Exercise 3: Corruption

for Data-oriented Programming Paradigms

Group 36: Bruna Dujmovic

January, 2021

Corruption

Corruption

"Improper and usually unlawful conduct intended to secure a benefit for oneself or another." – *Encyclopaedia Britannica*

- still prominent in many countries around the world, especially as political corruption
- **measurable?** – *Corruption Perceptions Index (CPI)*
  - published yearly by Transparency International
  - measures public sector corruption on a scale of 0 to 100
  - aggregate of other scores collected from a number of different sources
  - based on perceptions of the level of corruption in the public sector by business people and country experts
- **predictable?** – using country characteristics

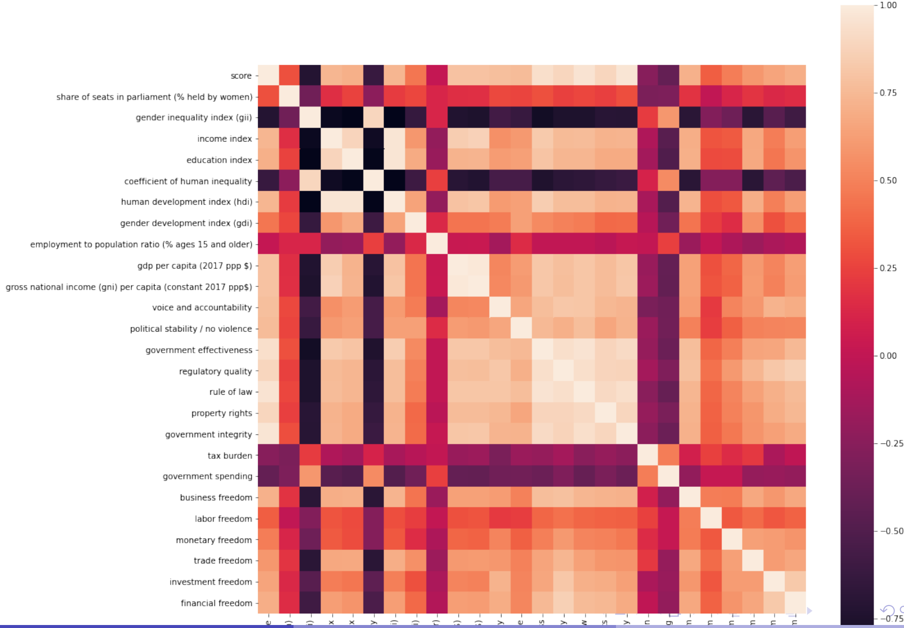
Data set

- 1328 entries and 30 columns (25 predictor variables)
- data for 2012-2019 considered
  - changes to CPI methodology in 2012
- country indicators and indices from multiple sources:
  - Human Development Reports (HDR) indicators
    - e.g. Human Development Index, Inequality index, Education index
  - Index of Economic Freedom (IEF) measures
    - e.g. Government Integrity, Property Rights, Tax Burden
  - Worldwide Governance Indicators (WGI)
    - e.g. Government Effectiveness, Rule of Law, Regulatory Quality

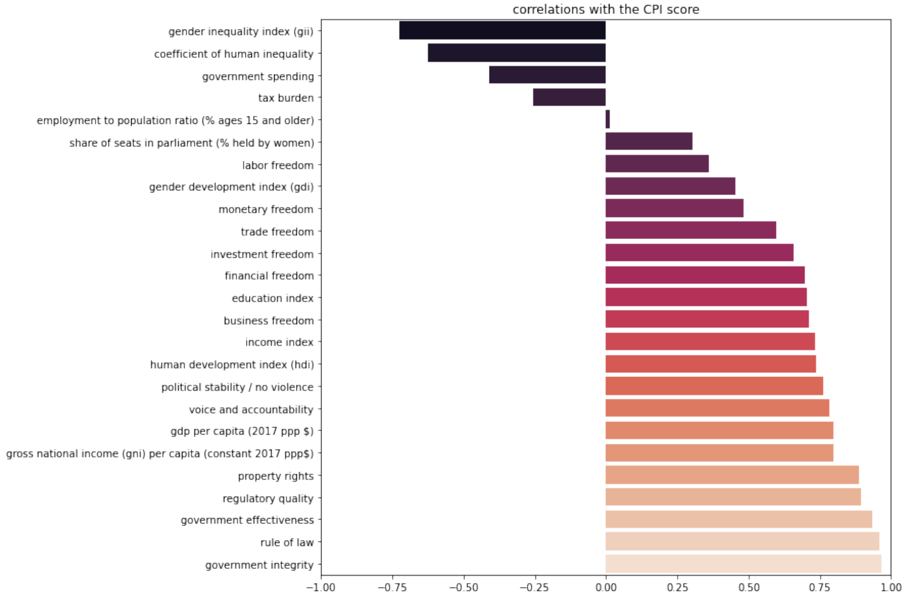
Preprocessing

- countries without score data for each year in the 2012-2019 time-span were removed
  - avoids using imputation on the target variable
  - data for 166 remained
- small amount of missing values
  - **imputation** based on *k*-Nearest Neighbors
  - mean value from 5 nearest neighbors found in the training set
- predictor variables are all numeric – no need for one-hot-encoding or similar transformation
- variables with different value ranges
  - **min-max scaling** applied when needed by the model

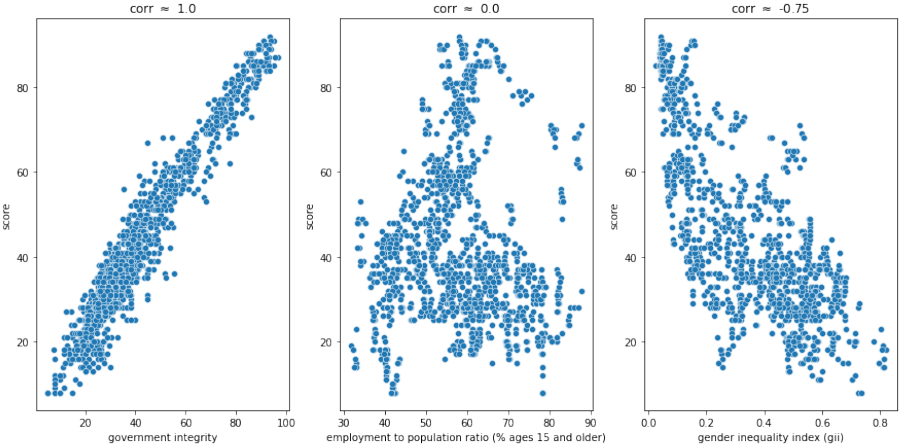
Visualizations – correlation matrix



Visualizations – correlation with target



Visualizations – scatter plots & correlation coefficients



Setup

- 70-30 train/test split
- GridSearchCV
  - 5-fold CV
  - model for each combination of fold and hyperparameters
  - best model taken for further analysis
- standard **performance metrics** for regression
  - R^2, Mean-Squared Error (MSE)
  - visualizing performance with scatter plots
- **baseline** model
  - one-variable linear model
  - government integrity – very high positive correlation with target

Models

- **LogisticRegression**
- **ElasticNet**
  - alpha = [**0.1**, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0],
  - l1_ratio = [0.0, 0.125, 0.25 , 0.375, 0.5, 0.625, 0.75, 0.875, **1.0**]
- **RandomForestRegressor**
  - max_depth = [5, 10, **15**],
  - min_samples_split = [2, 3, **4**],
  - min_samples_leaf = [**2**, 3, 4]
- **SVR**
  - kernel = ['linear', **'rbf'**],
  - C = [0.001, 0.01, 0.1, 1, 10, **100**],
- **MLPRegressor**
  - hidden_layer_sizes = [**(30,)**, (40,), (50,)],
  - activation = ['identity', 'logistic', **'tanh'**, 'relu'],
  - solver = [**'sgd'**, 'adam']

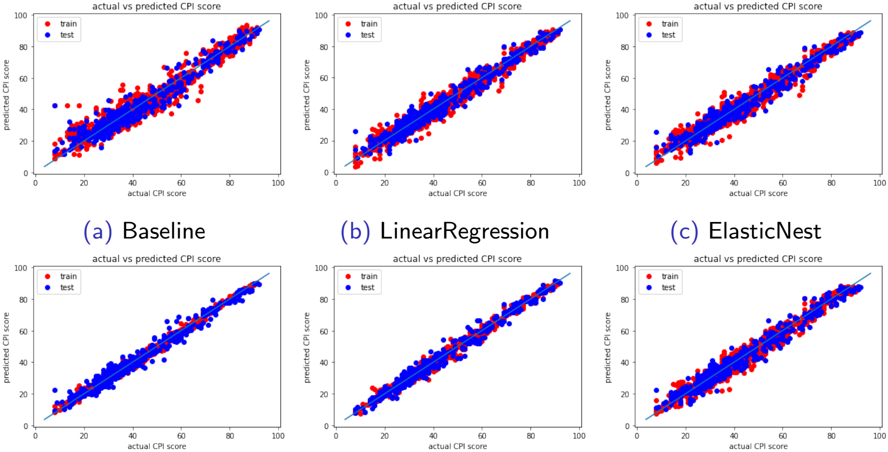
Results – comparison

- best performance by SVR, followed by RandomForestRegressor
- all models surpassed the performance of the baseline model

Test set model performance		
	MSE	R^2
Baseline	26.1042	0.9319
LinearRegression	15.9717	0.9583
ElasticNet	16.4886	0.9570
RandomForestRegressor	9.8694	0.9742
SVR	**7.6442**	**0.9800**
MLPRegressor	14.7483	0.9615

Table: MSE and R^2 test set scores.

Results



(a) Baseline

(b) LinearRegression

(c) ElasticNest

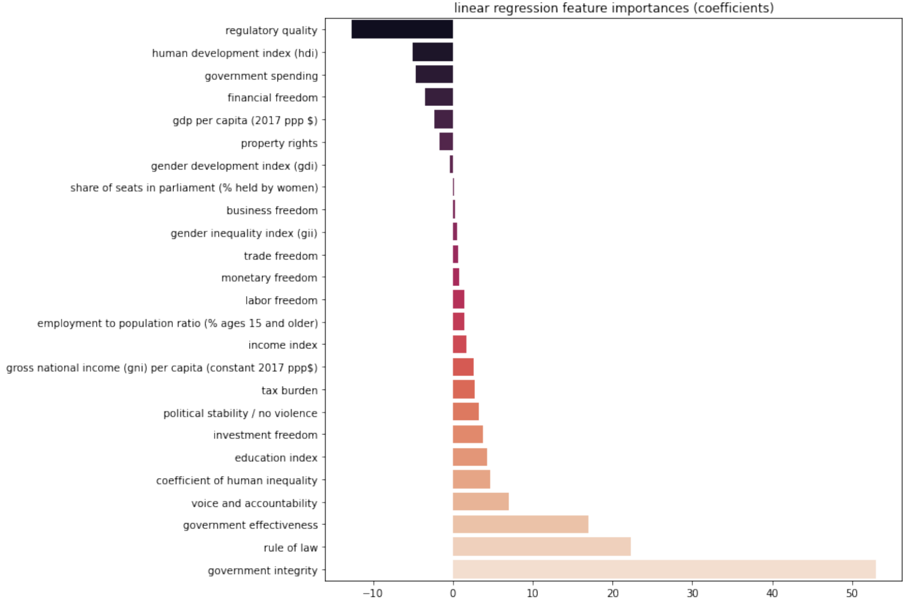
(d) RandomForestRegressor

(e) SVR

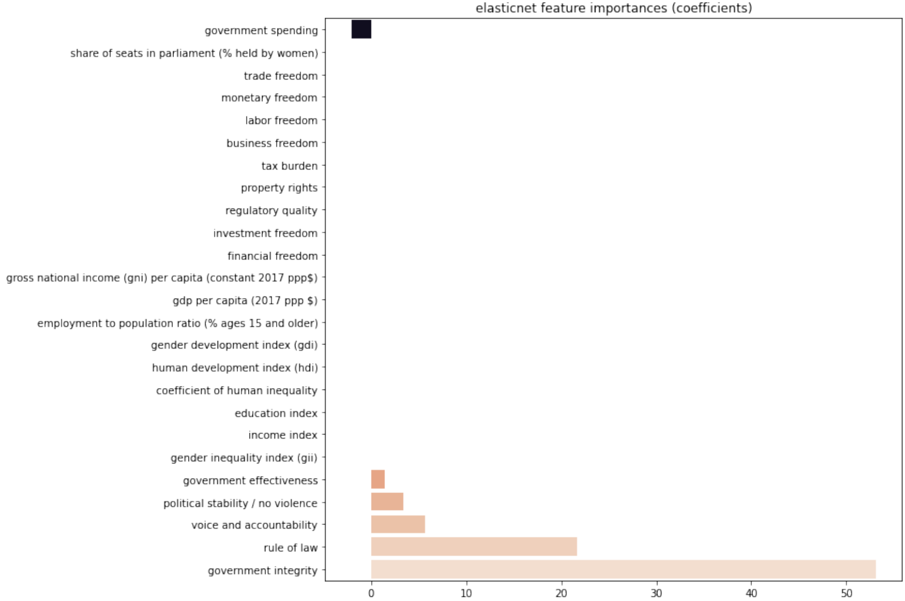
(f) MLPRegressor

Figure: Scatter plots of actual vs. predicted CP scores for training and test sets.

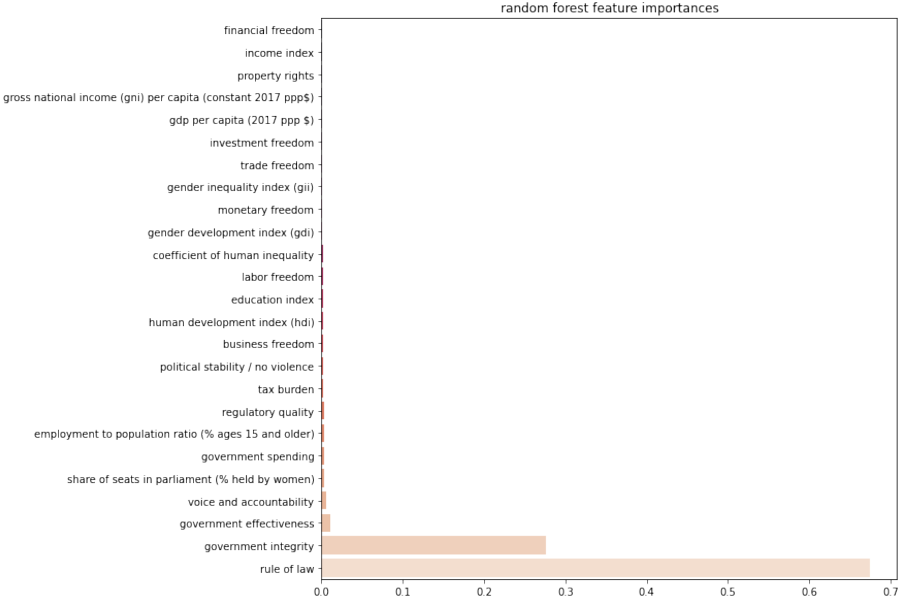
Feature importance – LogisticRegression



Feature importance – ElasticNet



Feature importance – RandomForestRegressor


random forest feature importances
financial freedom
income index
property rights
gross national income (gni) per capita (constant 2017 ppp$)
gdp per capita (2017 ppp $)
investment freedom
trade freedom
gender inequality index (gii)
monetary freedom
gender development index (gdi)
coefficient of human inequality
labor freedom
education index
human development index (hdi)
business freedom
political stability / no violence
tax burden
regulatory quality
employment to population ratio (% ages 15 and older)
government spending
share of seats in parliament (% held by women)
voice and accountability
government effectiveness
government integrity
rule of law