Predicting Fraud in International Development

Github repo: https://github.com/csking1/world-bank-project.git

Charity King
Emily Webber
Greg Adams

1. Background and Introduction

The World Bank creates and funds hundreds of development projects every year. Yet when they receive complaints of fraud, collusion and corruption, they do not currently have a data-driven way to prioritize their investigations.

The DSSG World Bank project is a multi-year project focused on detecting that fraud, collusion and corruption in the World Bank contracts. This project will be an extension of the past team's' efforts for predicting corruption and fraud in World Bank contract bids.

Corruption is one of the biggest impediments to economic growth in 60+ developing countries and drains about $900+ billion from the developing world every year. World Bank anti-corruption efforts currently rely on staff oversight and investigations based on complaints filed via the public Integrity Complaint Form.

For FY15, there were 323 preliminary inquiries. Ninety-nine were selected for full investigations, and 81 were closed. Approximately 61 inquiries were substantiated for corruption. Those corrupt contracts were worth $523 million. We hope to be able to increase the efficiency of these investigations, thereby decreasing investigators' caseloads and increasing the likelihood the World Bank is able to catch fraud.

Currently, the world bank investigates every complaint it gets. Based on those initial investigations, it decides whether to launch a full-scale inquiry into the project, work with operational staff to address the issue, or close the case. Investigators look for complaints with credible information regarding sanctionable practices that significantly impact either the success of large projects or the reputation of the World Bank Group.

This approach is time consuming and subjective. Follow-up investigations are largely based on the reports of the Integrity Vice Presidency (INT) officials, and the window from contract award to investigation is often as long as six years. The investigations themselves usually take between 12 and 18 months. Reducing this lag via automation has the potential to save the World Bank both time and money, allowing them to focus on high-impact areas rather than manually searching for fraud.

In order to automate the fraud detection process, we've developed a machine learning classification algorithm that estimates the likelihood of a given contract being fraudulent. We will then use this likelihood to estimate the expected value of an inquiry and generate a prioritized list to investigate. This prioritized list should serve as a roadmap for administrators and investigators to allocate resources to those inquiries which are most likely to reveal high-impact fraud.

Even a minor improvement in the likelihood of an investigation successfully uncovering fraud has the potential to save the World Bank huge sums of money. Because World Bank investigations are currently tied to the complaint procedures, and because it takes so long to start and complete investigations, the bank is currently running a relatively high risk of working with corrupt suppliers. Reducing that risk with automated fraud detection could save the bank

money and protect it from the reputational damage it currently tries to avoid with its investigations.

More than that, providing a reliable, consistent way for the World Bank to uncover corrupt practices makes their projects less risky. Even a marginal change in the risk of funding a project could help necessary development happen. As an organization dedicated to the promotion of shared prosperity, any improvement in their ability to implement poverty reduction and development projects is both beneficial and necessary.

2. Related work

The Difference in Practice. Investigators do not presently have way an automatic linkage between the contracts, projects, and investigations information. They currently have to manually switch between systems, and our pipeline allows a full three-way integration. Our system provides an immediate merging of these three sources of information that give detailed insight into the contracting process, the project's goals and expectations, and the details of an investigation for each project.

Furthermore, the average lag time from the signing of a contract to the opening of a case using the Bank's existing system is roughly 6 ½ years. Our approach has the potential of reducing this to a matter of days, and ideally influencing the contract signing stage. If a contract shows early signs of being fraudulent based on our algorithm, or if the project over time seems to indicate signs of becoming fraudulent, the Bank can change the terms of the contract or take action to save the project.

The Difference in Method. Because this project has been done by recent DSSG teams, we'd like to clarify how our approach is slightly different. We decided on "case outcome" as our binary variable, instead of "allegation outcome" like last year's DSSG team. This is because a given project can have multiple complaints, and each complaint can be either substantiated or unsubstantiated, while each investigation has only a single result. This means we did not use repeated project IDs in the training set, but instead treated each project as an atomic entry with a binary outcome.

This means that our training set is roughly one half the size of the 2015 team's set, because each entry is a unique reference to a project rather than one of up to five repeats. Because most of the fields for each project come from the contracts and projects data, we didn't think this limiting to a single entry introduced a major flaw into the process. Furthermore, the existence of multiple entries of the same project would overweight this in the classifier selection, each field in these rows would be weighted N times instead of only once.

The Similarities in Method. We are highly indebted to last year's DSSG team for their excellent work in the project, and in many ways our work is another step made from their efforts. The only piece of software we used directly from their team was an entity resolution script written in Python that addresses the problem of multiple names in the contracts csv being spelled differently. Their code is very well documented, and especially their full_pipeline.sh script was helpful in our charting out our approach to the problem.

3. Problem Formulation and Overview of Our Solution

The World Bank receives complaints from around the world that lend insight into whether or not a contract the Bank has awarded was successful in its goal of relieving poverty. Because of limited resources, unfortunately the Bank is only able to investigate roughly half of those complaints. From August 1, 2011 through March 1, 2015 the Bank received 4025 complaints, and opened investigations on 1917 cases. Within those two thousand, 905 were found to be unsubstantiated while 1012 were found to have cause for complaint. The Bank investigates half of its cases, and these investigations find grounds for further action only half of the time.

The time from contract approval to the Bank to receiving a complaint can vary from as low as a few months to as high as twenty years. On average, complaints found to be unsubstantiated came in ten years after the Bank approval for the beginning of the contract, while complaints found to be substantiated have a lag time of only 6 ½ years. Cases not investigated have a lag time of 5 years.

The information regarding project amount is striking. A secondary data source we're using is from the World Bank's public API; it describes all contracts that have been awarded worldwide. There are 207,397 of these contracts, dating from January 21, 2001 to today. The average contract amount out of those two hundred thousand is only $831,039. It seems immediately clear that if a contract is less than this average, it will very likely not be investigated by the Bank. It also seems clear that older contracts are less likely to be found as fraudulent.

Conditional Averages on Investigations Data Set

	Substantiated	**Unsubstantiated**	**Not Investigated**	**Total**
Avg Time from Contract Signing to Case Opening	6 ½ years	10 years	5 years	6 ½ years
Avg Project Amount	$8,561,403	$7,189,675	$562,975	$110,827,358

4. Data Description

We pulled information from three primary sources: projects, contracts, and investigations. The projects and contracts data comes from the World Bank's public API, with the contracts csv totalling roughly 207,000 entries and projects at 16,000. The investigations data is privately held by the Bank and totals nearly 8,000 lines split between two csvs. We'll only discuss the first of these two.

The Banks' investigations data has a remarkable number of repeats. The first of the two csvs has just over 4,000 rows, but only 171 unique cases listed. While a lower number is to be expected, because the Bank does not investigate every complaint, the repetition in Case ID is

striking. A case is opened when the Bank receives a number of complaints, so a given case encompasses multiple projects, companies, and contracts.

Surprisingly, a three-way join of these sets on project id leads to only 380 unique project listings. This is the case for a few reasons. First, the Bank can receive multiple complaints for a single project. This means their unique identifier in the investigations set is “Allegation ID”, not “Project ID.” Removing all repeated Project IDs leads to an investigations set of only 390 unique projects. The number of unique complaint IDs is only 460. The number of unique World Bank IDs in the investigations set is higher, with 891 unique entries. There are 883 unique contract names in the Banks’s investigations set, which is still considerably lower than the 4000.

Secondarily, the dramatic reduction in size of the data sets simply indicates that the Bank is overwhelmingly likely not to receive a complaint on a project. This is very good news; only 0.19% of contracts receive a complaint, and only 2.375% of projects receive a complaint. Projects are more likely to receive a complaint simply because there are less of them, a given project can have multiple contracts.

It’s likely the Bank’s investigations data would have better modeling results if there were more reference to unique investigations. By tying so many complaints to so few project IDs, it’s difficult to capture the variety of the complaints for modeling. Our final section closes with recommendations for the Bank’s system of collecting and storing data to resolve some of these questions and gain more utility out of their own information.

5. Details of our Solution: Methods, Tools, Analysis, Models, Features

Current Feature Generation: All World Bank column data had to be transformed in one way or another. Perhaps a reflection of our model results, most of the columns were of a categorical nature and every column was represented as a string. The process that evolved for feature generation with a varied dataset that may have to undergo multiple transformations was as follows. Ultimately, we created a mostly sparse matrix comprised of approximately 1,526 columns.

I. Create a transformation function:
    A. Logging or normalizing
    B. Binning
    C. Categorical Dummies
    D. Categorical to Numericals (mostly binary, also handles over two values)
II. Choose Transformation for each column variable
    A. Column exploration for range of values, missing data, data format problems, etc.
    B. Imputing data if necessary or designating that a value was missing via a separate column value creation
III. Create Global Transformation Lists
    A. Various Transformations lists, i.e. BINNING_LIST, DUMMY_LIST, LOG_LIST, BINARY_LIST, etc.

B. Create a drop list to clean up columns either not being used currently as well as columns that maybe are redundant or offer no value. A pandas unique.() search on columns such as ‘lendprojectcost’ and ‘ibrdcommaamt’ indicate values of either 0 or NaN. The other important aspect of the drop list is to add original categorical columns such as columns used for dummy variable creation since the original columns were mostly in unfit states for machine learning algorithms.

IV. Feature Loops: In combining all of these steps together, we simply implement one function in our gen.py file that transforms our dataset into usable X, y vectors for our various models. The data is transformed as such:

A. CSV to Pandas Dataframe read-in
B. Dataframe Conversion of string to numeric format when applicable
C. Fix predictor variable
    1. This step is important in tailoring what the outcome will be. The variable that we used as our predictor, ‘case outcome’ has 5 different values:;
        a) 'Substantiated'
        b) 'Unsubstantiated'
        c) 'This case is closed to be merged in case No. C-FY2015-2527-ZR as both cases belong to the same project. '
        d) "INT has determined that the allegations against Mr. tervino do not fall within INT's mandate of investigation under SR 8.01. INT has notified Karim E. Kemper accordingly. No further action. "
        e) 'Unfounded' 'Referred'
    2. We ultimately decided on only predicting whether a case was substantiated or not.
D. Loop through each Global List within each corresponding transformation function
E. Feature Importance Output: Provides an overall importance distribution of the features matrix and output on the top K most important features utilizing a Random Forest Classifier.

Transformation Type	Column Variables
String to Number Typedef	All applicable columns, numbers represented as strings
Categorical Dummy Transformation	Region', 'Fiscal Year', 'Procurement Type', 'Procurement Category','Procurement Method', 'Product line', 'Major Sector_x', 'Supplier Country', 'Supplier Country Code', 'resolved_supplier', 'allegation_category', 'allegation_outcome', 'allegation_type',

	'complaint_status','country', 'lead_investigator', 'major_sector', 'procurement_method_id', 'regionname', 'vpu', 'prodline', 'lendinginstr', 'lendinginstrtype','board_approval_month', 'borrower', 'impagency'
Log (Normalizing)	contract_amount'
Binary	‘caseoutcome' (Y variable), 'project_amount'
Binning	('project_amount', 50 bins)

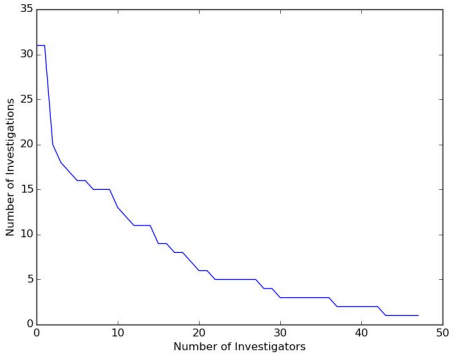
Top K=10 Performing Features using a Random Forest Classifier, MDI score
1. feature 985 ‘vpu_ECA’ (0.091976) Note: VPU_ECA designates the World Bank’s Vice Presidential Unit Europe and Central Asia region
2. feature 936 ‘lead_investigator_Samuel (Sam) BWANA’ (0.081429)
3. feature 791 ‘allegation_type_b.Abuse of position Other’ (0.027449)
4. feature 277 ‘Major Sector_x_Finance’ (0.021938)
5. feature 954 ‘major_sector_Global Information/Communications Technology’ (0.018492)
6. feature 755 ‘allegation_outcome_The allegations involve the supply of ARV products for contracts’ (0.016090)
7. feature 941 ‘lead_investigator_Thilda Outhuok’ (0.011192)
8. feature 761 ‘allegation_type_Collusion - Bid manipulation - PIU staff’ (0.010480)
9. feature 906 ‘lead_investigator_Christopher KIM’ (0.010399)
10. feature 964 ‘procurement_method_id_CQS’ (0.009773)

Feature 1 'vpu_ECA' Analysis:

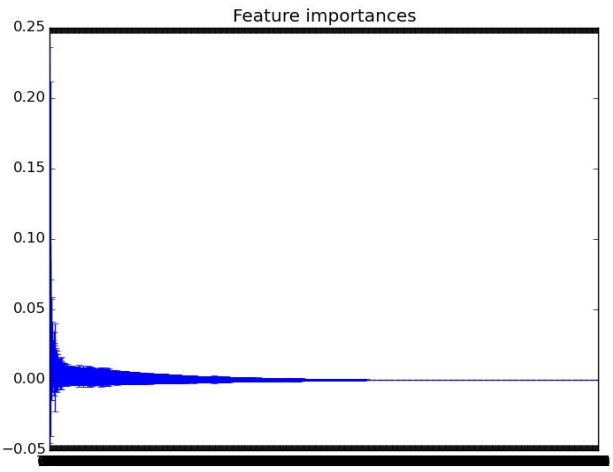
57/380 rows were designated as ECA. Of the 57, 51 were classified as "substantiated". It is unclear however why ECA was selected the top feature as other VPU designations have similar if not higher correlations to the outcome variable. Other designations for substantiated ratios were as follows:

Feature 2 'lead_investigator_Samuel (Sam) BWANA' Analysis:

Of the 48 Investigators that have headed investigations, Samuel Bwana is tied at 10th place with 11/380 investigations with the two top investigators, Simon ROBERTSON and Thilda Outhuok each having 31 investigations. Of Mr. Bwana's 11 investigations, 10 were "substantiated." Again, there are stronger correlations with the outcome variable and other investigators. Ms. Thilda Outhuok had 31/31 of her investigations substantiated.



Feature Importance: The majority of our feature creation has resulted in a "sparse" matrix comprised mainly of 0s in a memory-efficient capacity. With over 1500 features generated, we utilized the built-in sklearn feature importance algorithm that utilizes a Random Forest Classifier to measure the Mean Decrease Impurity (MDI) for each variable. This method directly measures the impact of each feature on accuracy of the model. The general idea is to permute the values of each feature and measure how much the permutation decreases the accuracy of the model. Ultimately, features that have no effect on accuracy rates will result in 0. Of our 1,526 features, 441 features are 0 and approximately 1,496 are under .005. The graph below designates how the most important features represent a very small fraction of the feature distribution with an asymptotic decline towards 0.



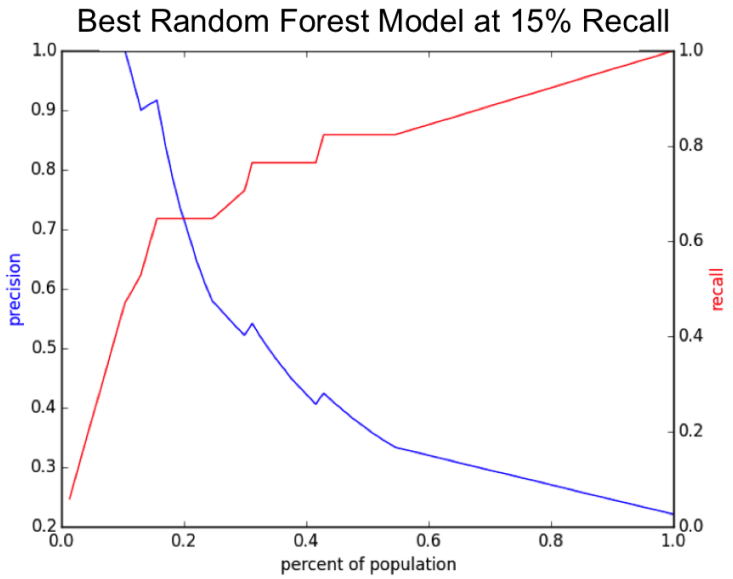
Features for the future: Feature generation proved to be much trickier than anticipated, specifically in creating and validating meaningful features that contribute to decreasing a dataset's MDI, especially on such a small dataset of 380 rows. There was data information that was not used simply due to running out of time. It was recommended to us from the last team to create a feature based on when a case was opened and closed in terms of number of days. This feature would have been created as a difference of days and potentially binned.

Another feature not implemented was creating and utilizing a normalized contract/project amount using the Purchasing Power Parity to normalize for differences in national GDP and living costs.

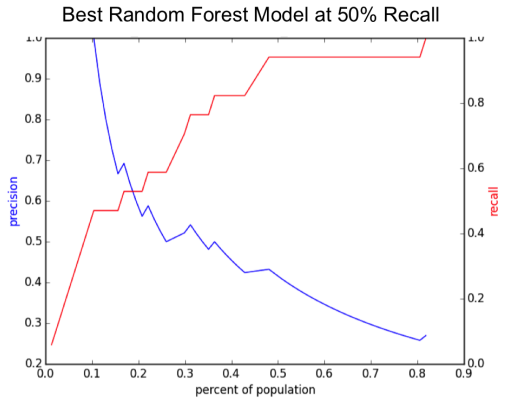
6. Evaluation

Our best model overwhelmingly has been Random Forests. We were able to find the most consistent results and the best ROC curves using this classifier. Many of the other classifiers yielded extremely odd ROC curves, which we think may have been problems with overfitting or perhaps our pipeline. We optimized for precision at various levels of recall, and found that 15% recall gave us the best and most believable precision: 91.6%. Accuracy at this level is roughly 90%, with an AUC of 82.9%. Exact specification of the model and precision-recall curve is below.

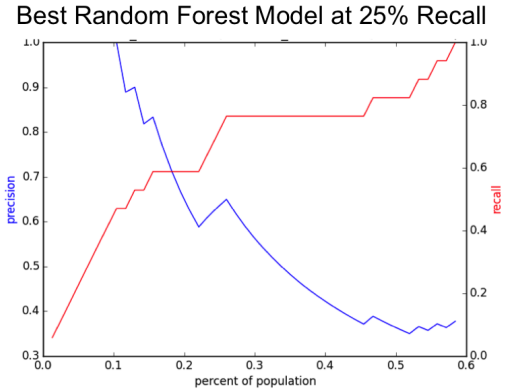
```
RandomForestClassifier(bootstrap=True, compute_importances=None,
            criterion='gini', max_depth=100, max_features='log2',
            max_leaf_nodes=None, min_density=None, min_samples_leaf=1,
            min_samples_split=10, n_estimators=10000, n_jobs=-1,
            oob_score=False, random_state=None, verbose=0))
```



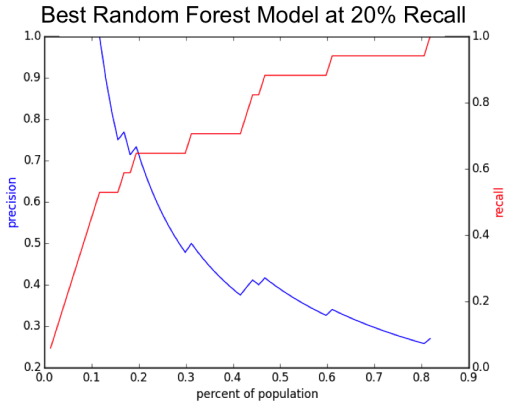
Our initial choice for recall was 50% to reflect the Bank's' own likelihood of responding to a complaint. At this level precision is quite low at 41%, but AUC is still strong at 86%. Interestingly, exactly the same Random Forest specification was selected as the best performing under each threshold for recall.



With a drop in recall to 25%, our precision jumps to 65%. The AUC statistic is perhaps at its highest here at 87.55%.



With recall at 20%, the precision of our classifier increases only slightly to 68.75%, with AUC still solidly in the 80s. Below is a precision-recall curve for this model. It's clear to see that precision consistently follows the pattern of a very steep drop for the top 20% of the sample, with a moderate decline afterward. The stepwise progression of recall is unique to this model specification.



As soon as we specify 10% recall, our precision hits 100% with disconcerting consistency. We recommend choosing levels of recall higher than 10% to avoid this problem. The drop from 20% recall to 15% yield incredible results in terms of precision, from 60% to over 90%, so we chose recall at 15% to use for optimizing precision. This jump in precision is reflected in almost all of the graphs, so this is no surprise.

7. Discussion of Results

a. Predicting Substantiated Complaints

We set out to predict which complaints received by the World Bank would be substantiated. Further, we hoped to predict, at the time of signing, which contracts the World Bank would receive complaints about, investigate, and find corruption.

Despite initial problems in predicting In building our models, we found some success in predicting corruption. Some of the initial problems had to do with each model's ability to predict unsubstantiated allegations, but the random forest models proved able to compensate for those problems and score highly on each evaluation metric across varying levels of recall.

b. Feature Importance

One of the strengths of random forest classifiers is their ability to effectively identify the features with the most predictive value. For the model we chose, two features stood out by a large margin: "vpu_ECA" (referring to the World-Bank-defined region of Eastern Europe and Central Asia) and "Lead_Investigator_Samuel" (presumably referring to an investigator at the World Bank). With MDI scores almost four times the third most important feature, these variables seem to have strong predictive value.

The predictive value for corruption of a specific region of the world is not necessarily surprising. Corruption, both in terms of local definition and prevalence, varies widely throughout the world. It is interesting, however, that Europe and Central Asia would be so predictive of substantiated claims of corruption, given that no other region is in the top ten. It merits further research--outside the scope of this project--why the ECA region would be so predictive.

As for the predictive value of "Samuel" as the lead investigator, we do not have enough data to make strong claims about the reasons. Speculatively, there are a variety of possible reasons that an individual investigator may be involved in a large number of cases found to be unsubstantiated: specialty in terms of region or industry, experience level, case-assignment patterns or random chance could all contribute. In order to identify whether this lead investigator has a causal relationship with findings of unsubstantiated complaints, we would need to analyze data on his assignments specifically, which is (again) outside the scope of this project.

c. Evaluation Concerns

Some of our initial problems with model-building involved predictions that were unusually and unrealistically high. While this may have been an issue with the code we were using, it could also have stemmed from the small dataset we used. With only 300 lines in the training set (and 77 for testing), it was easy to overfit the model and difficult to avoid the impact of random noise. Regularization helped, but the size of the dataset itself is simply too small to confidently

predict a much larger potential set of outcomes. The size of the dataset is also the most likely reason for the rather jagged precision and recall curves above.

9. What does this mean for the World Bank?

From a data point of view, we recommend the Bank consider its investigations on a contract-by-contract basis, rather than a project-by-project. This is because a given project can have many contracts, and only one of these contracts might prove to be corrupt, but the entire case will usually be treated in a single manner. This will increase the usability of the Bank's proprietary information by increases its ability to join with the larger projects and contracts data sets, and fully exploit the richness of the set.

From an operations point of view, we would like to test our findings by organizing two on-the-ground teams of investigators. One team will be the control group, who are given a standard list of contracts to investigate that follows their standard procedure. The second team, the intervention group, will be given a ranked list of contracts to investigate that has been optimized using our classifier and prioritizing the complaints based on highest expected value. We'll know the usefulness of the classifier after we see the results of both teams in multiple trials. On average, the classifier will be useful to the extent it improves both the number and dollar amount of contracts found to be fraudulent that would not have been otherwise identified.

This type of A/B testing is common in technology firms that have access to large amounts of data which are updated frequently, but is less common in more traditional organizations. While in some cases there are ethical and budgetary concerns regarding the creation of a control group, in the case of the World Bank there seems to be no harm in the creation of two otherwise identical teams that are simply given contracts to investigate in different manners.

We hope this type of data-driven collaboration across offices within the Bank, such as between groups that award contracts and those that investigate complaints, will encourage better data-sharing habits that increase the efficiency of Bank operations, reduce caseload, and ultimately ensure the quality and reliability of international development projects.

10. Limitations, caveats, and future work

This project is designed to have two potential applications. First, in the prioritization of existing allegations that have not been investigation. Second, in the flagging of a contract as having a high likelihood of being fraudulent. In the first case our methodology seems to be reasonable. In the second case, however, we notice that further specifications will be required in the gathering of contracts and projects data. Currently we pull everything available on the project and contract at the time of the query, but this is not useful for predicting the status of the contract *at the time of the signing*. To generate an early warning system, we would need to pull the data on the contracts and projects at the time of the signing, and not seven years after the fact.

In the future we'd like to use more of the information we already have available to us. Presently we're dropping roughly 20 columns because they are textual data that we're not able

to process, but with more time we'd be able to generate insights from these. We'd also like to generate more aggregate features, such as the number of days between a contract signing date and its first complaint, the number of contracts per project, and the number of complaints received per contract.

In terms of processing the data, we'd like to work with how we're joining the sets, perhaps by experimenting with outer joins to grab more of the information. Especially it would be useful for the Bank to link each complaint with a contract ID, and not just a project ID, as this will make the data they already have more useful for modeling.