. Our team worked with the World Bank Vice Presidency for Integrity (INT). The INT "investigates allegations of fraud, corruption, coercion, collusion, and obstructive practices related to World Bank Group-financed projects...Since 1999, INT has investigated and closed nearly 3,000 cases...To ensure the independence of its activities, the Vice President for Integrity reports directly to the President of the World Bank Group."

2.1 Contracting

Contracting at the World Bank is a process that involves many entities. The World Bank receives money from economically advantaged countries such as the United States of America and the United Kingdom. These countries may earmark their funding only to be used for particular projects (these are called *agreement types*, for example in the GEF agreement, a donor country has designated that the funding they provide can only be used for environmental projects). Economically disadvantaged countries (e.g. India, Uganda) then request funds from the World Bank for a particular

[1]The WB INT data set currently contains 74K unique company names, many of these names can be resolved to one another (e.g. PWC and Price Waterhouse Coopers), the DSSG2015 dataset in collaboration with researchers funded by the WB at the University of Cincinnati provides a resolution of many of these names such that the final data set contains 53k unique company names, a 30% reduction.

[2]Today, the WB INT investigators must use multiple systems in order to view the contracts, projects, and investigations data contained in the single data file produced by the DSSG system. The ability to view prior investigative results, project details and contract details all in one file would be extremely valuable to the INT team.

project (e.g. road infrastructure improvement). If the World Bank approves this funding request, a particular *implementing agency* (e.g. the Ministry of Finance) within a *country* then posts a Request for Proposal seeking contractors (often referred to as *suppliers*) to complete the project. Suppliers respond to these requests for proposal with written bids. Each request for proposal has a particular *procurement method* (e.g. National Competitive Bidding). Further discussion of different procurement methods can be found in Section **??**. The implementing agency in the country then evaluates the bids per the requirements of the procurement method and selects a winning contract. This contract selection (and the related supplier(s)) is then submitted to the World Bank for review. If the World Bank approves the contract, work begins.

Figure 1: Contracting Process Overview.

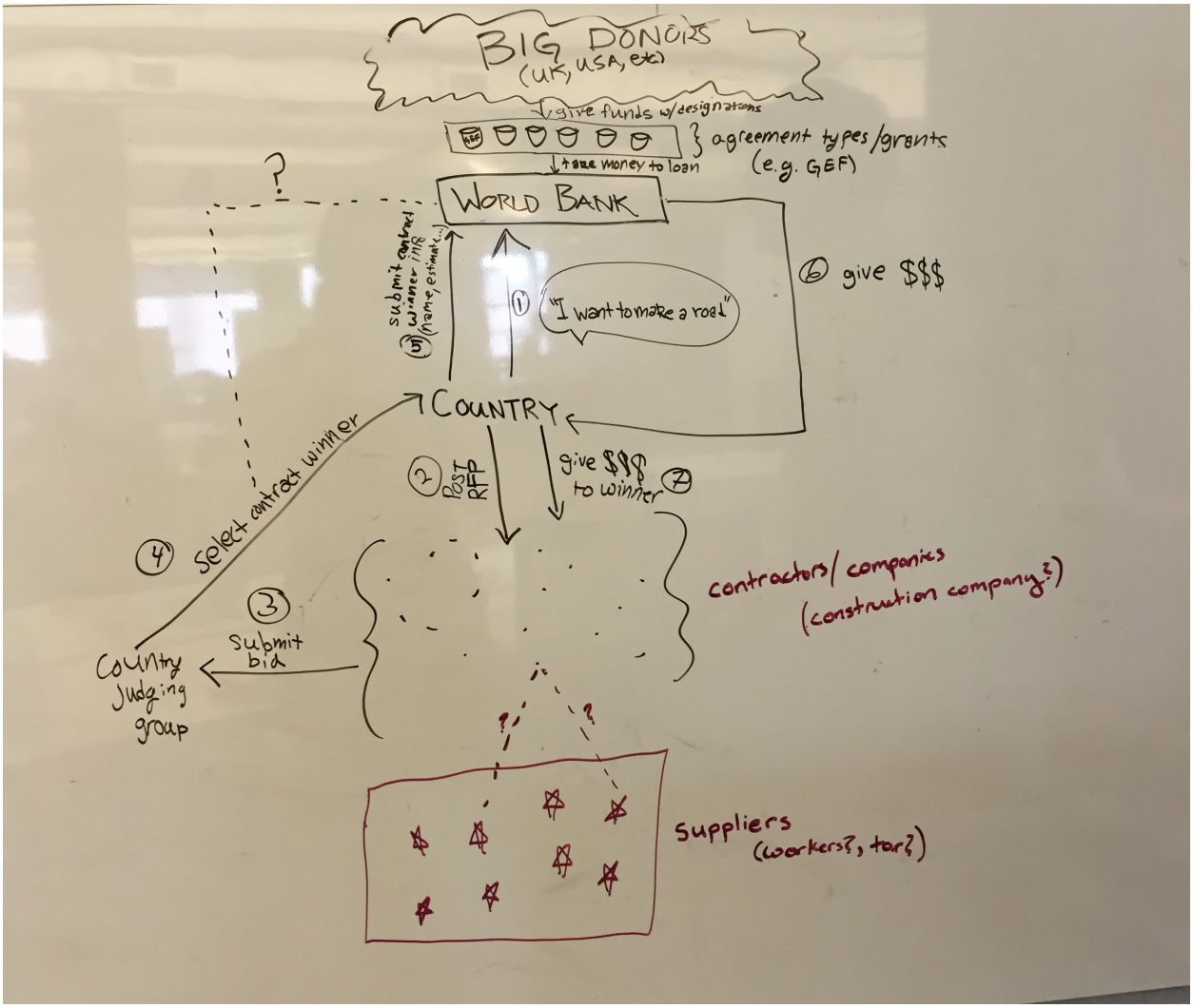


Figure 2: Contracting Process White Board Sketch.

2.1.1 World Bank Relationship Structure

The hierarchy of the World Bank contracting process includes the World Bank itself, each development project, the governmental implementing agencies which administrate the contracts, each contract award, and the suppliers that carry out the contracted work. A summary of this hierarchy can be seen in Figure 3.

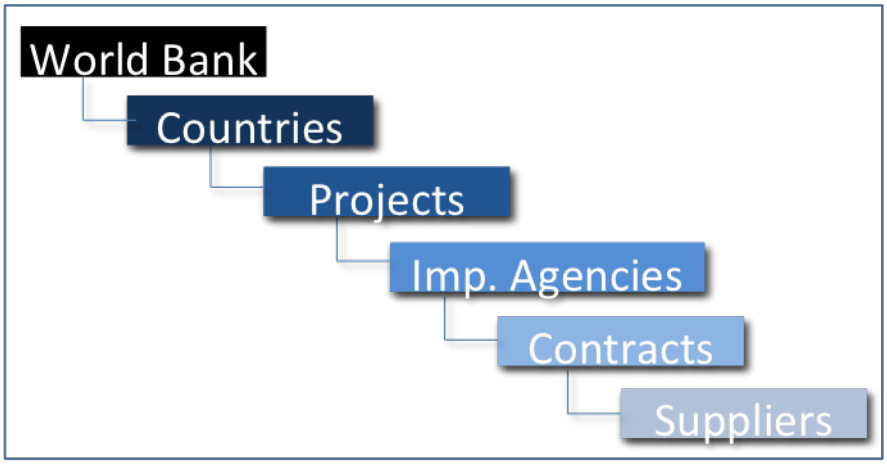


Figure 3: The theoretical hierarchy structure that was used to model the relationships between different entities involved in the World Bank contracting process.

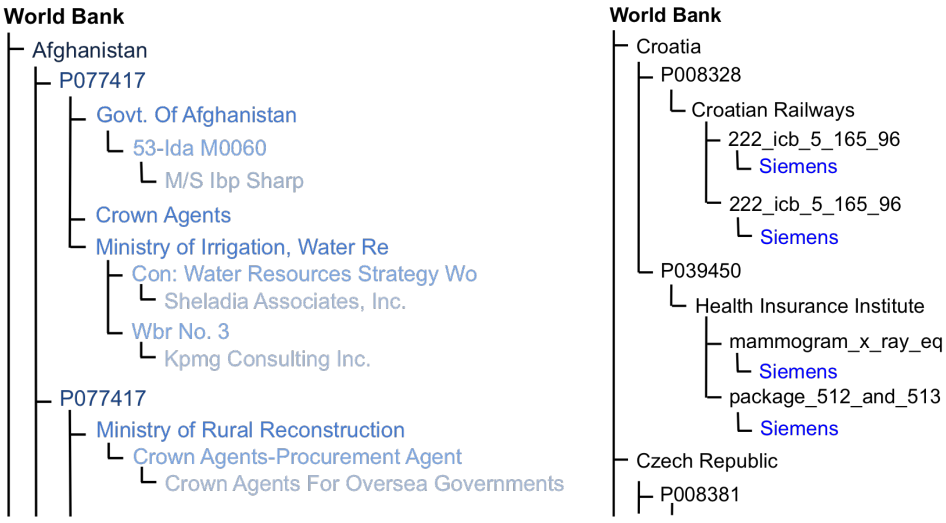


Figure 4: (Left) The top of the hierarchical tree generated from analysis of the World Bank contract data set. (Right) An excerpt of the tree structure showing the locations of a particular company (Siemens) within the full hierarchy.

To explore the relationships within the World Bank, a hierarchical tree structure was built from the contracts data set. Each data entry in the data contains the following fields needed to establish the full hierarchy:

- Country
- Proj_ID
- Implemtg_Agency

- Contract_ID
- Supplier

2.2 Investigative Process

The investigations process begins when someone who knows of or suspects wrongdoing submits a *complaint* regarding a contract or supplier to the World Bank. The World Bank then enters this complaint into the *pre-investigation database* and the *Management Report* file. If a complaint is determined to have enough information to move forward, the World Bank opens a *case*. Once a case is opened the World Bank investigates the allegations (this may include "missions" or visits to the country and supplier in question). An investigation will end in the determination of an *allegation outcome* which is entered into the Management Report. This allegation can end in a settlement or a sanction against the supplier.

Figure 5: Investigation Process Overview.

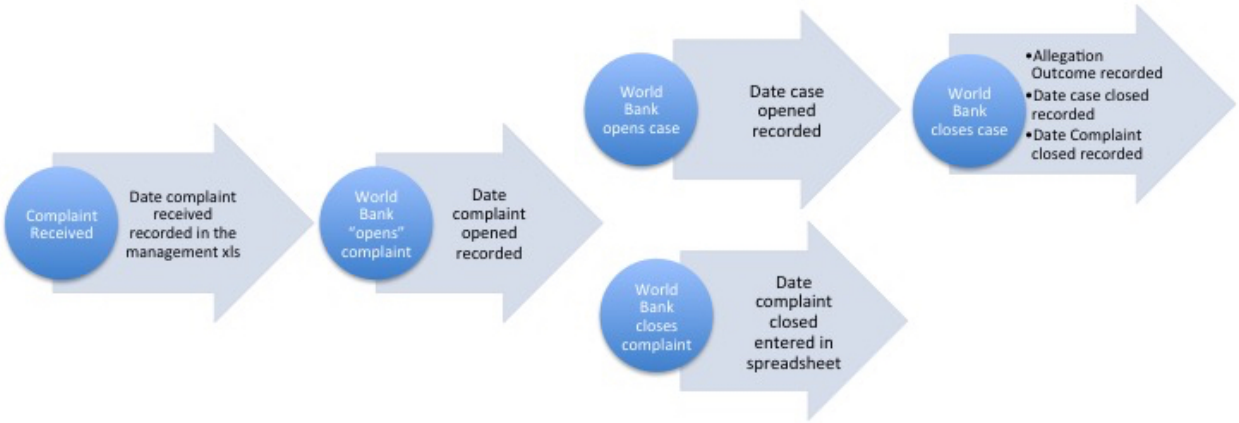


Figure 6: Investigation Process Recording in Management Reports File

2.2.1 Practicalities of Current Investigative Process

In the current process investigative process there is no linkage between the World Bank contracts data and the case management system investigations data. Investigators must switch manually between systems. Additionally, in their day-to-day work investigators are primarily looking at spreadsheets and documents, such as correspondence. There is currently not much visualization used to summarize information for the investigators.

More specifically, typically, an investigator will look at an allegation in the case management system, and then do research using the World Bank Intranet. This research will include the following steps:

1. Check finances @ worldbank.org for information on this contract/allegation.
2. Consider the World Bank's exposure to this company
    - What other contracts this company is carrying

- Size of company

3. Look in the project website to determine what other suppliers won contracts

3 Data

3.1 Data Summary

The data sets used in this analysis include:

- Contract data
  - **Source:** https://finances.worldbank.org/Procurement/Major-Contract-Awards/kdui-wcs3?
  - **Size:** 200,000 World Bank contracts
  - **Time period:** 2000 to present
  - **Fields include:** contract amount, region, supplier, and sector.
- Project data
  - **Source:** http://data.worldbank.org/data-catalog/projects-portfolio
  - **Size:** N World Bank projects
  - **Fields include:** total project budget
- Investigation data
  - **Source:** Private data provided by the World Bank INT
  - **Size:** N World Bank investigations
  - **Time period:** 2011 to present
  - **Fields include:** allegation category (fraud, corruption, collusion) and investigation outcome (substantiated, unfounded, unsubstantiated).

3.2 Data Stories

In order to better understand the World Bank data and what it contained, we performed a series of exploratory data analyses.

Story I: The contracts data set contains a total of $\sim$200,000 contract records spanning the years 2000 to 2014. The development projects for which work was contracted occurred in 168 countries around the world. These contracts were awarded to suppliers in 198 different countries (see Figure 7). In 76% of cases the contract was awarded to a supplier local to the borrower country. This proportion varied across the different borrower countries in the data set. The large majority of contracts awarded in South American countries went to a domestic supplier, while the proportion is lower in African countries (see Figure 8).

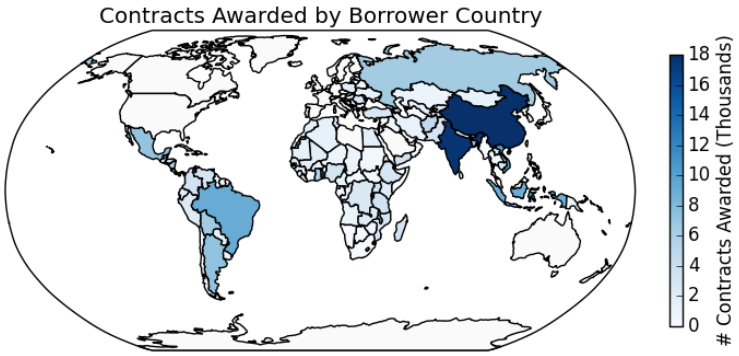


Figure 7: Number of contracts awarded to each borrower country over the time period 2000 to 2014

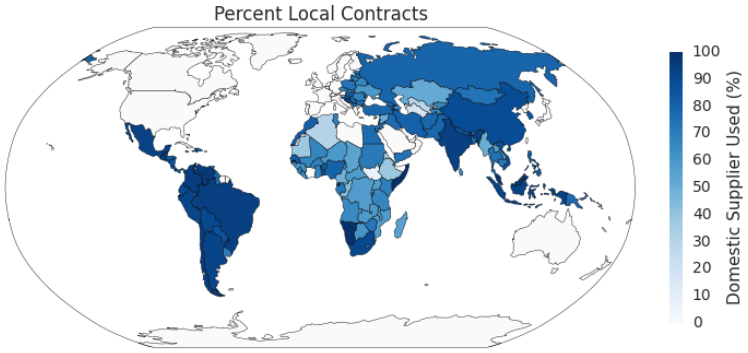


Figure 8: Percent of contracts in each borrower country that were awarded to local companies over the time period 2000 to 2014

Story II: The number of contracts awarded in each year varies over time.

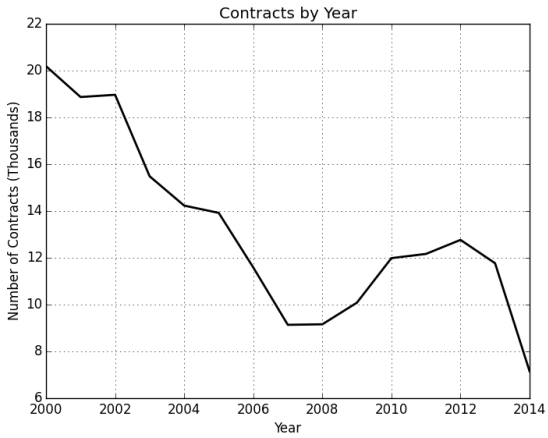


Figure 9: The number of contracts awarded each year by the World Bank. A long term decline in the number of contracts awarded was reversed in 2007. The incomplete data from the final year of the data set (2014) appears as an artificial downturn.

Story VI: The majority of complaints present in the investigations data set have been closed, but complaints received after 2012 are largely still ongoing.

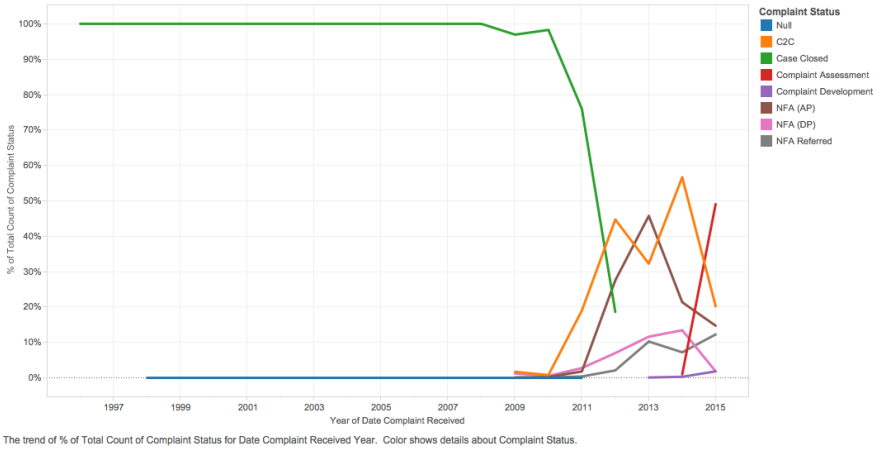


Figure 10: Status of complaints in the investigation data set by year that the complaint was received. The large majority of complaints received prior to 2012 have been closed out of the investigative process. The complaints received after 2012 are largely still ongoing in various stages of investigation.

Through our exploration we identified a number of anomalous data issues that were interesting to the World Bank.

Story III: There are 89 contracts that were awarded with values less than $100USD (prices adjusted for inflation). Of these contracts, 66% (59 contracts) were awarded in Venezuela. See Figure 11.

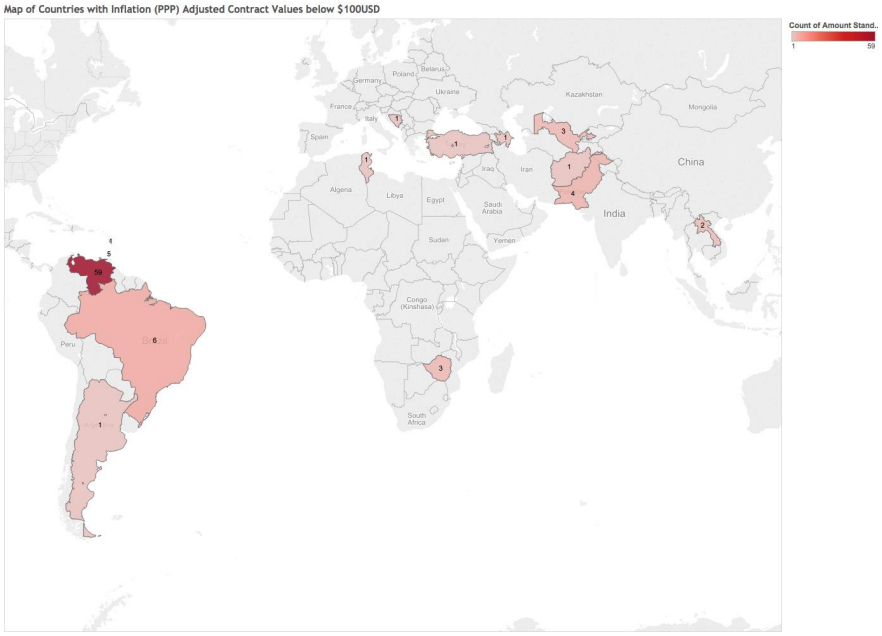


Figure 11: Number of contracts in each country with value below $100 USD

Story IV: There are 282 contracts that were not allowed to have a bidding process with domestic preference, yet had a domestic affect. See Figure 12.

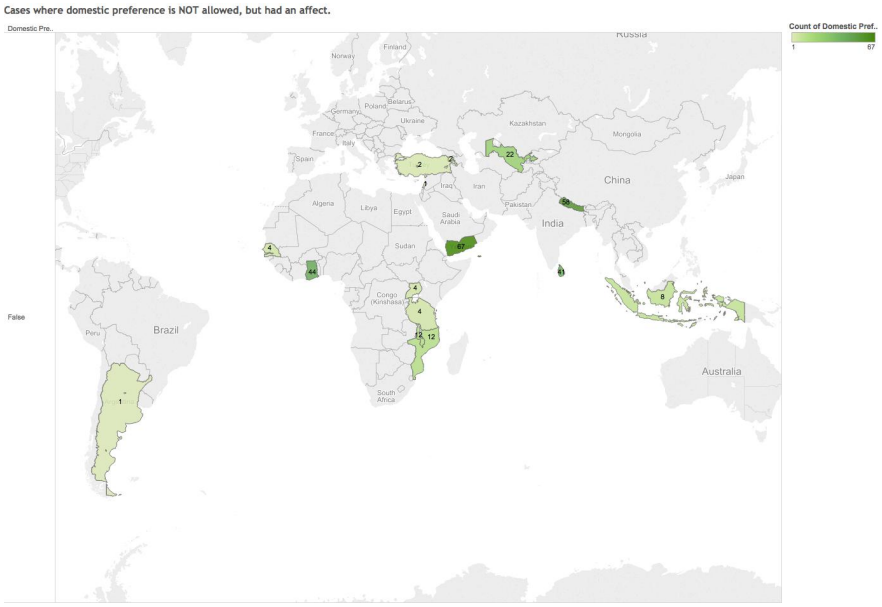


Figure 12: Number of contracts in each country where domestic preference allowed was false, but domestic preference was true

Story V: The investigation data set includes all possible combinations of the values of "Allegation Outcome" and "Outcome of Case When Closed." It is most common for these two variables to agree, but in some cases "Outcome of Case When Closed" is Unfounded even when "Allegation Outcome" is Substantiated. The "Outcome of Case When Closed" field indicates whether any wrongdoing was found in the investigation even if the specific company was not found to be involved.

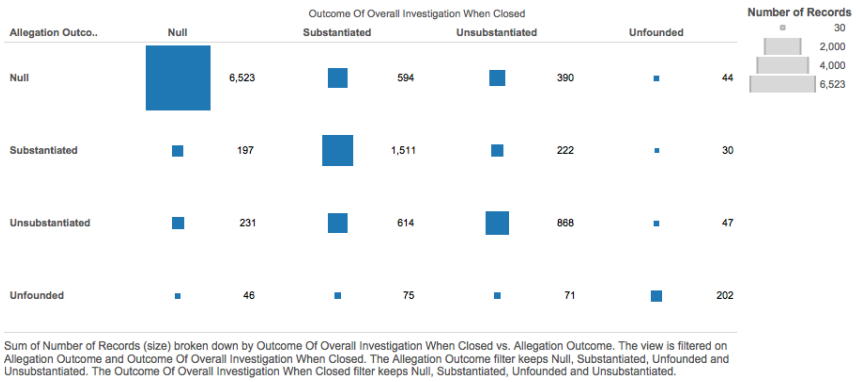


Figure 13: Comparison of "Allegation Outcome" and "Outcome of Case When Closed" variables showing the counts of each combination of values of the two variables. This shows that the majority of fields in the investigation set do not have either type of outcome. When such an outcome is present there is most often agreement between the two. In a significant subset of cases, there is disagreement between the outcome variables. In the case where there is a substantiated outcome of the case despite an unfounded allegation outcome this is due to a corrupt behavior other than the one alleged being proven. The case of a substantiated allegation outcome that doesn't result in a substantiated case outcome is not understood.

4 Features/Data Abstractions

Feature	Source	Type	Description
date_diff	Derived	int	Number of days between contract signing and award dates
amount_standardized	Derived	float	Amount standardized to 2010 USD via PPP
project_proportion	Derived	float	Percentage a given contract is of the project
total_project_amount	Raw	float	Total cost of project
#_supp_awards	Raw	int	Number of supplemental awards for a given contract
competitive	Derived	boolean	Whether a contract was awarded under competitive procurement
objective	Derived	boolean	Whether a contract was awarded under objective procurement

Table 1: Contract data features

We used features of our data to develop analyses and models for predicting fraud, corruption and collusion in World Bank contracts. The features that we developed and used are described below.

- **Objective Procurement**

  The method of procurement for a given project is originally labeled in the data as a text string. We categorize these methods as a binary: objective or subjective. See list below for procurement methods and labels.

  There are World Bank rules regarding what types of procurements can be non-competitive. We suspect that whether a contract winner is selected objectively or competitively may be an important abstraction for our model.

  - Single Source Selection: N/A
  - International Competitive Bidding: objective
  - National Competitive Bidding: objective
  - Quality And Cost-Based Selection: subjective
  - Quality Based Selection: subjective
  - Individual: subjective

- Least Cost Selection: objective
- CQS - Consultants Qualification: subjective
- Limited International Bidding: objective
- SHOP: objective
- Direct Contracting: N/A
- Service Delivery Contracts: unknown
- Selection Under a Fixed Budget: subjective

- **Competitive Procurement**

  The method of procurement for a given project is originally labeled in the data as a text string. We categorize these methods as a binary: competitive or non-competitive. See list below for procurement methods and labels.

  There are World Bank rules regarding what types of procurements can be non-competitive. We suspect that the use of non-compliant procurement methods may be a relevant feature for detecting corruption and collusion. The World Bank rules are listed below.

  - **Procurement Methods and Competitive Classification**
    * Single Source Selection: non competitive
    * International Competitive Bidding: competitive
    * National Competitive Bidding: competitive
    * Quality And Cost-Based Selection: competitive
    * Quality Based Selection: competitive
    * Individual: competitive
    * Least Cost Selection: competitive
    * CQS - Consultants Qualification: competitive
    * Limited International Bidding: competitive
    * SHOP: competitive
    * Direct Contracting: non competitive
    * Service Delivery Contracts: unknown
    * Selection Under a Fixed Budget: competitive

- **World Bank Rules for Non-Competitive Procurements**

  Sent by Elisabeth A. Wiramidjaja.

  > Direct Contracting or single source procurement is a method of procurement of goods that does not require elaborate bidding documents. The supplier is simply asked to submit a price quotation or a pro-forma invoice together with the conditions of sale. The offer may be accepted immediately or after some negotiations. Direct contracting may be resorted to by concerned procuring entities under any of the following conditions:
  >
  > a) Procurement of items of proprietary nature which can be obtained only from the proprietary source, i.e. when patents, trade secrets and copyrights prohibit others from manufacturing the same item;
  >
  > b) When the procurement of critical plant components from a specific manufacturer, supplier or distributor is a condition precedent to hold a contractor to guarantee its project performance, in accordance with the provisions of its contract; or
  >
  > c) Those sold by an exclusive dealer or manufacturer which does not have subdealers selling at lower prices and for which no suitable substitute can be obtained at more advantageous terms to the Government.

- **Currency Standardization**

  To allow direct comparison between different contracts across different countries and a large time-span, we must account for the effect of inflation and different purchasing powers. Therefore, each contract award amount (in U.S. dollars) was converted to the local currency and then converted to the price purchasing parity U.S. dollar amount. These conversions were done using historical data from:

- **Internationality**

  Another feature of potential interest is a supplier's level of internationality. As a proxy for internationality, we calculate the percent of a given companyś contracts that are in each of its top five countries.

- **Relative Award Values**

  In order to establish the relative scale of each contract, we use as an abstraction the award value compared to the total funds allocated to the parent project (this abstraction is represented as a percentage).

- **Historical Supplier Features**

  In addition to features related directly to static properties of the contract under investigation, features were generated to summarize the historical behavior of the contract's supplier. The strategy for these features was to calculate the percentage of a supplier's previous World Bank contracts that had a specific property over a particular aggregation period. For example, what percent of a supplier's contract in the past 3 years were in Africa or what percentage of all of the supplier's previous contracts were related to the agricultural sector? As an additional variation a set of ranked features were developed which didn't name the value of categorical variable. For example, what percent of a supplier's previous contracts were in its top country, second most common country, etc. This set of features was intended to capture a generic sense of how international a supplier was or how broad it's range of project sectors was.

  In total, five different categorical variables were considered:

  1. Country
  2. Region
  3. Major Sector
  4. Procurement Type
  5. Procurement Category

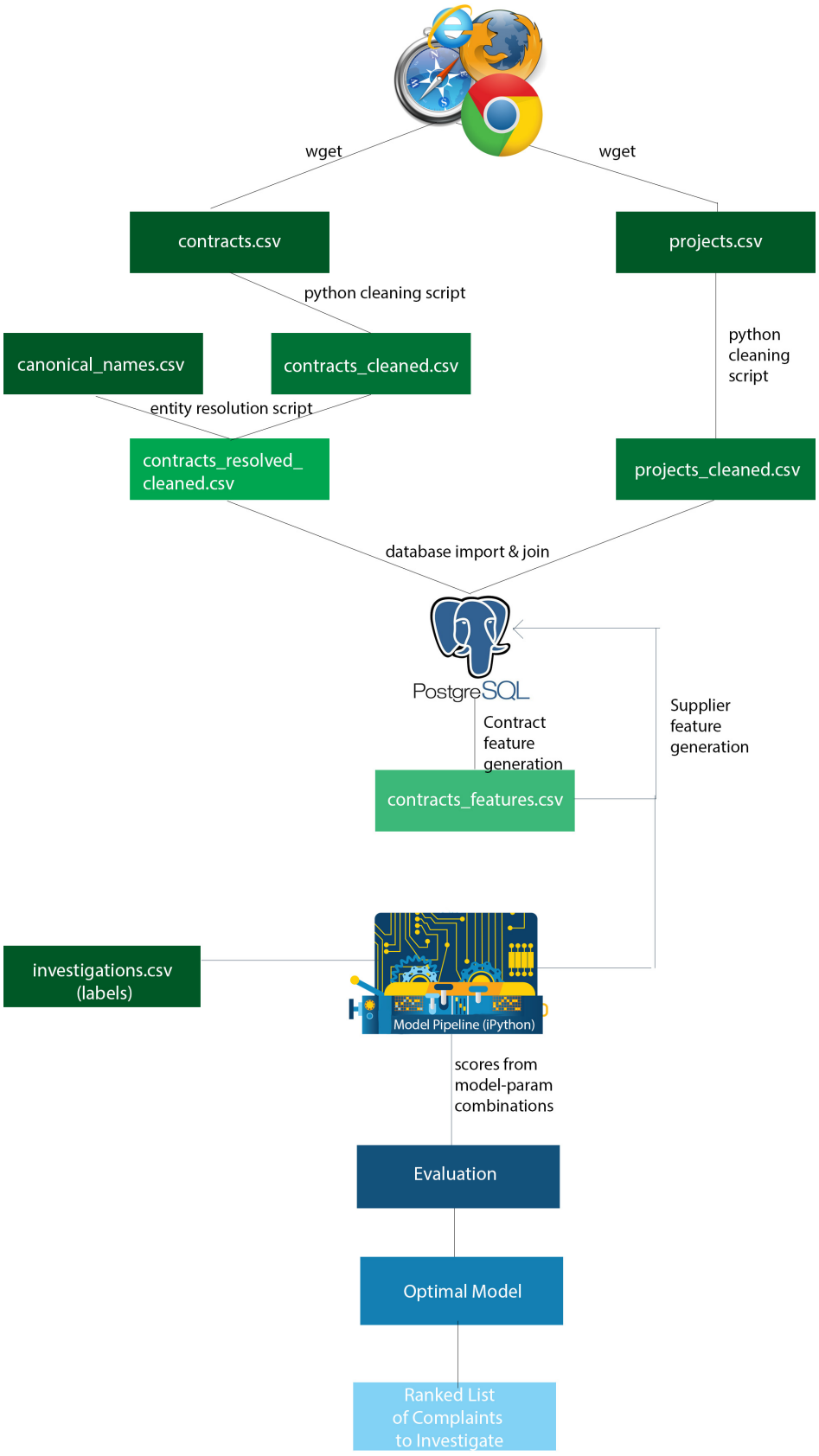
  Over four different aggregation periods:

  1. 1 year
  2. 3 years
  3. 5 years
  4. Full history

  Additionally, two versions of this full set of aggregations were carried out. One tracked the previous contracts by the amount of money awarded and the other by raw contract counts.

  Between all combinations of the different categorical variables, the aggregation periods, the named and ranked features, and the amount/count variation, a total of $\sim$4000 historical supplier features were created on our labelled data set. The total amount of features will depend on the number of values each of the categoricals takes in the data set (i.e. 7 regions, 10 major sectors, etc.).

5 Data Pipeline



5.1 Data Exploration

We developed a python script to enable the exploration of data files provided by the World Bank. This script reads in a given Comma Separated Value (cvs) or Microsoft Excel (xls) file and outputs the information listed below. Full documentation is available on Github in the README.

- The names and types of all columns in the data file.
- The length (number of rows) in the data file.
- For each column the following are output:
  - The number of unique values in the column.
  - The percentage of values in the column that are missing.
  - For columns containing numeric entries: the min, max, mean, median, and standard deviation.
  - For columns containing date entires: the range that the dates span.

Technical Requirements
The script is entitled data_exploration_script.py
Example command:

```
python [LOCALPATH]\/data\_exploration\_script.py -f [contracts\_file]
```

5.2 Data Cleaning

We developed a python script to clean the data files provided by the World Bank. This script reads in a given Comma Separated Value (cvs) or Microsoft Excel (xls) file. The script addresses the data challenges detailed below. Full documentation is available on Github in the README.

- **Dates**

  Dates within and across data files are specified in different formats (e.g. '09-10-2000', '9/10/2000'). We created a series of nested try/except statements to account for these formats and standardize all dates to the python datetime format.

- **Missing entries**

  Allows users to specify additional strings to parse as 'nan' values. That is, if the file denotes missing entries with 'NC', the user can specify this information in the data cleaning script run command in order to properly parse the file. The data cleaning script removes also any identical rows

- **Column names**

  The data cleaning script standardizes all column names to lower case and snake case (e.g. project_name). Users can specify specific columns to rename and provide new names, as needed. Users can also specify columns to remove, as desired.

- **De-duplification**

  The Major Contracts data files that we received from the World Bank contained some instances of sets of rows which contained identical entries in each column except amount. We chose to remove these rows and create one row with an amount that is the sum of the amounts in the previous nearly-identical rows.

Technical Requirements
The script is entitled data_cleaning_script.py
Example command:

```
python [LOCALPATH]/data_cleaning_script.py -f [contracts_file] -n [symbol to parse as NAN contained
    as string] -r [[column to rename],[new column name] sequence should be contained as string] -o
    [cleaned contracts file name]
```

5.3 Entity Resolution

The contracts data we received from the World Bank contained 73,858 unique company names. Many of these company names are actually representing the same company (e.g. Price Waterhouse Coopers, PWC, PriceWaterhouse-Coopers, Inc.). These companies represent suppliers, features of these suppliers are important in our data analysis. Thus, it is important for us to resolve companies that should be considered the same. This process is called entity resolution. A team of researchers at the University of Cincinnati led by Eric Rozier generated an entity resolution list which we used to resolve the entities in the contracts data.

We wrote a script which took in a csv file containing the contracts data and a csv file which contained the entity resolution list. We then replaced the suppliers in the contracts data with their resolved name (e.g. PWC was replaced with Price Waterhouse Coopers). The entity resolution csv list had the following format:

Original name	**Semantically Resolved Name**	**Syntactically Resolved Name**
PWC	Price Waterhouse Coopers	Price Waterhouse Coopers
Price Waterhouse Coopers	Price Waterhouse Coopers	Price Waterhouse Coopers
PriceWaterhouseCoopers, Inc.	Price Waterhouse Coopers	Price Waterhouse Coopers

Table 2: Example Entity Resolution List/Canonical

The Semantic and Syntactic name resolutions were derived differently. Semantic names were derived by . Syntactic names were derived by . Semantic names were used in our analysis per researcher recommendation.

After our resolution, we had 57,355 unique suppliers (entities) in our contracts data.

Technical Requirements

The script is entitled entity_resolution.py. Full documentation is available on Github in the README.
Example command:

```
python [path to data_pipeline_src directory]/entity_resolution/entity_resolution.py -c [contracts
    data file] -e [entity canonical file] -o [name of contracts file with resolved entities]
```

5.4 Contract Feature Generation

We wrote a python script to generate the features listed below from the contracts data. For a more detailed explanation of each feature, please see the data abstractions section (2). This script also uses information contained in the projects data located here: . In order to run this script you must first load both the contracts and projects data into a database and join the two tables. Full_pipeline.sh has the relevant code - you can either run the full pipeline or leverage the commands contained within the script manually.

Technical Requirements

The script is entitled contracts_feature_gen.py. Full documentation is available on Github in the README.
Example command:

```
python [path to data_pipeline_src directory]/contracts_feature_gen.py -f [contracts data file] -p
    procurement_method [name of procurement method column in your data] -wf [output file]
```

5.5 Supplier Feature Generation

The pipeline next generates the aggregated features related to supplier history that were described in Section 4. This aggregation is performed with the supplier_feature_gen.py script.

Example command:

```
python [path to data_pipeline_src directory]/supplier_feature_gen.py -cf [contracts data file] -if
    [labelled contracts data_file] -y [number of years to aggregate (0 = full history)] -cat
    [categorical variable]
```

In order to iterate over all possible combinations of aggregation periods and categorical variables run the feature_loop.py script.

Example command:

```
python [path to data_pipeline_src directory]/supplier_feature_gen.py -cf [contracts data file] -if
    [labelled contracts data_file]
```

5.6 Prediction

We use a Gradient Boosting Classifier (from scikit-learn) with 1000 estimators, min_sample_split of 15, learning_rate of 0.1 and max_depth of 160 to assign probability scores to all *unlabeled allegations* in the investigations file provided by the World Bank. We use all contracts level features as well as supplier feature table 59 (which contains non-country specific count cumulative supplier features). The script can also be used to predict which *contracts* are likely to result in substantiated outcomes if investigated.

Technical Requirements

The script is entitled predict.py. Full documentation is available on Github in the README.

Example command:

```
python [path to data_pipeline_src directory]/predict.py -tf [training table name] -pf [OPTIONAL,
    prediction table name] -ac [OPTIONAL, allegation category] -fl [location of feature set log
    file] -wf [output file] -pred_id [identifier for prediction table] -train_id [OPTIONAL
    identifier for training table]
```

Training table name:	this would be the labeled contracts PostgreSQL table name
Prediction table name:	include if you want to predict on contracts rather than allegations, add the unlabeled contracts file here
Allegation Category:	include if you want to make predictions only for a particular allegation category - this is for contracts predictions
Train ID:	if for allegations typically "alleg"
Prediction ID:	if for contracts typically "cntrcts"

BASH script (full_script.py has more information on correct options). If you are predicting on contracts you should use prediction_loop.py to iterate through predictions on contracts for all possible allegation categories.

6 Modeling

The features described in Section 4 were used as an input into statistical learning models (a.k.a. machine learning models) to predict the likelihood of substantiating a claim of fraud, collusion, or corruption in a contract. We used supervised a machine learning technique, where the label is the outcome of past investigations. An investigation which was successfully substantiated was considered to be a positive outcome, while any investigations considered to be unsubstantiated or unfounded were considered as a negative outcome. Investigations with any other outcome were included in the modeling. The labelled historical contract data contained complaints ranging from 2011 to 2014 which referred to contracts ranging from 2000 to 2014.

To ensure that the machine learning model will generalize well, such that it does not overfit or under-fit to the training data and performs well on unseen future data, we created train and test split based on time window. The training sets were created with six months incremental time window (2008-2014), and the test set contains year data after specific training period. This allowed us to simulate the effect of using each model for prediction at different points in the past and evaluate the performance of the model on the known results of the investigations.

Train Through	Test Start	Test End	Train Size	Test Size
31/12/2007	1/1/2008	31/12/2008	59	44
31/06/2008	7/1/2008	31/06/2009	73	36
31/12/2008	1/1/2009	31/12/2009	103	16
31/06/2009	7/1/2009	31/06/2010	109	15
31/12/2009	1/1/2010	31/12/2010	119	13
31/06/2010	7/1/2010	31/06/2011	124	17
31/12/2010	1/1/2011	31/12/2011	132	22
31/06/2011	7/1/2011	31/06/2012	141	24
31/12/2011	1/1/2012	31/12/2012	154	36
31/06/2012	7/1/2012	31/06/2013	165	42
31/12/2012	1/1/2013	31/12/2013	190	34
31/06/2013	7/1/2013	31/06/2014	207	20
31/12/2013	1/1/2014	31/12/2014	224	6

Table 3: Train/Test Splits

Below is the basic workflow of supervised machine learning approach:

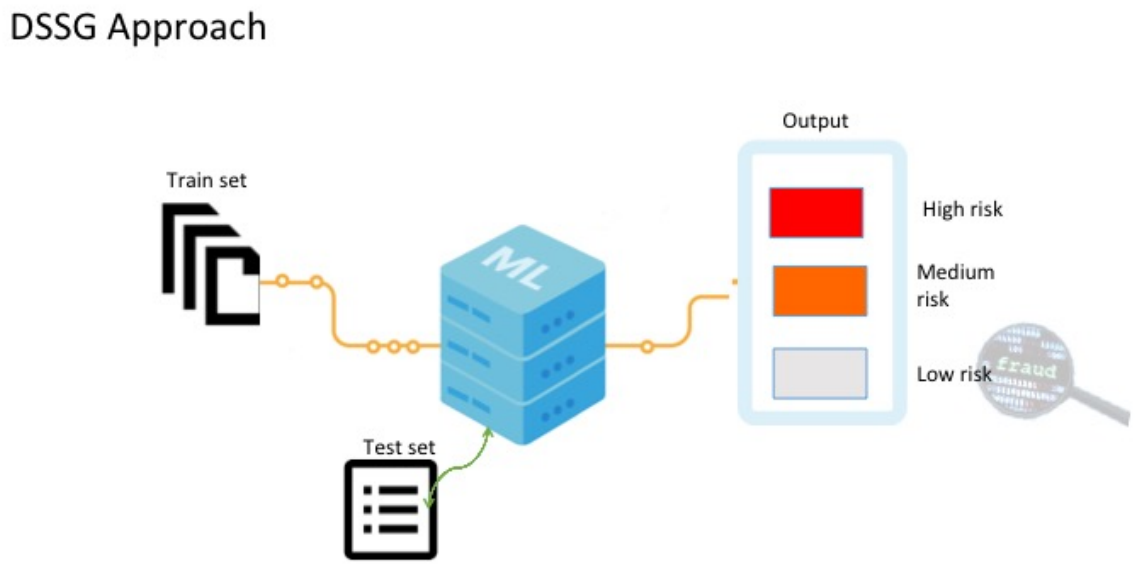


Figure 14: Workflow of supervised machine learning method

For each temporal train/test split a series of different specific models were used. The models which were considered are summarized in Table 4.

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 4: Models and parameter space explored to find the model with the optimal performance.

In order to iterate over all possible combinations of feature sets, models, model parameters, and feature sets the follow pipeline structure is used:

```
data_splits = [2008,2009,2010,2011,2012,2013,2014]
features_sets = [[set1_cols],[set2_cols],...]
models = [RandomForestClassifier(),AdaBoostClassifier(),...]
param_sets = [ [n_classifiers=100,max_depth=40], [n_classifers = 500, max_depth= 80], ...]

for training_data, testing_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)
```

6.1 Feature Sets

In addition to varying the machine learning models and their parameters in order to find the optimal predictor, different combinations of features selected. This served two purposes:

1. Checking for any reduction in model performance from including certain features which would be an indication of overfitting
2. To evaluate the relative importance to the model of certain types of features

The full set of feature combinations that were used can be seen in Table 5. These feature sets were selected with the aim of checking the following types 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 which did not take into account the specific country of a contract or company were considered.

2. **Aggregated contract count and amount feature**

   Feature 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 relevent information for detecting patterns of corruption, different feature sets were used which contained each of these aggregation time periods on its own.

Feature Set	Description
48	All features
49	All features with no contract specific info other than the allegation
50	All features except those that name countries
51	Only ranked features (rather than named)
52	All features with amounts but not counts
53	All features with counts but not amounts
54	All features with full history
55	All features with 1 year history
56	All features with 3 year history
57	All features with 5 year history
58	Non-country features with amounts but not counts
59	Non-country features with counts but not amounts
60	Non-country features with full history
61	Non-country features with 1 year history
62	Non-country features with 3 year history
63	Non-country features with 5 year history
64	No supplier history
65	No supplier history or named countries
66	All features plus the supplier name
67	No contract specific features except the allegation and supplier name
68	No supplier history, with supplier name
69	No supplier history, with supplier name, no named countries

Table 5: Feature Sets

6.2 Evaluation Metrics

The two primary evaluation metrics used to rank the performance of different models were:

1. **Precision**

   Precision is defined as the percentage of predicted positive values which are true positives. In our case, this translates to the proportion of complaints that the model predicts to be "Substantiated" that are actually substantiated. This metric is valuable to World Bank as a measure of success because it provides a measure of how effectively it directs their investigative resources. Because each investigation is a labor intensive process, ensuring that each investigation has high likelihood of success is a key measure of model performance.

   The DSSG team provided the World Bank with a ranked list of investigations, rather than a direct recommendation of which complaints to investigate. The precision of the model output can be measured as a function of the number of investigations are that pursued from the ranked list, moving down the list from top to bottom. Because the resources of the investigators are limited, achieving high precision in the top portion of the list is

important. For this reason, each model was evaluated on its precision when the top 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50% are investigated.

2. **Area Under Curve (AUC)**

   This metric is a measure of the balance achieved by the model between the true-positive rate and the false-positive rate. The Receiver Operating Characteristic (ROC) curve shows the trajectory of true-positive and false-positive rates as the probability threshold for a positive assignment is varied. An example can be seen in Figure 15. As the threshold is lowered a greater proportion of the test examples are predicted to have a positive result, whether these be true or false positives. In a well-performing model, a high probability score should correspond to positive test examples. In this case the true positive rate quickly approaches one before the false positive rate starts to increase. Therefore, an ROC curve with higher area under the curve (AUC) is better.

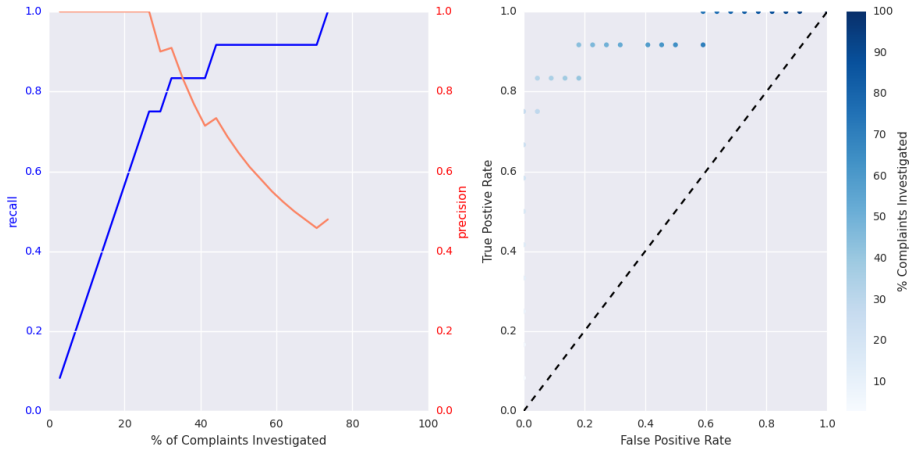


Figure 15: (Left) A precision-recall curve showing the trade-off between the success rate of substantiation (high precision) and maximizing the number of wrongdoers found (high recall). (Right) The ROC curve. This shows a different view of the same trade-off: if the threshold for predicting substantiation is lowered (and more cases are investigated) the number of false-positives will increase. Because the true-positive rate reaches 80% before suffering from any false positives, this model is performing well in the top portion of the ranked list.

6.3 Results

The best performing model type according the criteria of precision averaged across all train and test sets was the Gradient Boosting Classifier. In particular, model parameters 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 16.

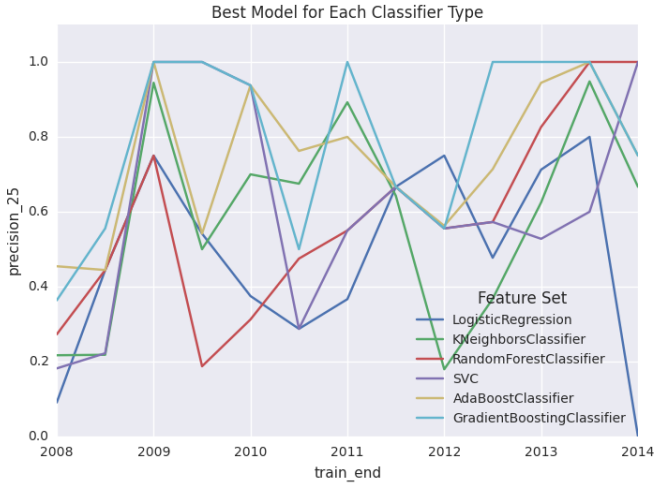


Figure 16: The maximum precision achieved in the top 25% of allegations for different model types across the different test/train splits. The gradient boosting classifier achieved the most consistent performance.

This model achieved an average precision of 73.1% in the top 25% of predictions. A summary of the other performance metrics of this model can be seen in Table 6.

Top n%	**Precision (%)**	**Recall (%)**
5	96.4	21.7
10	93.1	35.9
15	90.5	49.8
20	79.4	55.7
25	73.1	60.7
30	70.9	67.5
35	66.7	71.4
40	63.8	75.1
45	62.3	79.3
50	60.0	82.0

Table 6: Model Performance

The performance of the model across the different train/test splits in terms of precision in the top 25% and 50% can be seen in Figure 17. The performance varies significantly from test/train split to the next. This is likely caused by the low number of samples in each of the test sets, which ranged from 6 to 44 samples, causing large statistical variations.

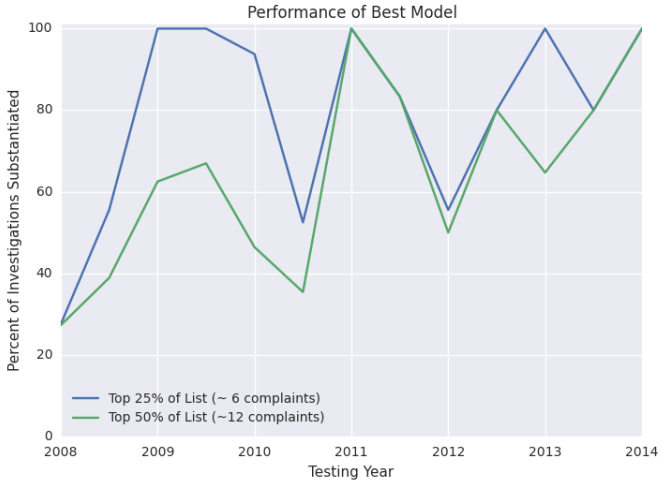


Figure 17: The success rate of targeting substantiated cases that our model would have achieved in each year in the past if the top 25% and 50% of the recommendations were investigated. The performance varies significantly from year to year, which can be attributed to the small size of the our data sets in each year. The amount of data in each year is around 25 cases with known outcomes. Thus the measured performance is not robust to statistical fluctuations. If we average over the performance across all years, we can expect around an 80% substantiation rate in the top 25% of complaints and a 65% substantiation rate in the top 50% of complaints

Because one of the top priorities of the World Bank was to target wrongdoing across all different countries, 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.

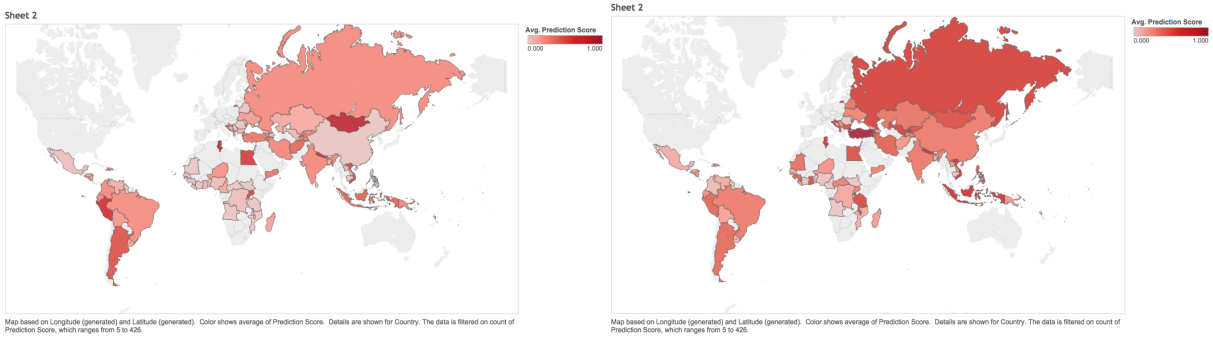


Figure 18: The average prediction score (i.e. probability of substantiation) in each country as predicted by two different models, one which includes country specific information (left) and one which does not (right). This shows that when the model is not provided with country information the resulting probability scores are more evenly distributed across the world. However, even when country information is not included as an input there is still some variation in the results by country indicating that the model is picking up on patterns which are more prevalent in some countries than others.

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 comparison of the performance of the two models can be seen in Figure 19.

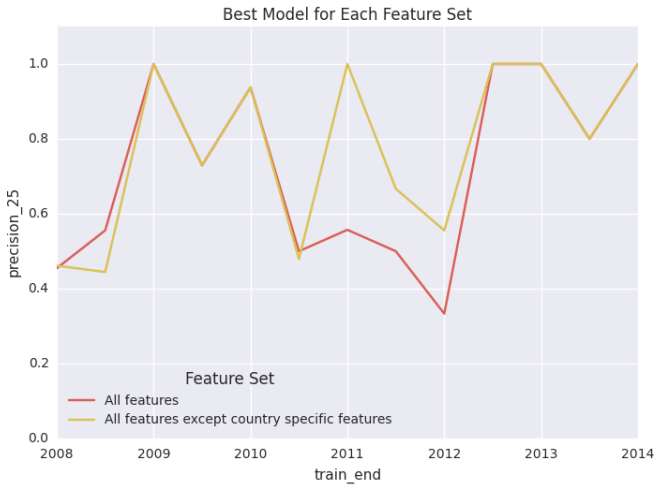


Figure 19: A comparison of the success rate of substantiation between a model which included country specific information and one which did not. The performance of the two models was largely similar, except for a period of several years where using the country specific information made the model worse.

6.4 Limitations

In moving forward to use this model in the real world, there are several limitations to keep in mind. Because the data set used in training and testing the model was fairly limited in size, the measurement of the model performance has a significant degree of uncertainty. In the series of test sets that were used for model evaluation, large differences were observed in the achieved precision.

An additional limitation is the limited scope of the data that was used for training. The data used was a subset of those World Bank contracts which:

1. Receive a complaint
2. Are selected to be investigated
3. Are related to an accusation against a company (rather than a government official, etc.)
4. Can be successfully matched to a World Bank contract ID

Thus, while good predictive performance was observed on this limited data set, the same performance is not guaranteed on other data sets. For example, when the model is run on the set of full complaints (not just those which are converted to cases) and the full set of contracts which do not receive complaints, care must be taken to evaluate the results.

For this reason, a field testing program will need to be designed. Three field trial alternatives are under consideration.

1. Give tool to analysts in certain trial regions to use in their case selection process. Determine impact via a "difference in differences" model over a set time period (say 18 months) compared with the non-trial regions.
2. Compare INT substantiation rates to model-only predictions (NO integration into human processes). The purpose of this would be merely to convince people on the efficacy of the tool.
3. Implement a non-complaint based, proactive process using the tool with a trial region and evaluate impact on substantiation rates

A more detailed document will be produced to outline the plan for going forward with a field trial program.

7 Technical Deployment

Our code pipeline is a set of python and SQL scripts which are all automated by a single BASH script. This code pipeline produces both flat files and tables which are loaded into a PostgreSQL database. We deployed our work to the World Bank team in two ways:

- Provided access to a fully documented Github repository containing the relevant code.
  - This repository contains a preliminary web dashboard in HTML, CSS, Javascript and PHP.
- Provided a virtual machine export. The virtual machine can successfully run the code pipeline and generates the data files aforementioned. This virtual machine will run Ubuntu and will have all necessary packages installed (Anaconda Python; python dependencies such as Pandas, a data science package; a PostgreSQL database, etc.). The virtual machine will be provided to the World Bank in OVF format.

The intended future plan is for the World Bank Business Intelligence Development team to build out an investigator dashboard within the pre-existing World Bank Business Intelligence Portal based on the wireframes shown below and to use the code pipeline and virtual box provided to design the back-end of this application. The Business Intelligence Development team can also leverage the rough web application prototype of the dashboard that is provided in our Github portfolio, if it fits their system requirements. We chose to deliver our work in this form, rather than in the form of a fully functional web application due to security constraints at the world bank, time constraints in the 12 week DSSG program, and desire to stay within the existing World Bank workflow by developing within the Business Intelligence portal, which is an already established application at the World Bank.

7.1 Investigator Dashboard Wireframes

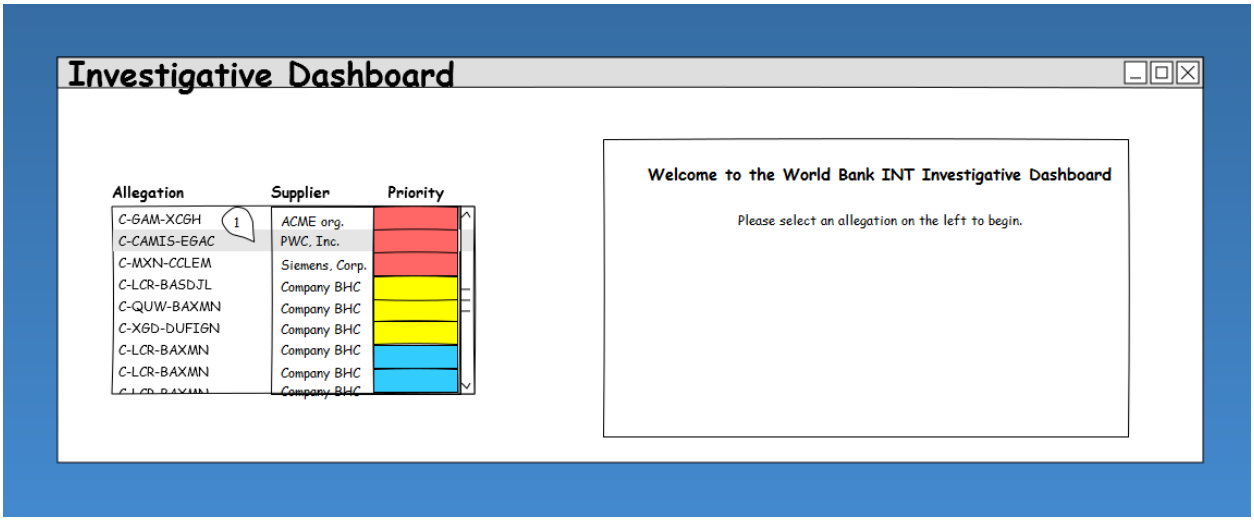


Figure 20: First frame of dashboard. User would click on an allegation to learn more. The ranked list will be the prioritized list of allegations generated by our model.

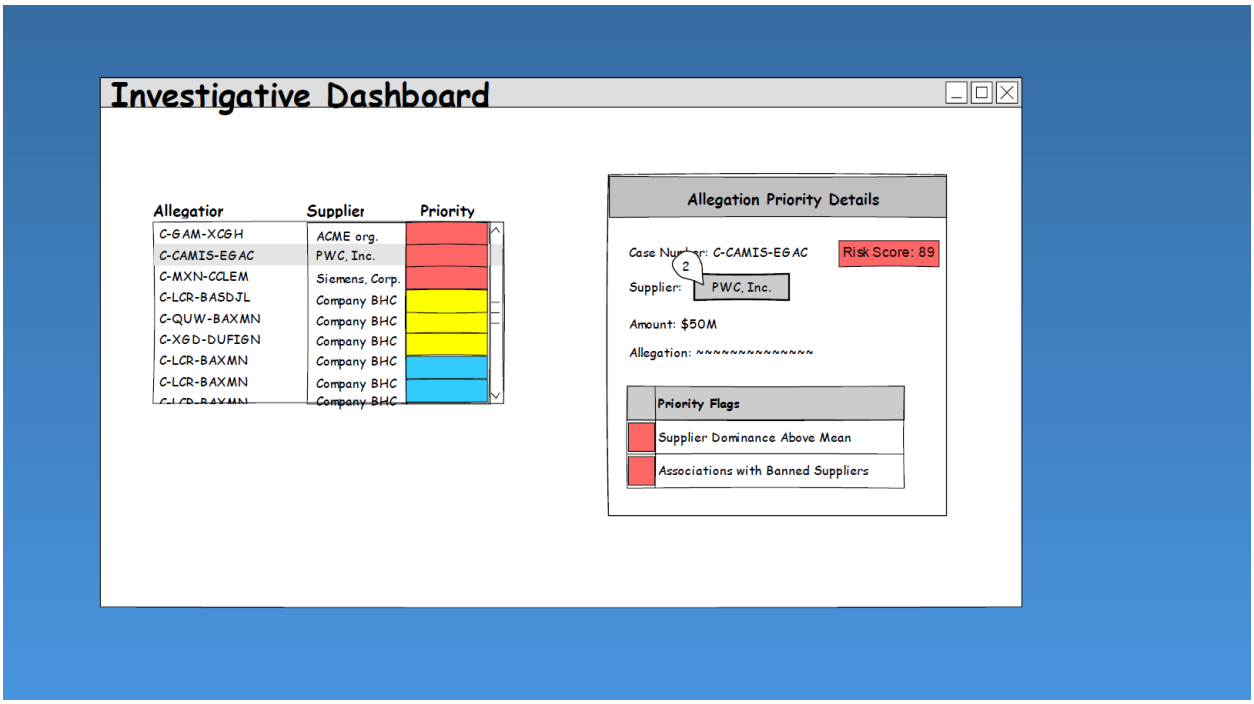


Figure 21: Second frame of dashboard. User can click on the button marked '2' to view supplier dashboard.

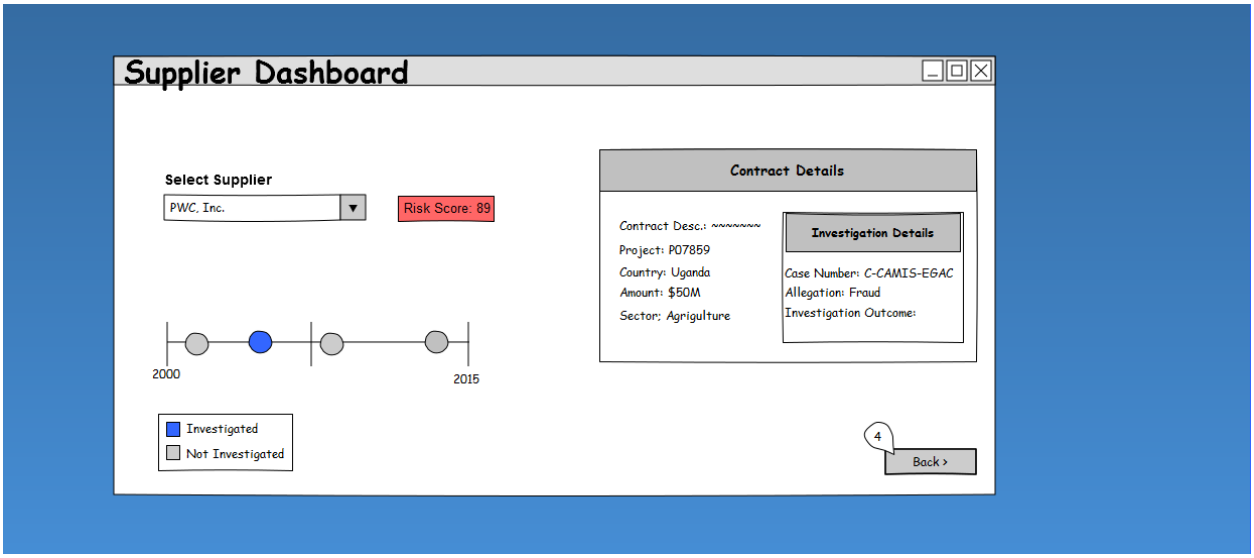


Figure 22: Third frame of dashboard. User can view supplier historical information - from our aggregated features- or return to previous frame.

7.2 Installation

7.2.1 From VirtualMachine

We have provided an ova file (virtual machine image) which you can install. This image contains our full pipeline, including all data dependencies (investigations data at time of DSSG2015 project, canonical file for entity resolution, currency conversion files as of 2015), all necessary packages, and a local PostgreSQL database. You can run our pipeline by clicking on the pipeline icon on the desktop (shown below) or by running bash dssg/Fraud-Corruption-Detection-

Data-Science-Pipeline-DSSG2015/WorldBank2015/Code/data_pipeline_src/full_pipeline.sh. You can install the virtual machine image using Oracle VirtualBox.

Installation Instructions

- Install Oracle VirtualBox: https://www.virtualbox.org
- In virtual box, File, Install Appliance, WBINT.ova
- Login credentials:
  - Username: dssg
  - Password: dssg2015

Space Requirements The OVA export is 2.7 GB. The virtual machine has 1024 RAM and 12 GB hard drive. Running the full pipeline (generating ranked lists for allegations *and* contracts) takes approx. 1 hour. Increasing RAM would increase this runtime speed.

7.2.2 From Source Code

You can also install our code pipeline directly from the source code available in this github repository. **Please note, the bash script that runs the full pipeline is designed for Linux/Unix OS**

Installation Instructions:

- Git clone https://github.com/eredmiles/Fraud-Corruption-Detection-Data-Science-Pipeline-DSSG2015.git
  Run full_pipelineśh to do EVERYTHING
- Install Anaconda python (see instructions here: http://docs.continuum.io/anaconda/install#linux-install) NOTE: Install Python 2.7 NOT Python 3.X
- Setup PostgreSQL database:
  - sudo apt-get install postgresql postgresql-contrib
  - Install according to these instructions https://help.ubuntu.com/community/PostgreSQL#Alternative_Server_Setup)
- Install pandas - conda install pandas
- Install seaborn - conda install seaborn
- Install csvkit - conda install csvkit
- Install psycopg2 - conda install psycopg2

Required Files

- The investigations file must be saved as a .csv file. Further, it must be named investigations.csv
- Canonical entity resolution file already included in pipeline_data (canonicalFileV2.csv)
- Download the zip file from here: http://data.worldbank.org/indicator/PA.NUS.PPP, save the ppp file within as ppp.csv in pipeline_data (or your DATA_STORAGE variable location, see below). Do the same for the zip file from here: http://data.worldbank.org/indicator/PA.NUS.FCRF, save the fcrf file as fcrf.csv
  - A 2015 version of these files is present in pipeline_data as ppp.csv and fcrf.csv

Modify full_pipeline.sh as follows

- The LOCALPATH variable should be the path to the WorldBank2015 directory that you cloned from git (e.g. /dssg/Fraud-Corruption-Detection-Data-Science-Pipeline-DSSG2015)

- The DATA_STORAGE variable should be the path where your data files AND THE INVESTIGATION FILE will live. there is already a directory called pipeline data in the github repo that contains the entity resolution canonical. You should use this directory (e.g. /dssg/Fraud-Corruption-Detection-Data-Science-Pipeline-DSSG2015/pipeline_data)

  NOTE: This is also where all output files will be stored.

- The CURRENCY_FILE_PPP variable should contain the full path where you want ppp.csv (a currency conversion file) to live.

- THE CURRENCY_FILE_FCRF variable should contain the full path where you want fcrf.csv (a currency conversion file) to live.

- If you are no longer using the local host database:

  - you must change the database connection in sql.py (Line 54), supplier_feature_gen (line 42) Example: host="localhost",user="dssg",password=password,dbname="world_bank"
  - copy example_config to config (e.g. cp example_config config), modify config to have the password for the database (e.g. the system password for your user, in our example the user password for dssg)
  - you must also modify model_pipeline_script.py (line 86) and supplier_feature_gen.py (line 217): create_engine('postgresql://[USER_NAME]:'+password+'localhost/DATABASE'. e.g. create_engine('postgresql://dssg:'+pas
  - you must create a config file in the directory from which you will run the script.

8 Future Work

There are many areas for development of our current approach which we couldn't address because of the limited time.

- **Data**

  Our current analyses were limited to using data from World Bank data sources. With additional time we could have explored using other public data sets, or sought out data from other development banks.

- **Pipeline**

  In the current pipeline, the entity resolution step is dependent on receiving an up-to-date flat file from the University of Cincinnati team. This process should be integrated into our pipeline such that the entity list is always up to date when the model is run.

- **Features**

  Given more time, there were several features that could have been generated from the data available to us.

  - Supplier's previous number of investigations
  - Supplier's previous number of substantiated investigations
  - Proportion of country's contracts taken by supplier
  - Proportion of country's contracts in given sector/category taken by supplier
  - Reimplementing last years network features

- **Modeling**

  A modeling approach which could be fruitful to try would be to apply ensembling methods to improve the achieved performance.

- **Evaluation and Interpretation**

  There are a number of steps which could be taken to provide more understanding of the results of the model. For example, we could try to apply apriori association rule to find out the feature association. Given more time, it would have been useful to look at which testing examples the model was performing well on compared to those it was misclassifying. Trying more combinations of feature sets would provide additional insight on which of the features are most predictive.

9 Data Questions

- **Q:** Is a certain time period between bid closing and contract award/sign date potentially important?

  A: Yes, similar to contract signing and award, consider different lengths of time relative to sector.

- **Q:** How might we distinguish between negative contract signing and award date differences (contract signing date prior to contract award date) which indicate suspicion vs. mis-entered data?

  A: I looked into the dates and found that the signature date before award date could be advance contracting or retroactive financing. It apparently happens often in WB projects, especially in emergency situations. -Betsy

- **Q:** In the management report data, what is the difference between the fields Allegation Outcome and Outcome of Overall Investigation When Closed?

  A: The discrepancy between the two is when a complainant alleges one or more things be we are able to prove an outcome of only some of those, or ones not originally alleged. For instance, we may receive an allegation of corruption, but only be able to prove a fraud.

- **Q:** Is it possible for domestic preference flag to be False and domestic preference affect to be True?

  A: No.

  ?domestic_preference_allowed? means that the evaluation process allowed for domestic preference to be an evaluation factor. Domestic preference means giving extra points in the evaluation process if a bidder is a national entity, as opposed to a foreign or international entity.

  ?domestic_preference_affect? means that the domestic preference factor affected the ranking of the awarded firm.

  Both these variables should have either ?Y? (yes) or ?N? (no) or a ?*? if left blank.

- **In the World Bank Major Contract Awards folder (e.g. Prior_Review_FY1986-1997.xlsx)**

  – Q: Is the following picture a correct representation/understanding of the investigations process and how information is recorded in the management spreadsheet? See figure 23

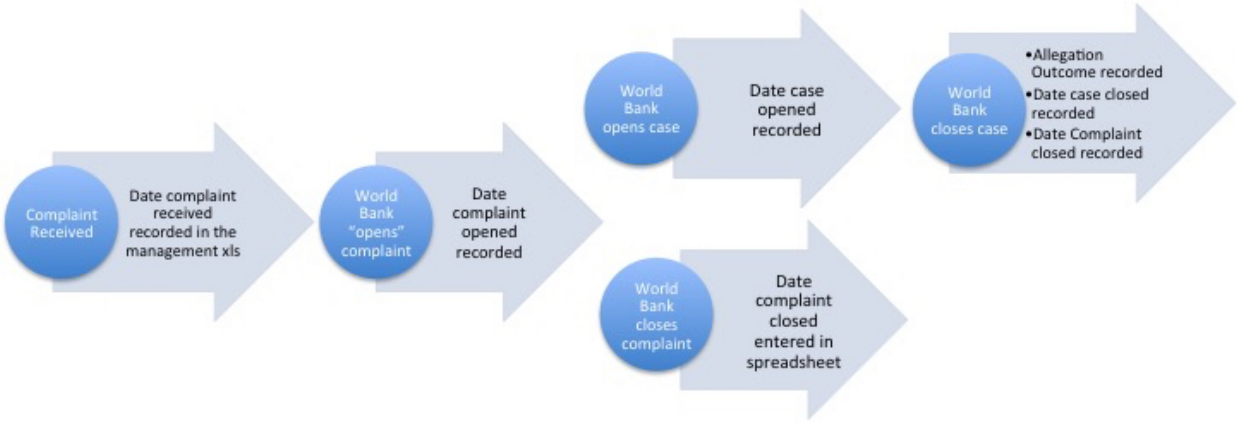


Figure 23: Investigations Process.

  A: Yes, this is a correct understanding.

  – Q: There is a column titled "Ln/Cr", an example entry in this column is "31700", what is this?
  A: This is the Loan or Credit number.

- Q: There is a column titled Unnamed: 5, an example entry in this column is "CLOSED(T) ROAD MAINT & REHAB.", is this a description of the project/project specification?

  A: This is a Project Name. The unique Project ID is directly to the left.

- Q: There is a column titled "Bor Ctr Ref #", an example entry in this column is "CONVENIO USBR & ADD", what is this?

  A: This is the Borrower Contract Reference. As mentioned, the WB has a unique contract identifier (a number, often begins with 12?. It is more common to refer to the contract by the Borrowers Reference which is often a more descriptive text label.

- Q: What is "domestic preference" and do you typically find that this attribute is related to fraud/corruption/collusion?

  A: Domestic preference refers to extra points being given to domestic firms (e.g. letting them money stay in the borrowing country and subsidizing experience for domestic companies). I don?t know if the presence of a domestic preference is related to fraud/corruption but that might be an interesting analysis.

- Q: It looks like F384 is a cost estimate form for the contract that must be submitted before work is started, with addendums for "supplemental awards" submitted when more money is needed, is this correct? Do supplemental awards tend to correlate with corruption/collusion/fraud in your experience or no relationship?

  A: The Form 384 is the document that records the actual contract award, not an estimate. Contracts may require extension or addenda. There is a strong suspicion that firms win on ?dive? prices (unrealistically low) and then make up the necessary revenue through addenda, by agreement with corrupt officials.

- Q: We noticed that there were a certain number of bids received (column title "Bids Rec'd") and a number of firms considered ("# Firms Considered"). In our Thursday meeting, we believe that you mentioned that if there is a long time between contract award date and sign date there was probably negotiation/controversy - would anything similar be true of differences between the number of bids received vs. firms considered?

  A: This data hasn?t been recorded since 2008 so it?s hard for me to know what the difference was, if any. For many contracts a large number of firms may express interest or purchase bid documents, and many fewer actually submit bids. Anecdotally it seems that when a large number of firms purchase docs but elect not to submit it indicates that they know that the award might already be in the bag for someone else. It would be good to know if this relevant, but since we haven?t collected the data in many years it might be hard. I believe we will be collecting this data again in the future.

- Q: In the "proc categ" column, there is an entry value "NC" - does this value mean that we don't know the procurement category? It think that is right though I didn?t see an example in the older file you mentioned.

  A: The category should be one of three items: Goods, Works, Consulting. There can be one or more methods of procuring each of these categories (e.g. QBS and QCBS for Consulting, ICB and NCB for Works)

- Q: What currency are the values in Amount? Is the value from the time of the contract award - e.g. 2001 dollars - or the current value (accounting for inflation)?

  A: Values are in USD and are the actual amounts awarded, not adjusted. If the contracts were awarded in another currency the amount was converted at the rate at the time. We noticed there were some negative values and very small ($0.04, $0.01, etc) in the amount column for some entries in Major Contracts - could you tell us a bit more about these small values (e.g. are these values human entry mistakes, are there really contracts with negative dollar amounts (and if so what does this mean), etc.)? Often contract addenda are recorded as new contracts but with no value (if no extra charge will occur.) Also, we have seen examples where two firms partner for a contract, and firm A is recorded as having the entire contract value, and firm B is recorded but with $0.00. I suspect a lot of human error as well.

- **In the National Procurement Data directory, for example in mex_cleaṅcsv**

  Q: There is a column titled "government_dependence" what is this?

  A: I?m not sure. This comes from a source identified in one of our hackathon events and one of the volunteers decided to download, clean and look at this data. I think you would need to go back to the original source (if it?s not identified in the file then it should be in the Google Docs spreadsheet of national procurement sites.)

- **In Management report-20150521-061422.xlsx (the complaints database)**

  Q: We noticed there is a complaint opened date and a complaint received date, and there are different amounts of time between these two dates. Is there any particular meaning to the differing lengths of time lapse between when a complaint is received and opened? That is, if there is a shorter period of time between when the complaint is received and when it is opened, should we assume that the investigator saw this item as ”more suspicious” and longer periods of time indicate ”less suspicious”?

  A: No, it really is more a function of how busy the intake office is and what the current metrics are. We now have to make decisions more quickly just because or management policies.