

## **Predicting Freedom Project Results**

Aidan Mayhue, Wissal Khlouf, Brian Hockett, Ian Kariuki, Claire Bassett, Rameez Rauf,  
Kalenga Mumba

### **Prediction Question**

How accurately can we predict a country's combined political rights and civil liberties score ("Total") using demographic, economic, and social indicators, and which factors are most influential? This analysis aims to support organizations like Freedom House or human rights NGOs in efficiently monitoring global freedom trends.

### **Model Performance**

#### **Principal Component Analysis**

Using a cumulative explained variance plot, the number of principal components needed to reach 95% of explained variance was found to be 18. Running linear regression on the principal components yielded an  $R^2$  value of 0.547 and root mean squared error of 19.892.

#### **Linear Regression With All Variables**

A linear regression model made using every variable in the data set achieved an  $R^2$  value of 0.668 and a root mean squared error of 17.016 on the test data, indicating moderate predictive capability for the 0-100 scale "Total" score (lower values indicate more restrictions).

#### **Linear Regression With Manually Selected Variables**

After dropping problematic columns containing split proportions (male/female, urban/rural), etc. that were affecting the results, another linear regression model was made. This model had a slightly lower  $R^2$  value of 0.659 and a slightly higher root mean squared error of 17.248.

## Model Comparison

The PCA model performed the worst of the 3, having the lowest  $R^2$  and highest RMSE. Based on this, the 95% explained variance threshold was too low, and a higher number of principal components would've been necessary to outperform the standard model. Of the two ordinary least squares models, the one with every variable performed the best, however the linear dependence of certain variables made the resulting coefficients difficult to interpret. The manually selected linear regression model performed marginally worse on  $R^2$  and RMSE, however the coefficients were much more interpretable, making it the most useful of the 3 models tested.

## Key Results

The following coefficients table comes from the linear regression run on the data set with manually selected variables. The displayed coefficients have the largest impact on the predicted freedom score according to the model.

Feature Category	Feature (Coefficient)	Rationale
Demographic	Percentage of population aged 60+ years old (2.598)	Older populations correlate with higher living standards and increased life expectancy, both of which are common in countries with more freedom
Educational Gender Equity	Gross enrollment ratio - Primary (female) (0.653)	Increases in the ratio of primary school students who are female predicts higher freedom score, as countries with a lot of freedom are more likely to allow women to be educated at a higher rate
Violence Metrics	Percentage of male and female intentional	High violence rates against females is most common in countries with a lower

	homicide victims, female (-2.152)	degree of freedom, where women are not treated equally
Labor Markets	Female unemployment rate (-1.088)	Female unemployment is much higher in countries where women's right to work is restricted, lowering their freedom score

## Interpretation

Based on the regression coefficients above, life expectancy/quality as measured by population ages, and the treatment of women seem to be the most predictive indicators of a country's freedom. Along with the percentage of the population aged 60+, the percentage between 0 and 14 was also a strong positive predictor of high freedom. Our interpretation of this result is that countries with high freedom typically have higher quality of life, which correlates with life expectancy and birth rate. Likewise the prevalence of high-importance variables relating to women's rights/treatment gives the impression that women's rights is one of the most predictive indicators of a country's freedom.

These regression coefficients are very rational, and in many ways expected. This was not the case for every variable, as several variables differ from the expected correlation. There were several instances of morally negative variables positively influencing the freedom of the country, such as sexual assault (+0.087) and homicide rates (+1.276). We speculate that since reporting rates tend to be higher in more affluent/free countries, higher reported sexual assault and homicide rates in "free" countries could be a consequence of higher rates of infractions being reported, rather than occurring. Another possibility we have considered is that the scoring system used by Freedom House fails to take variables of this type into account in their formula for "freedom".

While not every variable's impact on the regression could be understood and explained, the primary predictors of freedom were able to be ascertained. The final model, while not perfect, allows for the freedom of a country to be predicted using its economic and demographic features to a reasonable degree of accuracy, and could serve as an early warning system if it

yields a considerably lower freedom score for a country than that assigned by the Freedom House.

### **Practical Guidelines For Human Rights Organizations**

1. Prioritize monitoring of age and gender demographics as indicators of democratic health
2. Use changes in female unemployment rates and violence rates against females as early warning metrics
3. Combine predictive modeling outputs with qualitative information when tracking and predicting freedom