Improving the Precision of World Bank Investigations on Corruption and Fraud: A Machine Learning Approach

Machine Learning for Public Policy: Final Project

Carlos Grandet [*]
University of Chicago
cgrandet@uchicago.edu

Santiago Matallana [†]
University of Chicago
smatallana@uchicago.edu

Hector Salvador [‡]
University of Chicago
hectorsalvador@uchicago.edu

ABSTRACT

This paper provides a machine learning approach to labelling the World Bank contracts that are more prone to corruption and fraud. As a global institution, the World Bank awarded close to 200,000 contracts in the past 12 years, but less than 2,000 were investigated on malpractices. Lack of resources and personnel is likely a barrier preventing further auditing and sanctioning, thus allowing for the opportunity of many corrupt acts to go unnoticed. Thus, we aim to provide a model that allows the World Bank to be more efficient in their contract investigation by maximizing their ability to find the high-valued contracts incurring in corruption or fraud. We are proposing a supervised learning approach based on contract features related to supplier, project, country of origin, procurement information and major sector. We then train the data to find the best model that maximizes precision at a top N-percentile and test it with a subsample of labelled contracts. Our results with the testing data shows that our model has a precision of 50% on the top 20% of our sample with a threshold of 95%. Finally, we run the model with unlabelled data to provide a list to the World Bank of which contracts to investigate.

1. INTRODUCTION

Fraud and corruption is a major challenge to achieving the World Bank Group's (WB) goals of ending extreme poverty by 2030 and boosting prosperity in developing countries. Problems such as procurement above market prices, resources diversion and embezzlement of employees can limit its ability to provide aid to vulnerable populations. The problem of corruption is usually originated in the contracting process of the WB. The WB loans money to developing countries, which in turn conduct procurement processes to contractors. Not all procurement process follow the same structure, there are International Competitive Bids and there are also single source selections. Furthermore, the contracts can be for various areas (i.e. project management,evaluation, public works) and belong to different sectors within the World Bank major priorities (i.e. Water management,Infrastructure, Education). During the bidding and implementation processes, companies might engage in fraudulent behavior.

The WB currently relies on the complaints of whistleblowers and the domain expertise of contract investigators to identify malpractices. Procurement under WB-financed projects results in the award of about 20 - 30 thousand contracts in the past 5 years with a total value of about $20 billion each year. As a result of early detection, the WB has prevented approximately $138 million, across 20 contracts, from being awarded to companies that had attempted to engage in misconduct.[1]

We believe that the current approach relying on whistleblowers can be improved through a machine learning methodology. Nowadays, the WB is not proactively auditing contracts which is a barrier to deterrence and detection. The main reason is the lack of resources to conduct audits on large samples that might render unsuccessful results. Our project has the potential to provide an effective tool to target and select the contracts that are involved in corruption. This represents an opportunity for the WB to utilize their own data to improve their investigation procedures and also to become more effective in dectecting corruption thus increasing their reach and impact.

2. RELATED WORK

During the summer of 2015, one of the Data Science for Social Good (DSSG) teams worked with the WB's Integrity Vice Presidency (INT) to develop a data-driven model to increase the WB's success rate in finding corruption and fraud. In general terms the 2015 project had the following aspects

- Create a consolidated database from the World Bank data on projects, contracts and allegations
- Entity resolution for supplier companies
- Feature generation on supplier, procurement method and contract financial information

[*]MSCAPP 2017
[†]MPP 2016
[‡]MSCAPP 2017

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

[1]The World Bank Group Integrity Vice Presidency, Anual Update, Fiscal Year 2015:

- Model training looping through models and parameters to maximize precision

This project achieved an average precision of 75% on the top 25% of the sample. Our intention is to maximize the precision on the top 10% (see section General Approach for a more detailed explanation), we plan to achieve this through an intensive exercise of feature generation and the inclusion of ensemble models into our pipeline. These are two of the aspects that the team considered as next steps.

However, we will build on the data cleaning and entity resolution work of the DSSG team and use the consolidated database that merge the project, contract and allegations information, as well as did entity resolution for the suppliers names. To our knowledge this is the only related work that has been conducted.

3. GENERAL APPROACH

Our project aims to provide information to the WB officers about which are the contracts more likely to incur in misconducts. The objective is to improve the current method of corruption detection which solely relies on whistleblowers. While relying on whistleblowers is an usual effective strategy to detect corruption, it is also a limited one for it depends on a third party to provide information. Additionally, since it is an *ex-post* strategy one can only find the contracts and suppliers involved in a malpractice once the act has occurred which still causes losses and hinders the WB ability to efficiently allocate their resources.

3.1 Ideal corruption control policy

An ideal corruption control policy should aim to have a deterrent effect in which there are appropriate incentives not to commit corruption due to an expectation of a punishment. This could usually be achieved through randomized audits in which all contracts could be subjected to an internal investigation. However, these process might be costly and will not always yield corrupt contracts.

The ideal approach should have three conditions

1. A deterrent effect so that people are not tempted to be involved in malpractices.

2. An effective targeting of contracts that present corruption or fraud to maximize the available resources of the WB.

3. A prioritization of the contracts with the highest expected value

We believe that a machine learning predictive model approach can address both issues through an informed selection of the contracts that are more likely to present malpractices. If implemented appropriately this will increase the WB's effectiveness in detecting and sanctioning corruption without wasting resources on contracts that did not presented such issues. While also creating a deterrent effect in which key actors of the procurement process are aware that the WB is conducting investigations actively and not only relying on whistle-blowers.

4. DATA DESCRIPTION

The DSSG team worked originally with three databases from the World Bank:

Contracts (access via API)
Provides information on contractor, project country, project sector, contract signing date, procurement method and type, contract amount, and links to information on the project to which the contract belongs.

Projects (access via remote server)
Provides information on the lending instruments, project status, approval date, total grant, project description, geographic location and goal.

Investigations (access via remote server)
Provides information on the projects that have an initiated investigation. They have data on type of complaint, resolution, sanction, parties involved, and length of investigation. The data is available by project, not by contract, which might be troublesome to trying to detect corruption in contracts.

The 2015 DSSG team consolidated this data into one single database that was used in the pipeline of the project. We used this data as our main source of information, once the features were generated these were the following characteristics of our data:

Table 1: Characteristics of model database

Size of n	567
Number of features	162
Percentage of positive finding of corruption	41.5%
Percentage of negative finding of corruption	1.3%
Percentage of unlabelled data	57%

An initial exploratory analysis gave us some insights into the data. Records are distributed across 173 borrower countries, with China holding the maximum number of instances: 11,423 (5.5%).

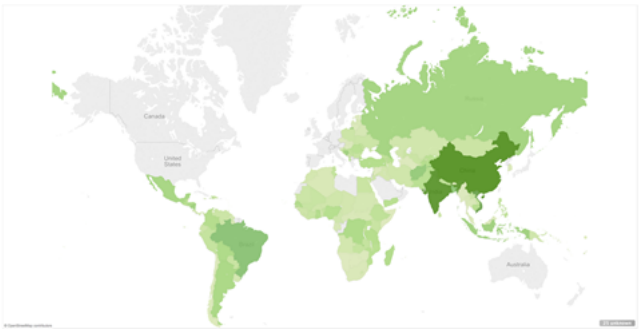

Figure 1: Distribution of contracts by country

Also, nearly every country around the globe supplies the goods and services associated with the contracts, as shown in Figure 2. Naturally, countries that do not act as WB borrowers, such as the US, Canada, Australia, and most West European countries, do act as suppliers. China also holds the maximum number of records by supplier country: 11,765 (5.7%).

Each record in the contracts dataset is labelled under one of 70 different procurement types, four procurement categories, and 15 procurement methods. In Figure 3 and 4 we

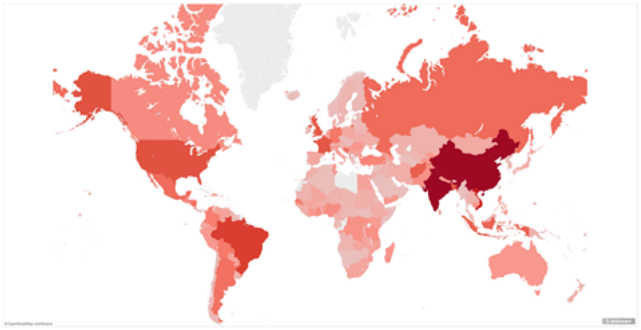

Figure 2: Distribution of supplier by country

observe how are those contracts distributed. Most contracts are destined to management, infrastructure and rehabilitation. Also, 24% of the contracts are obtained through international competitive bidding and 20% through national competitive bidding, which amounts to 44% of contracts obtained through a competitive method.

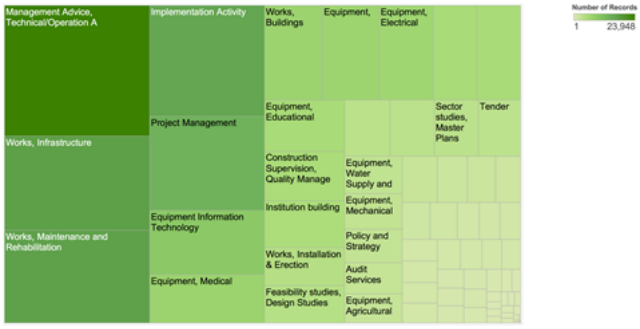

Figure 3: Distribution of contracts by sector

However, there is also a significant amount of contracts that were granted to single selections or direct contracting. This amounts to 28% of the contracts.

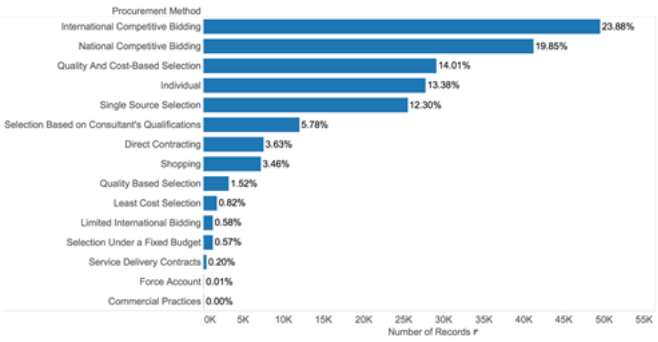

Figure 4: Distribution of contracts by procurement type

5. MODEL CHARACTERISTICS

5.1 Feature generation

Before the feature generation we did some data preprocessing to be able to run the models. We did the following:

- Drop repeated data on contracts
- Transformed data from string variable to categorial or dummy
- Drop columns that weren't useful for analysis such as project ID

Feature generation allows us to provide better inputs to our model and improve our predictions. Building upon the work done by the DSSG group in summer of 2015 we create additional features regarding characteristics of the contracts, projects and loans. We expand upon that work and create features in 5 categories: contract amount, procurement type, sector, supplier, and country characteristics.

Contract: Contract amount, Contract larger than average country contract all years - Dummy, Contract larger than average sector contract all years - Dummy, Contract larger than average supplier contract all years - Dummy, Total value of contracts for country in the year of the contract - Float, Ratio of contract amount over total amount per country - Float, Ratio of contract amount over total amount per sector - Float, Ratio of contract amount over total amount per supplier - Float,

Procurement type: Percentage of contracts obtained through international bidding by country - Float, Percentage of contracts obtained through international bidding by supplier - Float, Percentage of contracts obtained through international bidding by sector - Float, Percentage of contracts obtained through single source selection by country - Float, Percentage of contracts obtained through single source selection by supplier - Float, Percentage of contracts obtained through single source selection by sector - Float, Contract obtained through a method not present in 90% of the contracts - Dummy, Contract obtained through a method not present in 90% of the contracts by country- Dummy, Contract obtained through a method not present in 90% of the contracts by supplier- Dummy, Contract obtained through a method not present in 90% of the contracts by sector- Dummy,

Sector: Contract belonging to a specific sectors (water management, infrastructure) - Dummy

Supplier: Rate of concentration of projects for specific supplier - Float, Supplier for a different country than project - Dummy, Atomization of supplier (supplier contract amount divided by project amount) - Float, Diversification of supplier (how many sectors is he a contractor for) - Integer, Diversification of supplier (how many countries is he a contractor for) - Integer, Diversification of supplier (how many projects is he a contractor for) - Integer

5.2 Feature selection

In total we create over 100 features that extract different aspects of the data. In order to select which ones are the features that best fit the data and prevent overfitting we use a lineal Support Vector Machine model that uses a Lasso regression analysis to select which are the main features in the model.

In order to select the best features we run a loop changing the parameters of our Lasso method and randomizing

over the initial variable to select the best group of features. In total we obtain 23 features that fit the Lasso restrictions. We then transform our dataset to keep the features that best fit our model. The results of the feature selection are in Table 2.

Table 2: Selected Features

Percentage amount of civil works contracts grouped by major sector of contract
Percentage amount of consultant services contracts grouped by year of contract
Percentage amount of goods contracts grouped by supplier of contract
Percentage amount of International Competitive contracts grouped by year of contract
Percentage amount of Limited International Bidding grouped by major sector of contract
Percentage amount of shopping contracts grouped by country of contract
Percentage amount of single source selection contracts grouped by supplier of contract
Percentage amount of contracts that supplier has of total
Dummy if project is procured through Quality and Cost-Based Selection
Dummy if project is procured through Single-source selection
Dummy if project is in Education
Dummy if project is in Transportation
Amount of contract
Number of sectors a contract belongs to
Country mean contract amount
Currency converter
Number of contracts a supplier has in country of contract
Number of contracts a supplier has in sector of contract
Procurement category
Dummy for contract higher than the mean in year of contract
Dummy for supplier belonging to a different country

5.3 Model

We split the data into testing and training data, ordered by year. Thus 20% of our most recent data was used for testing and the rest was used for training, preventing using future data to predict the past.
Then we implemented a loop of classifiers that iterated through different parameters with the following models:

- Logistic Regression
- Decision Tree
- Linear SVM
- Naive Bayes

And the following ensemble methods:

- Gradient boosting
- Random Forest

The loop changes the parameters of each model and does a cross-validation with 3-folds. In order to determine which

5.4 Evaluation

In order to determine which is the best model we calculate three measures:

- Precision
- Area under the Precision-Recall Curve
- Precision at the top 10% of the contracts with the highest expected value (EV) where EV is defined as the product of the score times its monetary value.

We are interested in maximizing precision, which is defined as the ratio of predicted positive values which are true positives, given a certain number of possible investigations. (Intuitively, precision is the ability of the model not to label as positive a sample that is negative.) In the context of this project it is the ratio of contracts that our model predicts to have a misconduct and those that actually do. This provides a valuable measure of the current effectiveness of WB's investigative resources, and should drive future efforts in terms of high likelihood of success.

As noted, this approach hinges on the plausible assumption of scarce resources. That said, we do not know how many contracts the WB is actually capable of investigating. In absence of information on resource availability, we evaluate the precision of our model assuming that the top n contracts of a ranked version of the dataset are investigated.

Additionally, we rank the dataset prior to calculating precision by EV, defined as the product of the predicted probability of misconduct and the value of the contract. By incorporating expected value as a criterium for first ranking contracts we are addressing the fact that, broadly speaking, we will prefer to investigate a millionaire contract over a low value one, even if the latter has a very high predicted probability of misconduct relative to the former.

We use this EV measure because, as stated on our definition of an ideal corruption control policy, we need to be able to prioritize which are the contracts where corruption is more costly. Without any constraints, we would be interested in the model's ability of correctly detecting contracts with misconduct, given that it is most costly to omit investigating a fraudulent contract than to investigate a contract with no misconduct whatsoever. However, that approach assumes no resource restriction and the WB has limited resources to allocate to the investigation of contract misconduct.

A thorough description of the model results is found in the Results section.

6. RESULTS AND POLICY RECOMMENDATIONS

Our work offers the WB a data-driven criteria to prioritize investigation of reported contracts. Summing up, we predicted the probability of wrongdoing in flagged contracts, calculated their expected loss, and maximized the precision of our models subject to a resource constraint. With our best performing model, our approach yields a ranked list of reported contracts for the WB to focus on in order to improve the effectiveness of resource allocation. Were the

Table 3: Models results

Model	Precision	Recall	ROC-AUC	Precision at top 20%
LR	0.31	0.14	0.50	0.88
KNN	0.17	0.04	0.49	1
DT	0.17	0.07	0.48	1
SVM	0.19	0.23	0.50	0.11
RF	0.24	0.10	0.50	1
GB	0.12	0.04	0.47	1

WB's Vice Presidency of Integrity to implement our methods in a pilot, we would be able to validate ex post the impact of our model by running A/B tests on future investigations. Naturally, our approach should also be validated by comparing the outcome of current ongoing investigations with our prioritized list of contracts.

The parameters of the models where :
LR - {'penalty': 'l1', 'C': 10}
KNN - {'weights': 'uniform', 'algorithm': 'auto', 'n_neighbors': 10}
DT - {'max_depth': 5, 'criterion': 'gini', 'min_samples_split': 5, 'max_features': 'log2'}
SVM - {'C': 1}
RF - {'max_depth': 10, 'min_samples_split': 2, 'max_features': 'log2', 'n_estimators': 10}
GB - {'subsample': 1.0, 'max_depth': 3, 'learning_rate': 0.5, 'n_estimators': 10}

We selected the model that had the highest Precision at the top 20% and then as a second criteria chose the value that maximized the AUC of the precision-recall curve which in this case was Random Forest.

Once we applied the RF model to the testing data we evaluated our model and obtained the following results:

Table 4: Evaluation of testing data

Accuracy	.5
F1	.12
Precision	.5
Precision Top N	.5
Recall	.07
ROC AUC	.5

The objective of this model is to provide the World Bank with a list of contracts to be investigated for different thresholds. An example of the deliverable output is found under table 5 where we put the contracts that have the highest expected value (y fitted score x amount).

Table 5: Contracts with highest expected value

contract_id	yscore	expected_value
226	1	19.37
45	0.96	15.04
47	0.96	14.71
44	0.96	14.43
46	0.96	13.75
134	1	13.61
198	0.97	12.93
199	0.97	12.93
324	0.8	12.87

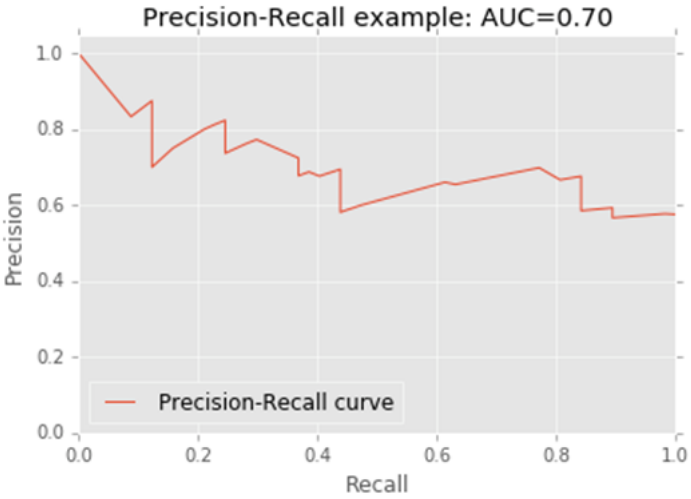

Figure 5: Precision-Recall curve for RF best model

7. LIMITATIONS AND FUTURE WORK

Limitations:

- Analysis is restricted to reported contracts. There are potentially other contracts with misconduct that are never reported.
- Train/test splits are done on the basis of contract award dates. We do not know if these are consistent with investigation dates.(We could, in fact, be training the past with data from the future.)

Future work:

- KNN classifiers to incorporate more features
- Experimentation with a randomized control trial in the field
- Loop over probability threshold and top n% of contracts to maximize precision and AUC
- Refine model to separately predict fraud and collusion (vs. misconduct)