Can Money Buy Freedom? Analyzing Global Trends in Prosperity and Political Rights

Aidan Mayhue, Wissal Khlouf, Brian Hockett, Ian Kariuki, Claire Bassett,

Rameez Rauf, Kalenga Mumba

School of Data Science, University of Virginia

DS 3021: Introduction to Machine Learning

Dr. Terence Johnson

May 9th, 2025

Abstract:

With the proliferation of globalization, concerns over human rights have become increasingly relevant as advances in technology allow for increased awareness of potential issues. Concerns over human rights violations have become a central problem in international discourse. This growing awareness underscores the need to utilize quantitative data-driven initiatives to track, analyze, and predict freedom status across different nations and their respective scores. The dataset mostly focuses on quantifying the degree of freedom in each country through metrics such as the number of homicides, the unemployment rate, GDP Per Capita, access to education, etc. For EDA, the data contained many missing values, all of which were imputed. A Principal Component Analysis (PCA) was performed on the data to find which variables most strongly influenced a country's freedom score. Based on the findings from PCA, the model was evaluated based on performance metrics such as R-squared and Root Mean Squared Error (RMSE). This model was also intended to determine how accurately a country's combined political rights and civil liberties score could be predicted using these factors. The main goal of this project is to assess the extent to which a country's freedom score could be predicted using a combination of quantitative and qualitative measures. The findings provide insights that may inform future research and policymaking on the relationship between economic conditions and democratic freedoms.

Introduction:

Many value freedom as a fundamental human right, yet there remains difficulty in defining freedom, and even more so measuring it objectively. Although political institutions and legal systems lay a foundation for how freedom is experienced, social and economic factors such as access to education, gender equality, and employment rate hold crucial in framing the quality in which these freedoms are realized in day to day life. With this in mind, organizations such as Freedom House and various human rights NGOs monitor global freedom trends to predict human rights violations and declining living standards. The aim of this research is to examine the methodology used by the Freedom House that were used to assign freedom scores and to analyze how underlying economic as well as social variables contribute to those ratings. Approaching the problem through predictive modeling allows freedom not to be seen merely as a political concept, but something that can be understood through real-world conditions.

The research found that certain indicators–such as life expectancy and the treatment of women given criteria of education and labor–were strongest among indicators of a country's freedom. Although some of the findings were as expected, such as increased education or better health outcomes correlating strongly with more freedom, other relationships were less intuitive. For example, the data showed that high crime rates, particularly gender-based violence, appeared to have a greater frequency of reporting in freer societies. This raised consideration whether certain "negative" indicators are actually reflections of more freedom attributed to better reporting infrastructure rather than worse social conditions. The research also showed the value of analyzing certain variables that may be overlooked in traditional freedom scoring systems. For example, gender-based indicators like female unemployment or female homicide rates concluded to be strong predictors of a country's overall freedom score. This gave valuable insight into how the treatment of women is not just a side issue, but an important part of what defines freedom for a nation.

A recognition of these patterns allows for more data-informed policy responses. As a result, several practical recommendations for human rights organizations to consider include:

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

The findings of the research can be applied as an early warning system, particularly in the case of it yielding a considerably lower freedom score for a country than that assigned by Freedom House, allowing for the support of more timely interventions. Additionally, the model presents more transparency as a foundation for better understanding what it is that motivates and threatens freedom across nations.

Data:

The dataset includes 1,150 rows and 51 columns, with each row representing a country in a given year. It combines a wide range of indicators that cover freedom, economic conditions, education, labor markets, crime, infrastructure, and demographics. Some of the key variables include a country's Freedom House status (Free, Partly Free, Not Free), its political rights and civil liberties scores, and a combined "Total Freedom Score." Other columns include things like GDP per capita, unemployment rates, enrollment in primary and secondary education, age distributions, internet usage, and crime rates such as homicides and assaults. This data is useful for trying to predict a country's level of freedom based on economic and social factors. For example, we can look at whether countries with higher education levels or better access to clean water also tend to be freer. The dataset includes both direct measures of freedom and a wide set of variables that might help explain why some countries are more free than others. We can also compare trends across regions or over time using the “Edition” column, which indicates the year of the observation.

At the same time, there are some challenges with using this data. Some variables have missing values, especially in the crime and economic categories. These had to be cleaned or imputed, which could affect our results. Some numeric columns were stored as text with commas, which we had to convert. There are also cases where variables are split unnecessarily, like male and female unemployment or urban and rural water access. We decided to drop some of these columns to avoid redundancy. Another issue is that the Freedom House scores are based on expert assessments, which could have bias or reflect regional politics. Finally, many variables in the dataset are highly correlated, which could cause problems in regression models if not handled carefully. While the data is far from perfect, it offers a strong starting point for studying the factors that relate to political and civil freedom.

Methods:

This project explores whether economic prosperity can be used to predict the level of political and civil freedom within countries. The goal is to understand if improvements in economic and demographic indicators are associated with increased civil liberties and political rights. This question is especially important for international development, where economic progress is often seen as a pathway to greater societal freedoms. The dataset used in this analysis contains observations at the country-year level. Each row represents a specific country in a given year, capturing a variety of economic, social, and demographic factors along with a composite score reflecting that country’s level of freedom. This freedom score combines political rights and civil liberties into a single numeric measure ranging from zero to one hundred, with lower values representing more restrictions.

To address this as a supervised learning task, we used regression analysis to predict the total freedom score. Although it would be possible to treat this as a classification problem by dividing countries into “Free,” “Partly Free,” and “Not Free,” a regression framework provides more nuanced, continuous predictions that capture variation within each group. Before applying any models, the dataset was cleaned and prepared. Many variables contained missing values. To address this, we imputed missing data using each country’s average value for that variable across all available years. If a country had no data for a particular variable, we substituted the average value for the broader world region. If both were unavailable, we removed the corresponding row from the dataset. After cleaning, all numeric features were standardized to prepare for dimensionality reduction.

We used principal component analysis (PCA) to reduce the number of predictors and mitigate multicollinearity. Many of the economic and demographic indicators are likely correlated, which can distort regression estimates. PCA transformed the original variables into a smaller set of uncorrelated components while preserving most of the variation in the data. We selected the number of principal components needed to capture at least 95 percent of the total variance and used these as inputs in a linear regression model. To evaluate the usefulness of PCA, we also ran a standard multiple linear regression model using the original features without dimensionality reduction. Both models were compared using R-squared and root mean squared error (RMSE), which indicate how well the model fits the data and how close the predictions are to the actual freedom scores. We considered a model successful if it produced an R-squared value above 0.5 and an RMSE between 10 and 18. These benchmarks reflect a meaningful ability to explain variation in freedom scores while maintaining reasonable prediction accuracy.

There were several anticipated challenges. Selecting the right number of principal components required balancing the trade-off between simplicity and information loss. If too few components were used, the model would overlook important variation in the data. If too many were included, the benefits of dimensionality reduction would diminish. Additionally, interpreting some model coefficients was not straightforward. For example, variables like sexual assault or homicide might be reported more frequently in free societies with better data infrastructure, leading to unexpected positive correlations with freedom. To explore this, we compared the PCA model with one using the original variables and considered contextual explanations for any surprising results.

Our findings will be presented through a table of regression coefficients and performance metrics for each model. These materials provide a clear comparison of the models and offer insight into which factors most strongly influence a country's level of freedom.

Results:

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.

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).

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.

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.

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 homicide victims, female (-2.152)	High violence rates against females is most common in countries with a lower 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 occuring. 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”.

Conclusion:

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.

Many of the shortcomings of the analysis are rooted in the incompleteness of the dataset. Future work should focus on improving data quality/quantity (reducing the number of missing values) and exploring additional variables to enhance the validity of these predictions. Due to the complex nature of the target variable, “Freedom Score”, omitted variable bias likely contributed to the inexplicable values found for certain regression coefficients. Adding more variables to the dataset would increase the interpretability of the regression coefficients, as well as improve the predictive power of the regression model.

Likewise, making use of a more robust feature selection process, such as LASSO, would be a worthwhile endeavor for future work. For this paper, Principal Component Analysis was the only algorithmic feature selection process used, and it yielded worse results than manual feature selection and simply including every variable in the dataset. Better feature selection would improve the performance metrics of the model and the interpretability of the regression coefficients.

Despite the issues, the data tells an important story that freedom scores are highly correlated with an older population, and treatment of female citizens. An aging population reflects a high quality of life and access to medical care. These findings suggest that monitoring demographic trends and gender equity can provide early warning signs of shifts in a country's freedom. Human rights organizations should prioritize tracking changes in female unemployment and violence rates, supplementing models with qualitative insights for a more comprehensive view. While the model is not without limitations, it demonstrates that economic and demographic factors offer meaningful predictive power for understanding global freedom. Future work should focus on improving data quality and exploring additional variables to enhance the validity of these predictions.

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

References:

Freedom House. (2023). *Freedom in the World*. Freedom House. https://freedomhouse.org/report/freedom-world#Data

United Nations. (2019). *UNdata*. Un.org. https://data.un.org/