

What characteristics of a european country can predict the level of corruption?

Data Oriented Programming Paradigms 2024W - Exercise 2

Marcus Pilgram, Sergej Bensack, Margharida Maria, Gerhard Karbeutz

Data Collections

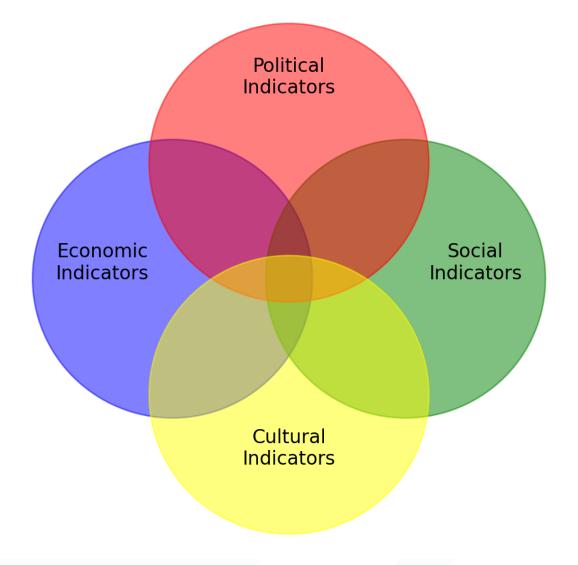


Our Dataset List

- Corruption Perceptions Index (CPI)
- World Happiness Report
- World Bank (WGI, WBI)
- UNDOC Data Corruption and Economic Crime
- Quality of Government OECD

Data Collections

Different Kind of Indicators





Corruption Perceptions Index



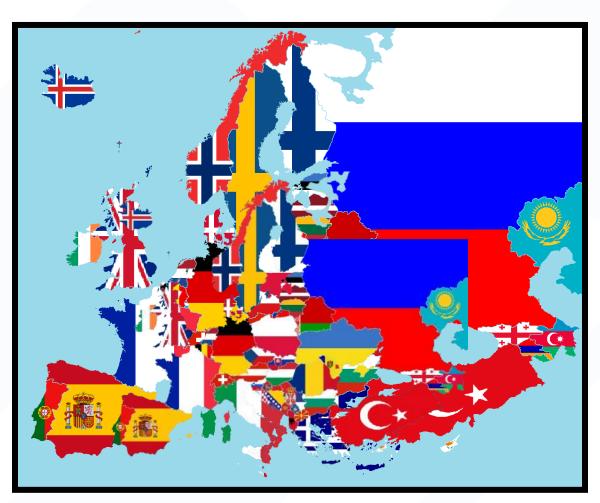
Key Points About the Index

- Measure: Assesses perceived public sector corruption (0 = corrupt, 100 = clean).
- **Sources:** Combines expert and business leader surveys (e.g., World Bank, WEF).
- **Scope:** Covers ~180 countries; highlights bribery, embezzlement, misuse of power.
- **Trends:** Updated yearly by Transparency International for global comparisons.

Data Characteristics

It includes 42 countries and ~ Features





2012

2013

Question change

To be more concrete



What characteristics of a european country can predict the level of perceived corruption (cpi) between 2012 and 2023?

Limitations

Corruption Indices come with limitations



Quantification Challenges:

Hidden and Informal Corruption

Subjectivity of Corruption Indices

Cpi relies heavily on expert opinions

Focus on Public Sector Corruption

Often overlooking private-sector, corporate, or transnational corruption

Data Cleaning and Preprocessing

TUWIEN

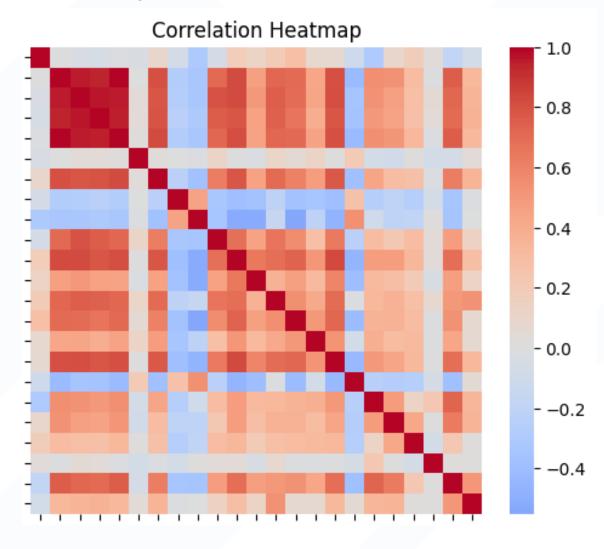
Handle missing values:

Impute missing data with median by country and feature

Exploratory Data Analysis (EDA)

Correlation heatmap shows multicollinearity

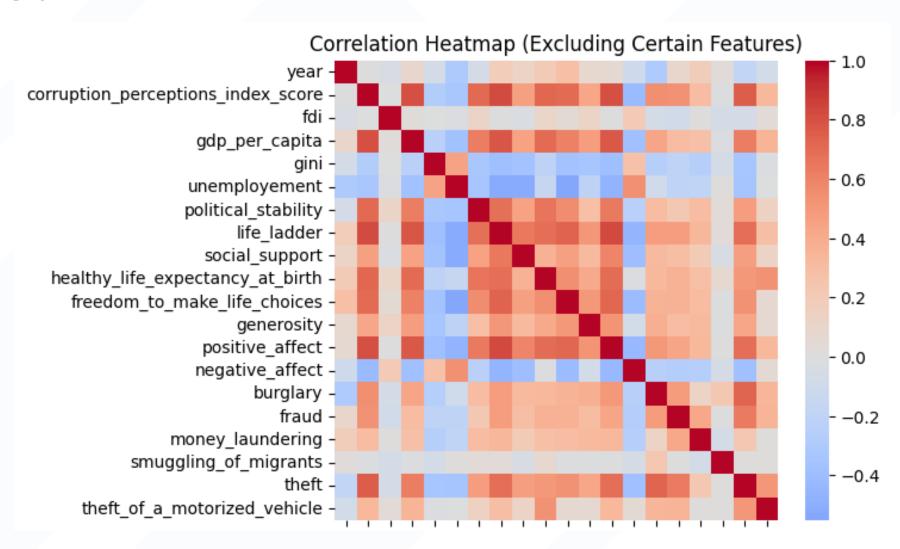




Exploratory Data Analysis (EDA)

TUWIEN

Dropped highly correlated variables



Model Developement

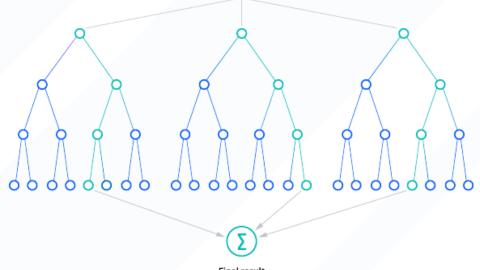
Develop predictive models for corruption levels



Baseline Models: Linear regression



Advanced Model: Random Forest



Final result

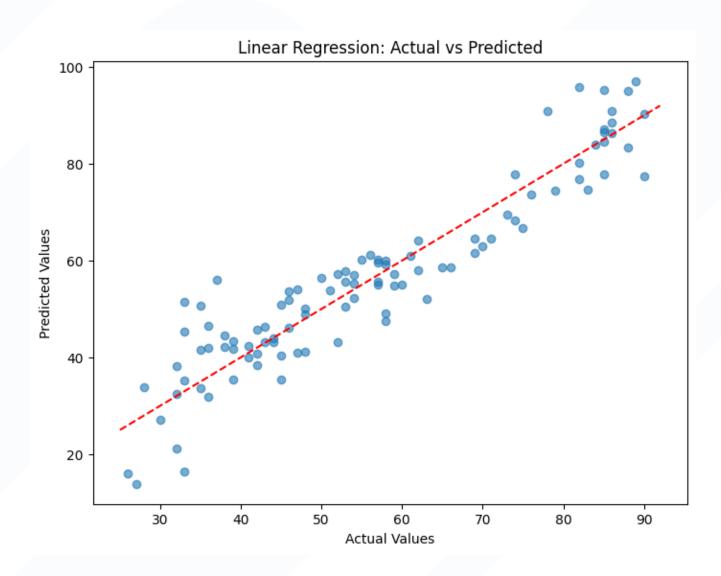


Linear Regression

Metric	Training Set	Testing Set
Mean Squared Error (MSE)	44.890	45.517
Mean Absolute Error (MAE)	5.528	5.301
R ² Score	0.857	0.865

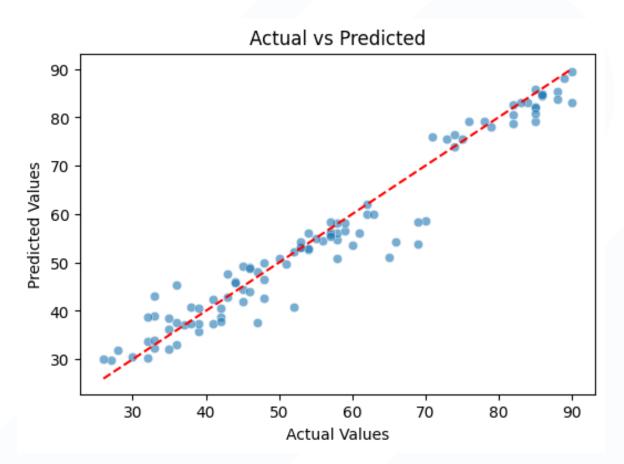


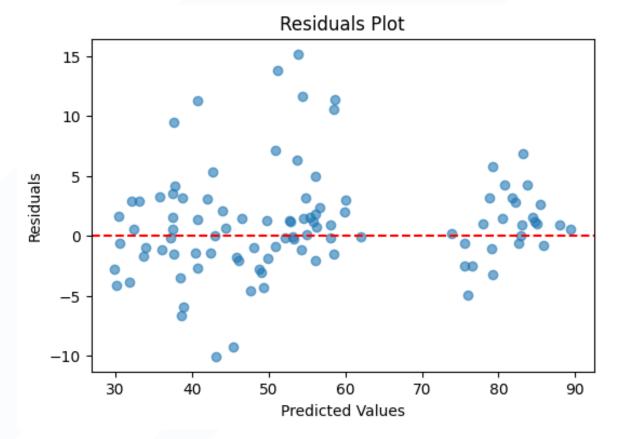
Linear Regression





Random Forrest



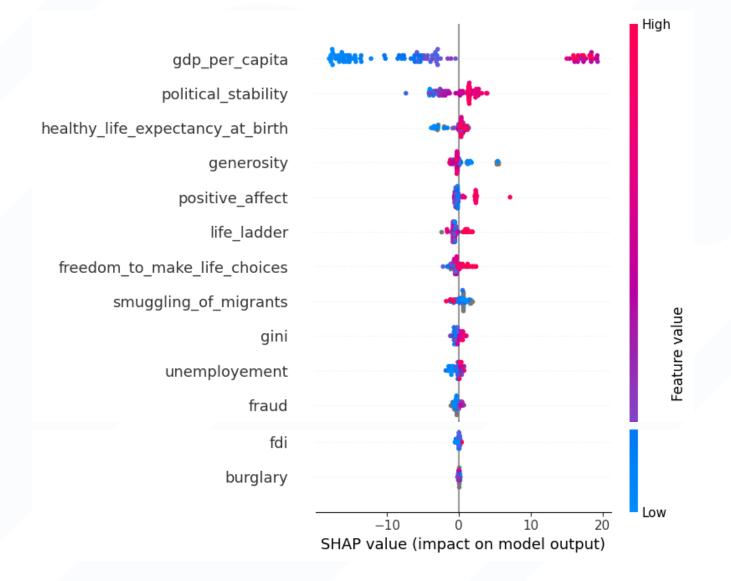


TUWIEN

Random Forrest



SHAP Values





Key Insights



- Well-Being Predicts CPI: Higher life satisfaction, healthy life expectancy, and positive affect correlate with lower perceived corruption.
- Economics Matter: Wealthier nations (higher GDP per capita) and lower unemployment link to higher CPI scores.
- Governance & Stability: Political stability correlates with less perceived corruption.
- Social Freedom & Support: Freedom, social support, and generosity also associate with higher CPI scores

Limitations and Potential Bias



- Correlation ≠ Causation
- Data Gaps & Missing Values
- Different Definitions & Reporting Standards
- Perceived vs. Actual Corruption
- Time Overlap & Aggregation Issues

Difficulties in Analysing the Data



- Data Analysis Challenges
- Multiple Data Sources Integration
- Incomplete Coverage Across Years
- Varying Quality & Consistency
- Risk of Overfitting
- Complex Model Interpretation