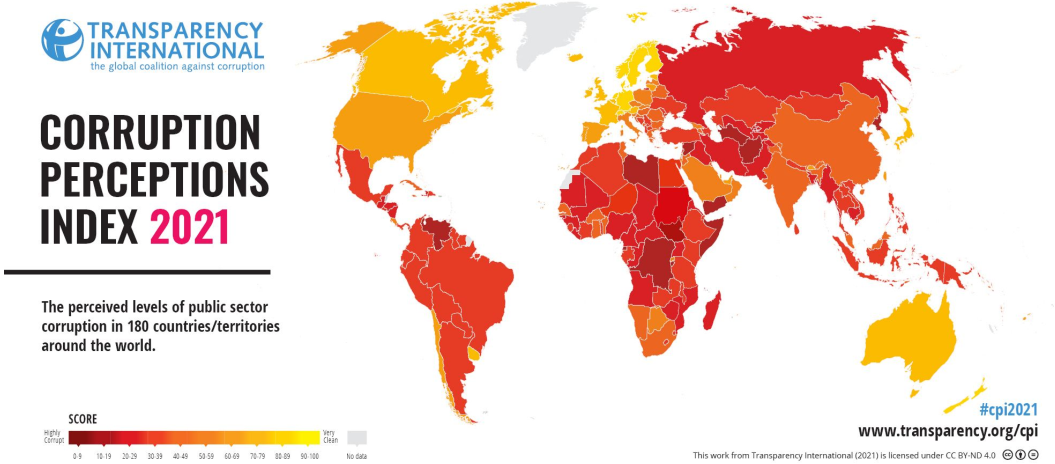
Prediction of Corruption Perception Index

Yuanhan Peng

Facts


TRANSPARENCY INTERNATIONAL
the global coalition against corruption

CORRUPTION PERCEPTIONS INDEX 2021

The perceived levels of public sector corruption in 180 countries/territories around the world.

Glencore recently paid $180 million to Congo to settle corruption allegations between 2007 and 2018.

131 countries have made no significant progress against corruption in the last decade according to the Corruption Perception Index.

Amidst the COVID-19 pandemic, many countries use this as an excuse to curtail basic freedoms and corrupt public funds used for national financial relief.

Sources: transparency.org, ca.finance.yahoo.com, www.ibac.vic.gov.au

Motivation

What is Corruption Perception Index (CPI)?

- Corruption Perception Index (CPI): an index that ranks a country's perceived levels of public sector corruption.
- Assessed by experts and business institutions through public opinion surveys.

How does CPI affect global economics and/or life quality?

- Illegal financial or unreported economic activities hidden from the public.
- **Financial stability**, **social inequality**, **market economy**, and **trust in public institutions**.

Why do we want to study this topic?

- We want to determine which specific features significantly affects a country's CPI
- Suggest steps countries may take to reduce corruption.

Source: https://www.transparency.org/en/cpi

Research Question

Main Question:

- Which machine learning model has the highest accuracy for prediction of Corruption Perception Index (CPI)?

Sub Questions:

- Which socioeconomic features are the most influential indicators of corruption perception index?
- What the government can do to reduce corruption?
- Which factors in Human Development Index and Economic Freedom Index contribute most in the prediction of CPI?

Answer to the research questions & Contribution

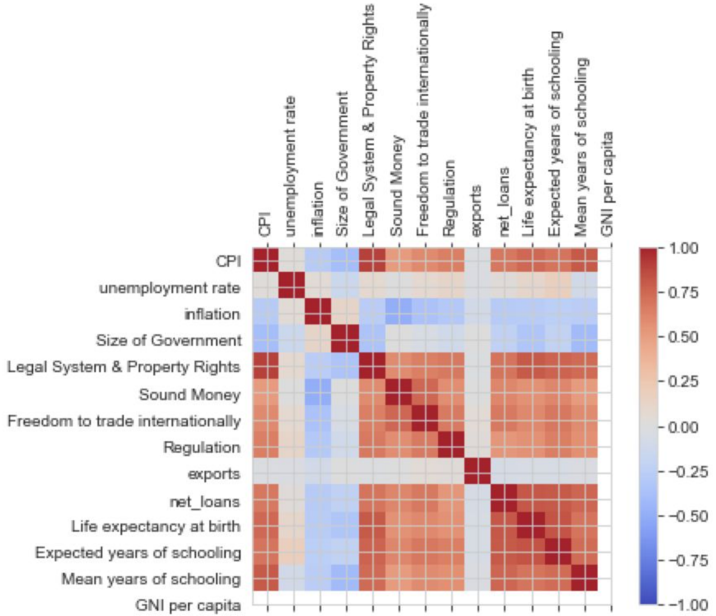
Expected answer to the research questions:

1. Top 5 Significant Features for CPI Prediction:

- GNI per capita, Legal system & property rights, Life expectancy at birth, Expected years of schooling, Size of the government

2. Suggestion to the Government (anti-corruption policy):

- Increase life expectancy and GNI per capita, legal system & property rights, expected years of schooling
- Decrease in the size of the government
- Reduce inflation



Contribution:

- Build a high performance machine learning models to predict a country's CPI with high accuracy
- Predict CPI using almost a decade of data including 196 countries
- Finding and ranking 5 key indicator that affect CPI
- Make suggestions to the government on the indicators that have significant effects on CPI to improve their anti-corruption policies formulation

Source: https://www.transparency.org/en/cpi

Short Literature Review

Sarabia, M., Crecente, F., del Val, M. T., & Giménez, M. (2020). The human development index (HDI) and the corruption perception index (CPI) 2013-2017: Analysis of social conflict and populism in europe: Znanstveno-strucni casopis.*Ekonomska Istrazivanja, 33*(1), 2943-2955. doi:https://doi.org/10.1080/1331677X.2019.1697721

Similarity:

- Using CPI as dependent variable and HDI as predictors in linear regression model

Difference:

- We include more predictors and observations

Lima, M. S., & Delen, D. (2020). Predicting and explaining corruption across countries: A machine learning approach. *Government Information Quarterly*, *37*(1), 101407. https://doi.org/10.1016/j.giq.2019.101407

Similarity:

- Predicting corruption via random forest
- Using cross validation

Difference:

- We also concern the linear relationship between dependent variable and predictors

Domashova, J., & Politova, A. (2021). The corruption perception index: Analysis of dependence on socio-economic indicators. *Procedia Computer Science*, *190*, 193–203. https://doi.org/10.1016/j.procs.2021.06.024

Similarity :

- Predicting corruption via random forest and linear regression
- Using MSE and R^2 to evaluate models

Difference :

- Unlike finding common features by cluster method, we rank the predictors

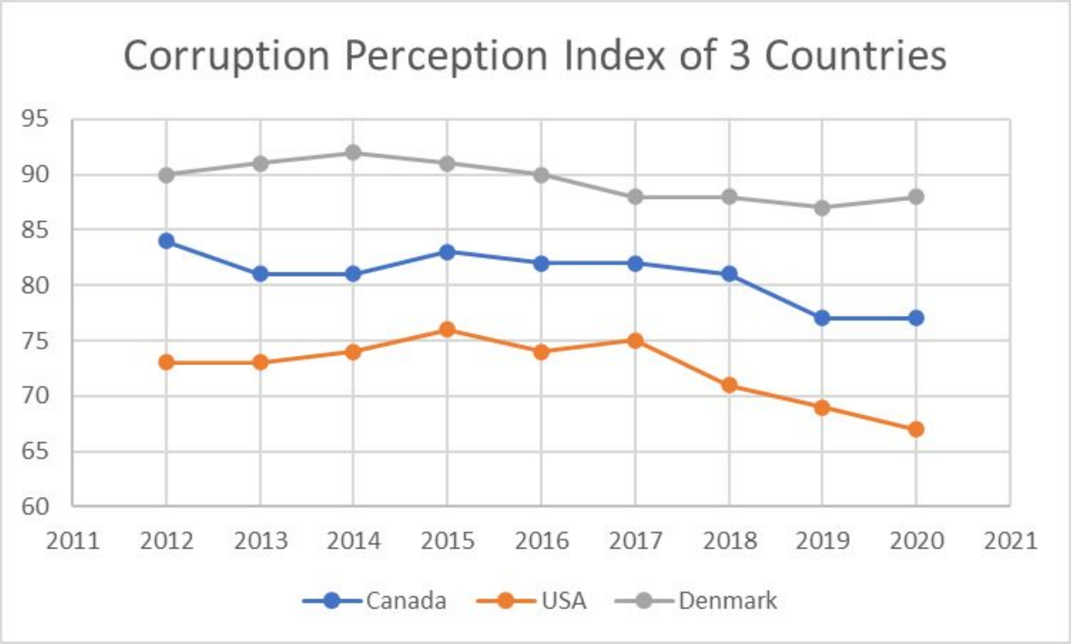
Data Sources

Transparency International: https://www.transparency.org/en/cpi	**Fraser Institute Economic Freedom** https://www.fraserinstitute.org/	**Human Development Reports** https://hdr.undp.org/data-center/documentation-and-downloads	**Our World in Data** https://ourworldindata.org
- Corruption Perception Index (CPI)	- Size of government - Legal Systems & Property Rights - Sound Money - Freedom to Trade Internationally - Regulation	- Life expectancy - Gross National Income per Capita (GNI per capita) - Expected Years of Schooling - Mean Years of Schooling	- Unemployment Rate - Government Primary Net Loans/Borrowing - Export Volume of goods and services

Countries: 196 (2012-2020)

Observations: 1764

Data and Descriptive Statistics



	CPI
Mean:	46
Median:	40
Min:	14
Max:	92
Std:	19

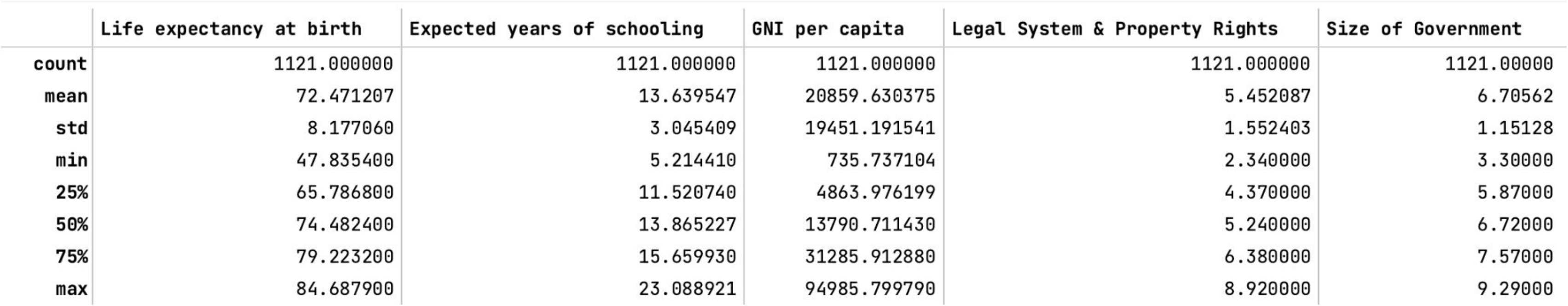
- Denmark is the country with the highest CPI
- Most countries' CPI decline since 2015

Data and Descriptive Statistics

List of 12 features:

(According to the corruption perception index: Analysis of dependence on socio-economic indicators):

- **Legal System & Property Rights**
- **Life expectancy**
- **Gross National Income per capita**
- **Expected years of schooling**
- **Size of government**
- General government primary net loans/borrowing
- Mean years of schooling
- Regulation
- Sound money
- Freedom to trade internationally
- Unemployment rate
- Export Volume of goods and services



	Life expectancy at birth	Expected years of schooling	GNI per capita	Legal System & Property Rights	Size of Government
count	1121.000000	1121.000000	1121.000000	1121.000000	1121.00000
mean	72.471207	13.639547	20859.630375	5.452087	6.70562
std	8.177060	3.045409	19451.191541	1.552403	1.15128
min	47.835400	5.214410	735.737104	2.340000	3.30000
25%	65.786800	11.520740	4863.976199	4.370000	5.87000
50%	74.482400	13.865227	13790.711430	5.240000	6.72000
75%	79.223200	15.659930	31285.912880	6.380000	7.57000
max	84.687900	23.088921	94985.799790	8.920000	9.29000

Method

Data Collection & Cleaning:

- CPI and other 12 features from 2012 to 2020
- Removing missing value

Data Splitting:

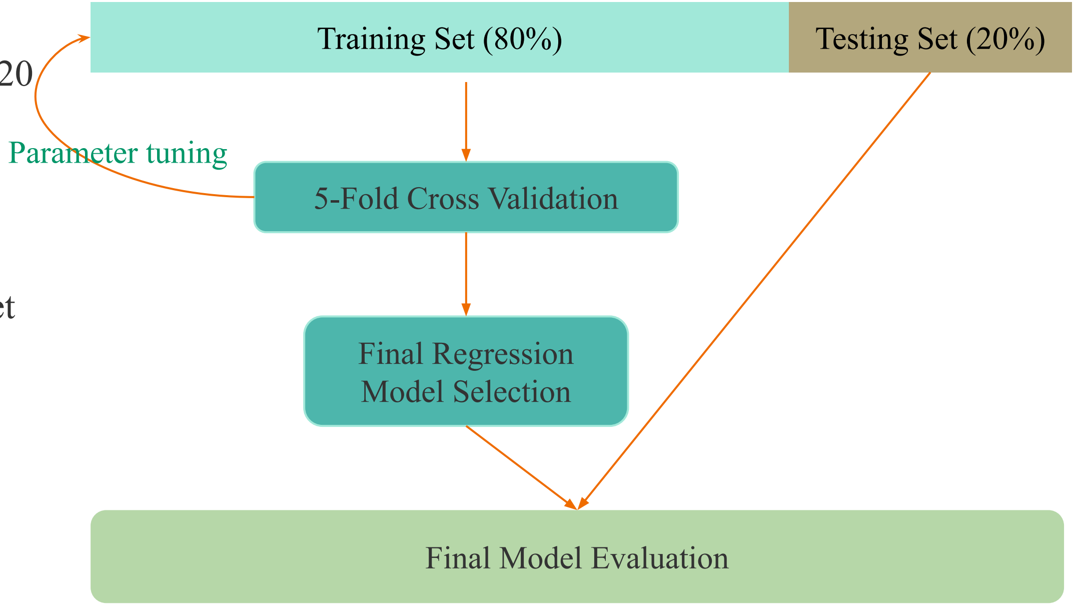
- Training dataset : Testing dataset = 8:2
- 5-Fold Cross Validation on Training dataset

Algorithms:

- Linear Regression with LASSO regularization method
- Decision Tree Regression
- Random Forest Regression

Evaluation:

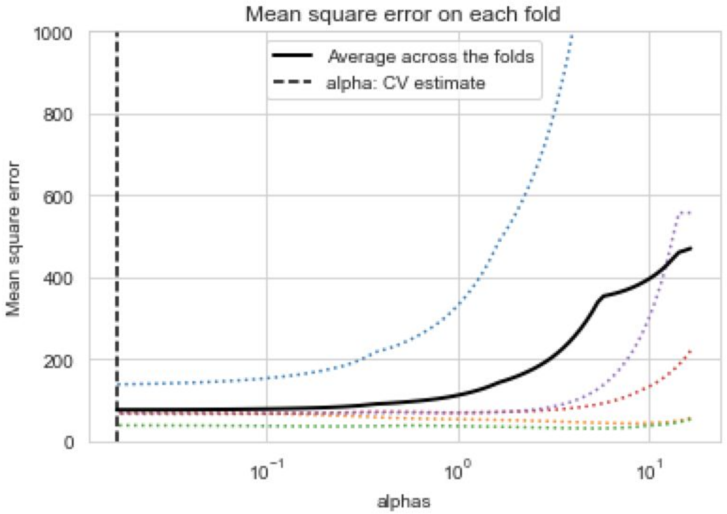
Mean Squared Error (MSE), Coefficient of determination(R^2)



Result - Model Performance (1121 variables)

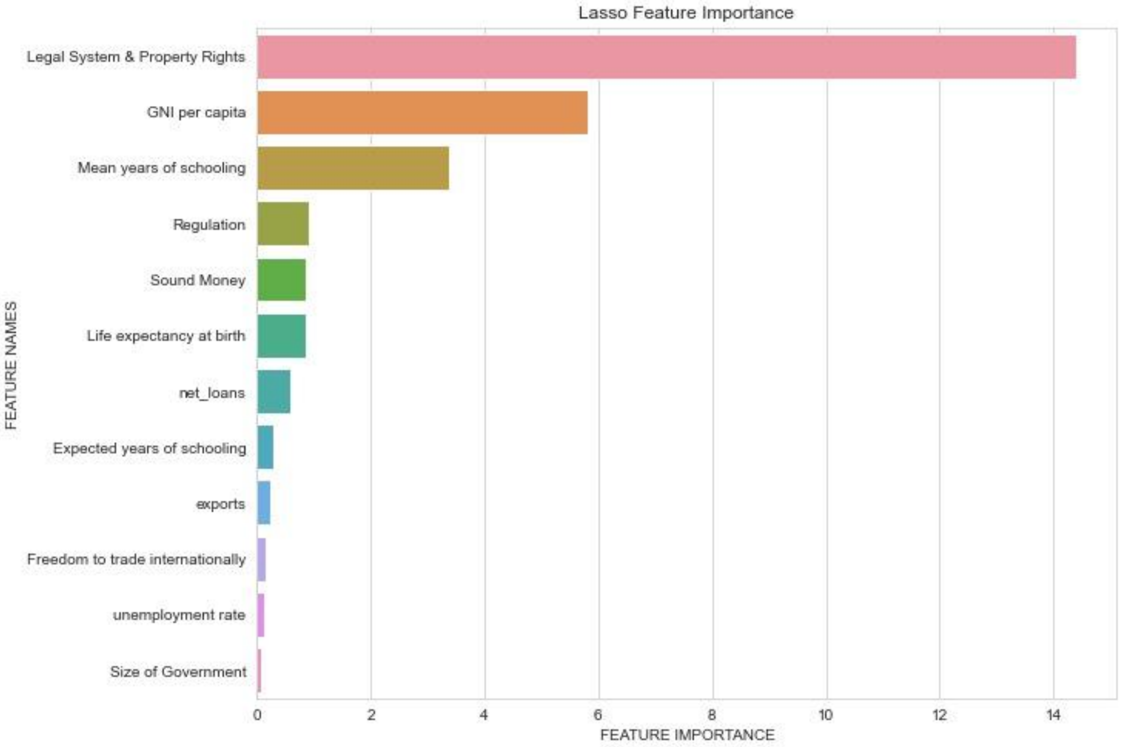
After **5-fold cross validation** on training dataset to tune models:

- Optimal alpha for the Lasso is 0.017
- Optimal ccp_alpha for the decision tree is 0.5
- Optimal parameters for the random forest is max_depth=70, max_features='sqrt', n_estimators=200



Model	MSE_Test	R^2
Lasso Regression	**77.65**	**0.84**
Decision Tree	93.93	0.80
Random Forest	82.46	0.83

Result - Feature Importance



Top 5 Features for Lasso:

- Legal system & property rights
- GNI per capita
- Mean years of schooling
- Regulation
- Sound Money

Conclusion

- **Data from 2012 to 2020:** 1121 variables + 128 countries (after cleaning)
- **Machine learning models:**
  - **Linear Regression with LASSO regularization method (84%)**
  - Decision Tree Regression
  - Random Forest Regression
- **Feature importance:**
  - Legal system & property rights (EFI)
  - GNI per capita (HDI)
  - Mean years of schooling (HDI)
  - Regulation (EFI)
  - Sound Money (EFI)
- **Suggestion to the government on Anti-corruption:**
  - Legal system & property rights, GNI per capita, Sound Money, Regulation, Mean years of schooling ⇧
    - Provide an independent and unbiased judiciary to protect the property rights of owners -> increase Legal system & property rights
    - Improve access to education -> increase mean years of schooling
    - Decrease inflation -> increase sound money

Q&A Session