Exercise 3: Corruption

Bruna Dujmovic, 12005502

January, 2021

The goal of this project was to predict a country's level of corruption based on other country characteristics. The Corruption Perceptions Index (CPI), which measures corruption in the public sector and is published yearly by Transparency International, was the target variable. Multiple country indicators and indices, with a focus on the ones related to governance and politics, along with more general, well-known measures for education, poverty etc., were used as predictor variables.

1 Data sets

The data set consisted of CPI score data, several Human Development Reports (HDR) indicators (e.g. Human Development Index, Inequality index, Education index, etc.), 12 measures intended to asses economic freedom obtained from the Index of Economic Freedom (IEF) data set (e.g. Government Integrity, Property Rights, Tax Burden, etc.), and the Worldwide Governance Indicators (WGI) data (e.g. Government Effectiveness, Rule of Law, Regulatory Quality, etc.).

A period of 8 years was considered (2012-2019), as older CPI data is no longer comparable due to changes in the methodology. The final data set had 1328 entries and 30 columns, 25 of which were the actual predictor variables (the remaining 5 were country name, ISO 3 country code, region, year, and the score itself).

2 Setup and tools

The project solution consists of a Jupyter Notebook with code implemented in Python 3.6+. Libraries such as Pandas and NumPy were used to access and preprocess the data. Results and data were visualized with the help of Matplotlib and Seaborn. Regression models for CPI score prediction were built using the machine learning library scikit-learn.

3 Data preprocessing

Countries that didn't have CPI score data for each year in the 2012-2019 time-span were removed to avoid using imputation on the target variable. Data for a total of 166 countries remained.

Data for the other variables was usually available. For the small amount of missing values, imputation was performed based on the k-Nearest Neighbors algorithm. Each sample's missing values were imputed using the mean value from 5 nearest neighbors found in the training set.

The 25 predictor variables were all numeric, so no one-hot-encoding or similar transformation had to be applied.

Although the variables themselves were usually defined/computed so as to have values in a certain range (0 to 1, 0 to 100, -2.5 to 2.5), their scales were quite different when compared to one another. Therefore, min-max scaling was applied when needed by the model.

4 Visualizations

Several visualizations were considered: plots of distributions of the target variable per year, a strip plot of target values per region and year, a correlation matrix for the data set, a bar plot for correlations with the target, scatter plots for different variables and the target. Perhaps the most important one is shown in Figure 1, as it points to the existence of several very strong correlations of variables with the CPI score.

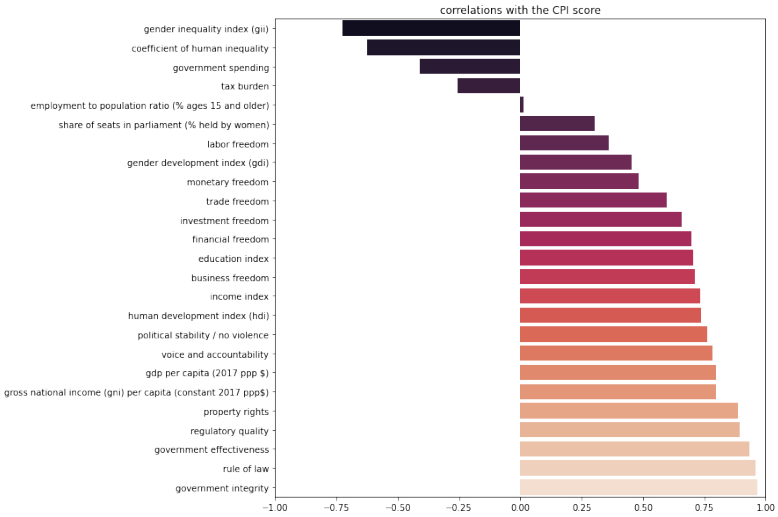

Figure 1: Correlations with the CPI score.

5 Models and procedure

Five regression models of different characteristics were used – logistic regression, ElasticNet, random forest, SVM, and MLP (multi-layer perceptron).

Since each model generally has multiple hyperparameters which impact its performance, different hyperparameter combinations were considered by performing a cross-validation grid-search over a given parameter grid (scikit-learn's `GridSearchCV`).

The data set was first split into training and test sets (70-30 split), after which the training set was split into 5 subsets. `GridSearchCV` trained and tested a model for each combination of the 5 subsets and model hyperparameters, of which the best found model was selected for further analysis.

Model performance was evaluated on the test set using standard performance metrics – R2 and Mean-Squared Error (MSE) – and visualized by plotting the actual and predicted CPI scores for the training and test sets.

A baseline model (one-variable linear model, with Government Integrity as the variable due to its very high positive correlation with the target) was constructed for comparison.

6 Results

The best performance was achieved by the SVM model (MSE around 7, R2 around 0.98), followed closely behind by random forest. The linear regression, ElasticNet, and MLP models had a somewhat worse performance, but still managed to surpass the baseline. Test set MSE and R^2 scores can be seen in Table 1.

An analysis of feature importances for the linear regression, ElasticNet, and random forest models clearly indicated that variables such as Government Integrity and Rule of Law can be used to predict the CPI score. The data set's correlation matrix also indicated that these variables have a very high positive correlation with the score.

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 1: MSE and R^2 test set scores.