Research Design

We have a city year panel that is 2018-2023

Predicted Outcome:

LGFV Interest-bearing Debt/GDP (JB)

OR

Non-Standard Financing Balance /LGFV Interest-bearing Debt(%) (this one already exists in the dataset) (Matt)

The first one is more reliable as the second variable has way more missing values.

We run both OLS and logistic

For logistic: LGFV Interest-bearing Debt/GDP=20% is the cutoff; Non-Standard Financing Balance /LGFV Interest-bearing Debt(%) = 30%

Alternatively, look at the distribution of two variables, and draw a boxplot first.

These two are continuous variables, we could choose a criteria and set a binary variable called default risk (high risk=1, low risk=0), or a categorical variable

(OR

k-means based on LGFV Interest-bearing Debt/GDP, Non-Standard Financing Balance /LGFV Interest-bearing Debt(%), debt ratio) – not for milestone

Feature Engineering:

GDP,

Growth Rate of GDP,

Real Estate Investment,

Per Capita Disposable Income,

Comprehensive Financial Resources,

LGFV Interest-bearing Debt,

Non-Standard Financing Balance,

Fiscal Self-sufficiency,

Budget Revenue,

Revenue of Government-Managed Funds,

*State-owned Land Transfer Income(or State-owned Land Transfer Income/Budget Revenue) – try out both, and tell me how they perform, and leave one variable in the feature engineering.*