What factors could detect the potential bad performance bank projects

Methods and Initial Results

Zhuo Leng
May/2017

1.Goal

The goal of my research is to study the factors that could detect the potential bad performance of bank projects.In other word, I'm going to select most relevant features to build World Bank Project Performance Prediction Model.

2.Context and Data Source

I will use data from two main sources:

World Bank Projects API
(http://search.worldbank.org/api/v2/projects)
It provides access to basic information on all of the World Banks lending projects from 1947 to the present. It includes information such as: project title, task manager, country, project id, sector, themes, commitment amount, product line, and financing.

World Bank Project Performance Ratings
(https://finances.worldbank.org/Other/IEG-World-Bank-Project-Performance-Ratings/rq9d-pctf)
This dataset contains all World Bank project assessment carried out by Independent Evaluation Group (IEG) in the 70s, spanning over 30 years. It includes more than 11,300 projects assessments, covering more than 9,600 complete projects.

For Data I: World Bank Projects API, it grants access to all of the World Banks projects, including closed projects, active projects and those in the pipeline. The queries a user can make are listed but not limited to the following: catalog source, country, income level, indicator, lending type, topic and regions. There is a pilot geocode data that shows the project location. However, this data are collected through the study of existing project documents and thus, need further data validation and quality enhancement.

Data II: World Bank Project Performance Ratings performed by Independent Evaluation Group (IEG), which perform the assessment on completed World Bank lending operations through reviews all implementation completion and results reports through ICR Reviews, and uses the same evaluation criteria as World Bank project teams. Building on that, IEG performs in-depth field-based evaluations, called Project Performance Assessment Reports (PPARs). The evaluation contains outcome ratings data on World Bank lending projects based on a six-point scale: Highly Satisfactory - Satisfactory - Moderately Satisfactory - Moderately Unsatisfactory - Unsatisfactory - Highly Unsatisfactory. Other variables in the evaluation include closing fiscal year, Global Practice, Region, and/or Country etc.

2.1 Summary Statistics

descriptive statistics table

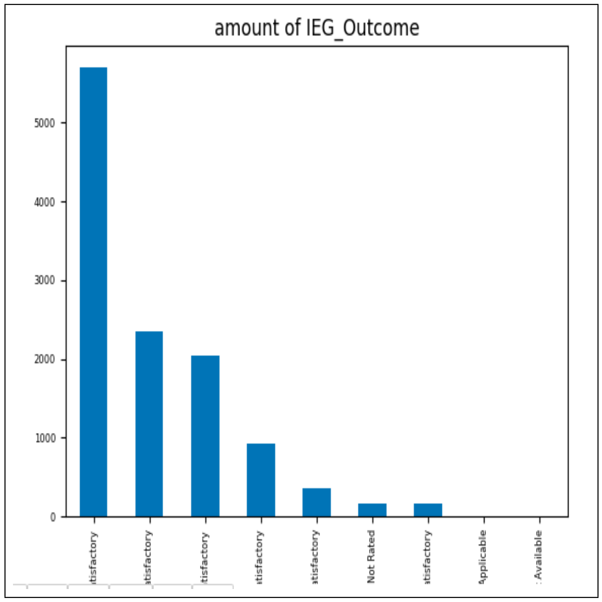
Table 1: descriptive statistics

variables	count	mean	std	min	max
Approval FYs	11726.0	1991.0	11.45	1948	2015
Lending Project Cost	6.23e+03	1.66e+08	4.99e+08	6.00e+00	2.69e+10.0
Net Commitment	1.15e+04	6.43e+07	1.18e+08	-7.46e+08	3.00e+09
Exit FY	11726.00	1994.07	76.77	0.00	2016.00
IEG$_{E}valFY$	111726	1999.19	10.55	1973.00	2017.00
ERR at Appraisal	4566.00	25.32	19.94	0.27	541.50
ERR at Completion	3979.00	22.23	28.85	-100.00	747.00
grantamt	1.17e+04	6.55e+05	4.69e+05	0.00e+00	1.19e+08
totalcommamt	1.17e+04	7.33e+07	1.25e+08	0.00e+00	3.00e+09
totalamt	1.17e+04	7.26e+07	1.25e+08	0.00e+00	3.00e+09
themecode	6.84e+03	2.07e+09	2.95e+09	2.10e+01	9.39e+09

The tables below present basic descriptive statistics for numerical features. We can already see some anomalies in the data, such as projects with a zero value for amount features and Exit FY with value zero.I will do data visualization of some categorical features in next step.

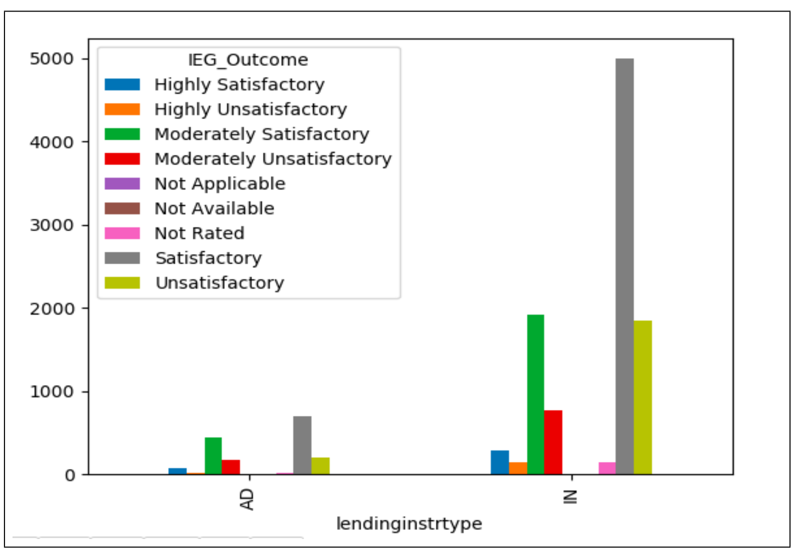
2.2 Visualization of Data

plot1: Distribution of target dependent variables(label):IEG Outcome



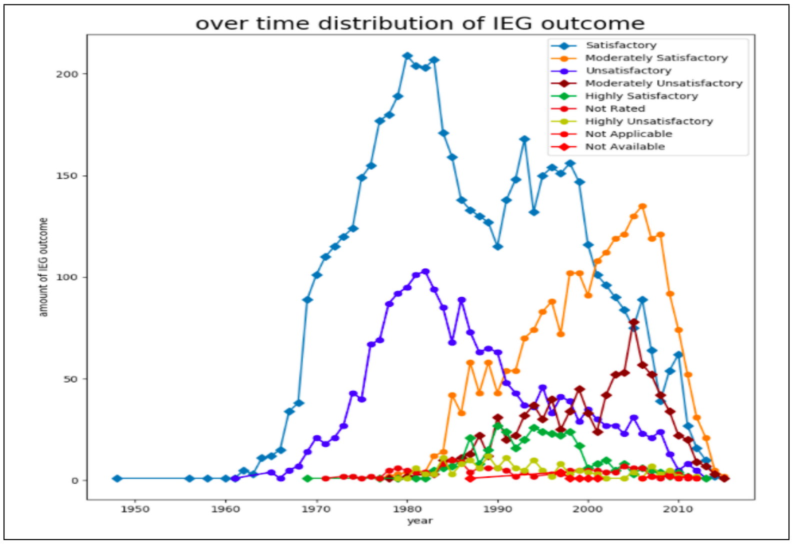
After we count the number of projects across outcomes, we could see that a large number of them have Satisfactory outcome (5697 projects), followed by Moderately Satisfactory.

plot2: Number of projects by lendinginstrtype



From the crosstab bar plot, we could know that the amount of IN lendinginstrype is far more larger than that of AD lendinginstrype. The proportions of each IEG Outcome seem similar in different lendinginstrype.

plot3 Over Time Distribution by Approval FY



Above is the distribution of different type of IEG outcome over approved by Fiscal Year. We could notice that from 1960-1980, the amount of all types of IEG outcome continuously increase. By Approved year of 1980, the amount of satisfactory project is far more than that of others. However, from 1980, the number of IEG decrease lead to amount of all IEG outcome decrease except for Moderately Satisfactory and Moderately Unsatisfactory. By around 2010, the amount of Moderately Satisfactory exceed Satisfactory, rank first among all kind of outcome, which means with increase of Approval FY, project performance are not satisfactory than before.

3. Research Methodology

Because world bank current project investigative process is complaint-driven and not accurate (2010), so I want to use machine learning system to detect world bank project bad performance by using classification system. Thus I could create automated mechanism to help them save cost and times as well as improve rate of substantiated investigation.

First I need to integrate the two data source and reduce the dimension and redundant data by implement PCA-Principle Component Analysis. Then the next step is to build a classifier to assign 'performance score' to World Bank Project. I will firstly try Random Forest, SVM, Decision Tree classifier. I will mainly do these four step:
1) Data integrate: linkage
2) Data pre-process: generate features using PCA
3) Modeling: SVM , RF, DT
4) Evaluation: AUC, Precision, Recall, f1

3.1 Principal Component Analysis

PCA is one method used to reduce the number of features used to represent data.It projecting data from a higher dimension to a lower dimensional manifold such that the error incurred by reconstructing the data in the higher dimension is minimized.PCA methodology could captures the dominant nonlinear features of the original data by transforming it to a high dimensinal feature space(Hotelling, H., 1933).In order to get rid of overfitting as well as reduced complexity and run-time, we could use PCA in our pre-process step to help us generate feature.

The first principal component of a set of variables x1, x2,xp,that has the largest variance.

$$Z1 = \phi_{11}X_1 + \phi_{21}X_2 + \ + \phi_{p1}X_p$$

By normalizing the features,

$$\sum_{n=0}^{p} \phi_{j1}^2 = 1$$

The elements of $\phi_{p1}$are known as the loadings of the first principal component, and combined together to form the principal component loading vector.

After I use PCA to reduce dimension, the principal components that explain 99% of the variance are kept, and the rest are thrown out.

3.2 Model Analysis

After I reduce dimension of my dataset by using PCA, I will use the new features to implement my machine learning classifier.I utilized three classifiers in my preliminary models this time:
1.Random Forest
2.SVM
3.Decision Tree

Then I will rank model excellence based of these four metrics (AUC, Precision, Recall and f1). Then I will continuously adjust my model parameters and add new features in it until I got the best model.

4.Initial Result

4.2 Pre-process

Our goal is to create a model which is able to predict and detect the potential bad performance projects of the World Bank at the time of the projects publication. The dependent variable of the data set we care about is the performance outcome evaluated by the World Bank. It is a categorical variable with three levels of satisfactory and unsatisfactory. There are 180 projects out of 11,717 that lack of evaluation. I drop these rows since it is a fairly small number and we do not have better way to fill them. We divide the performance outcome into two groups, good performance and bad performance. And we convert the dependent variable into a Boolean variable with True for good performance and False for bad performance.

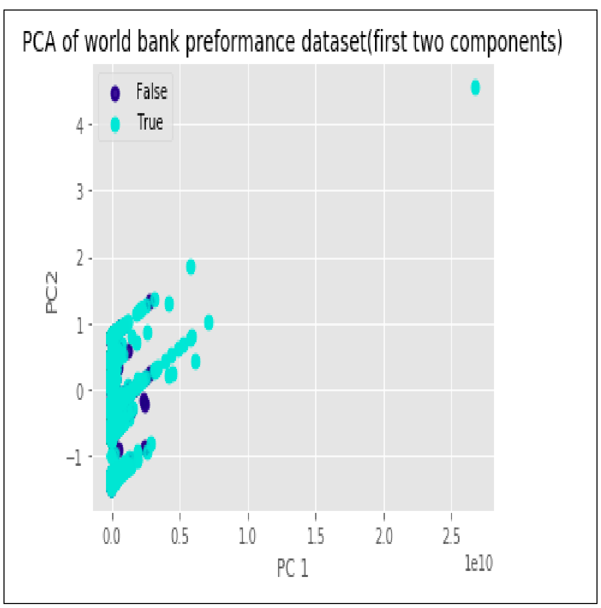
Because the variables of this dataset are around 100, after did data exploration, first I selected to use the following feature:

Table 2: first round feature selection

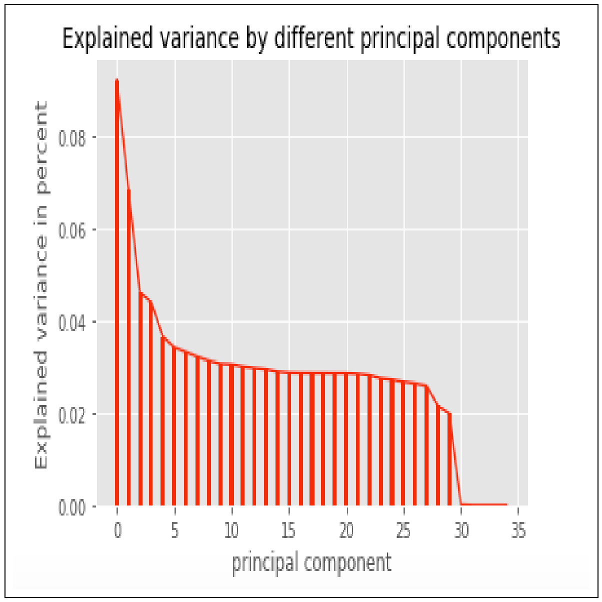
Name	Date	Porduct Line	lendinginstr	lendinginstrtype	lending project cost
Type	Date	Categorical	Categorical	Categorical	Numeric

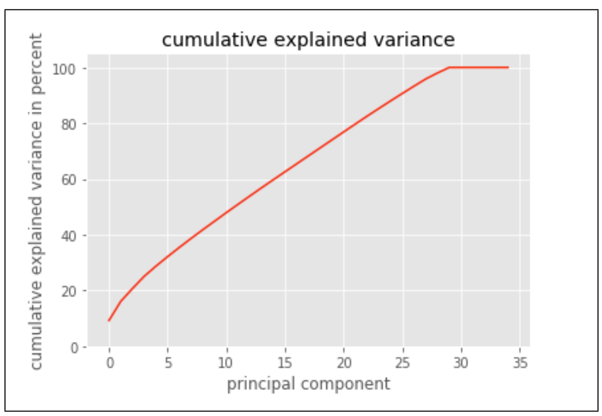
Because this is the initial computational result, this time I just selected 5 feature by myself to do the PCA , so it's not that accurate. For final version, I will use all features to do principal component analysis in order to reduce dimensions and extract important features.

4.2 PCA



I first do PCA of world bank performance dataset (first two components) to get basic sense of the whole data. False means Failure Outcome and True means Satisfactory outcome. We could see from plot that Satisfactory outcome could more be explain be pc1 positively.





By doing PCA on selected 35 features, We could know the first five principal components seems could highly explain variance in percent by different principal components compare with the rest of components. After 30th components, the feature seems could not explain variance. In this draft, I will not try to transform data after PCA. I just want to get a good understanding of data sahpe for this time. I will try that in final report.

4.3 Model Evaluation

I build a magic loop to loop through all initial parameters I created for my three types Classifier (Random Forest, Decision Tree, SVM) base on above part: RF: n estimators: [1, 10], max depth: [1, 5, 10, 100], max features: ['sqrt', 'log2'];SVM: c :[0.1,1]; Decision Tree:criterion: ['gini', 'entropy'],max depth: [1, 5, 10, 20, 50, 100], max features: ['sqrt', 'log2'], min samples split: [2, 5, 10]

The best evaluation result of each kinds of classifier is as below:

Table 3: Result Evaluation

Classifier	AUC	f1	Precision	Recall	Parameters
Decision Tree	0.7787	0.551119	0.0222395	0.803518	$\min_s ample_split : 2, max_depth : 100$ criterion:'entropy', 'max$_f eatures' :' log2'$
SVM	0.84225	0.499901	0.284046	0.996482	c:1
RF	0.843031	0.508309	0.00988301	0.992436	'criterion': 'entropy', 'min$_s amples_split' : 2$, 'max$_d epth' : 10,' max_features' :' log2'$

Overall, it seems decision tree is the highest performing classifier. Resulting implications need to be analyzed further, but we can hypothesize that this performance can be attributed to the shape of the data features we created.In absolute terms, performance for all models is low. This is likely rooted in proper feature selection.I will run with an assortment of features as a next step.

5.Reference

World Bank what we do[Online]. Available: http://www:worldbank:org/en/about/what-we-do.

Transparency International.Corruption perceptions index[Online]. Avail- able: http://www:transparency:org/cpi2010/results,2010.

The World Bank Group Integrity Vice Presidency [Online]. Available: http://siteresurces.worldbank.org

World Bank annual report. 2015.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417441, and 498520.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28, 321377

Abdi. H., Williams, L.J. (2010). "Principal component analysis". Wiley Interdisciplinary Reviews: Computational Statistics. 2 (4): 433459. doi:10.1002/wics.101