How country level indicator and project level measures affect bank project performance

Zhuo Leng | zleng@uchicago.edu | University of Chicago | github.com/zhuoleng1

Abstract

As the largest international development bank, the world bank group awards proximately 20,000 -30,000 contracts annually. The monitor and evaluation of these projects help projects achieve their objects and thus, are critical to the world bank operation. There are numbers of reason that lead to underperformance projects, such as fraud and corruption.

Giving the importance of project evaluation and prediction, I conduct a research on how country-level indicators such as GDP and life span and project level factors, such as project duration affect the potential performance of world bank projects. I use machine learning techniques to evaluate different models and eventually identify the best model to approach the research question. My findings offer important insights to help world bank improve rate of substantiated investigation.

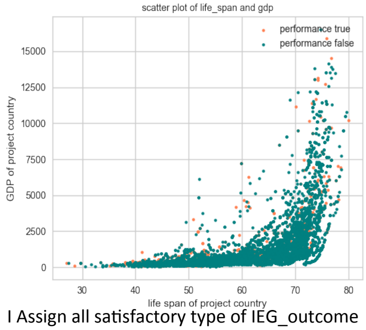
Data

- World Bank Projects API
- WB's IEG evaluation
- Worldwide governance indicator
- Worldwide development indicators

Feature Type	Features Selected
Numerical	leading project cost, ibrd commitment amount, ida commitment amount, total commitment amount, life expectancy at birth, the GDP index
Categorical	region, country name, product line, leading instr type, agreement type
Aggregated	total number of projects done for the past years by the each country, total commitment amount for the past years, average GDP, average life span index, cost-commitment ratio, project duration time

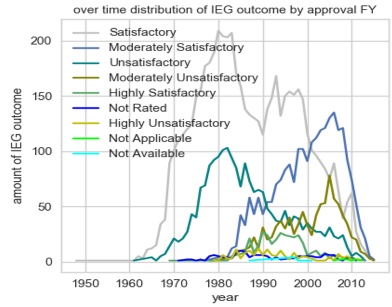
we get 11726 projects evaluation from 1948-2015 . 27% of project evaluated as 'unsatisfactory'.

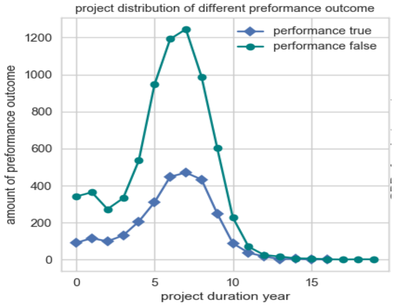
How Does country related indicator(GDP) influence IEG outcome(performance evaluation)?



I Assign all satisfactory type of IEG_outcome as True, and others as False.

How Does Project related feature influence IEG outcome(performance evaluation)?





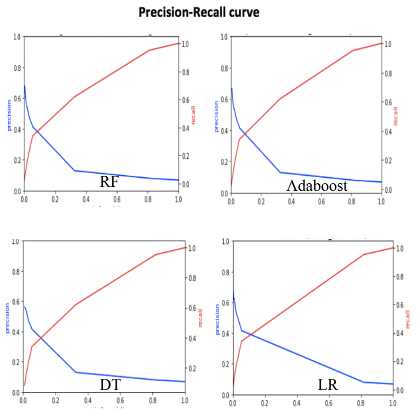
Above is the raw evaluation types. Below, I assign all satisfactory type of IEG_outcome as True, and others as False.

method

- Group IEG outcome(label): assign all satisfactory type as 1, and others as 0.
- Use machine learning system to build classifiers and generate 'performance score' of world bank project:
  - ➢ Random Forest
  - ➢ Adaboost
  - ➢ Decision Tree
  - ➢ Logistic regression
- Rank model excellence based of these four metrics (AUC, Precision, Recall and f1).
- Use random forest to select very importance features and use these features in logistic regression model to better interpret the coefficient.

Results

Precision-recall curve



From the precision-recall curve, RF and Adaboost are slightly better than other two.

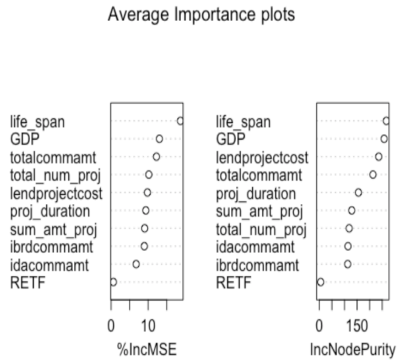
AUC-ROC evaluation

Model type	Parameter	AUC-ROC
RF	'max_depth': 50, 'min_samples_split': 10, 'n_estimators': 10, 'max_features': 'log2'	0.721645
DT	'criterion': 'entropy', 'max_features': 'sqrt', 'min_samples_split': 10, 'max_depth': 5	0.692314
AB	'algorithm': 'SAMME.R', 'n_estimators': 100	0.721576
LR	'C': 0.001, 'penalty': 'l1'	0.673455

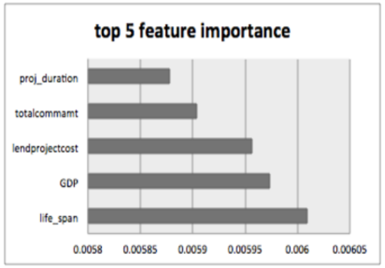
Random Forest classifier is the best model and it gets the highest AUC-ROC score.

Random Forest:

Select Importance Features



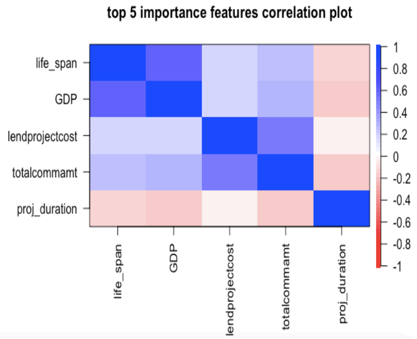
Top 5 importance feature



Model 1: Because GDP, lendprojectcost, totalcommamt are skewed, take log on these three variables

- glm(performance ~ life_span + log(GDP+0.1) + log(lendprojectcost+0.1) + log(totalcommamt+0.1)+ proj_duration)

High correlation between Life_span & GDP: lendproj ectcost & totalcommamt



Model 2: Add interaction between life_span and GDP, lendprojectcost and totalcommamt

- glm(performance ~ life_span+ log(GDP)+life_span*log(GDP)+ log(lendprojectcost) + log(totalcommamt) + log(totalcommamt)*log(lendprojectcost)

	z value	Pr(>|z|)
(Intercept)	-1.040	0.29854
life_span	-0.082	0.09344
log(GDP)	4.299	1.72e-05
log(lendprojectcost)	0.590	0.55525
log(totalcommamt)	-0.132	0.00894
proj_duration	2.770	0.00561
log(lendprojectcost + 0.1):log(totalcommamt + 0.1)	-0.373	0.04089
life_span:log(GDP + 0. 1)	-0.007162	0.002

#AIC decreases from 11426 to 11421. Model3 is preferred.

Interpret the coefficient

- glm(performance ~ life_span+ log(GDP)+life_span*log(GDP)+ log(lendprojectcost) + log(totalcommamt) + log(totalcommamt)*log(lendprojectcost)

(Intercept)	0.03695109
life_span	0.99870102
log(GDP + 0.1)	1.07817265
log(lendprojectcost + 0.1)	1.11779723
log(totalcommamt + 0.1)	0.97615142
proj_duration	1.02524497
life_span:log(GDP + 0.1)	0.99286315
log(lendprojectcost + 0.1):log(totalcommamt + 0.1)	0.99613901

conclusion

- Random forest is the best model for my feature set, among adaboost, LR and DT.
- Llife_span, GDP, these two country level indicator and lend project cost, total commitment cost, project duration these three project level indicator could highly affect world bank project..
- GDP is the most significant feature that could affect project performance. 1% increase in GDP results in 0.73 increase in log-ration of performance satisfaction likelihood. 1% increase in GDP results in 1.08 times higher performance satisfaction likelihood.
- 1 unit(year) increase in project duration results in 1.02 times higher performance satisfaction likelihood.

Limitation

- GDP variable is the mean GDP between the start date and end date. However, projects' success depend to little on the starting condition.
- More country level indicator data, such as inflation rate, employment rate and crime rate etc., is necessary
- After join country and government data, the number of projects reduced about ¼ because of missing value.