Data science final project

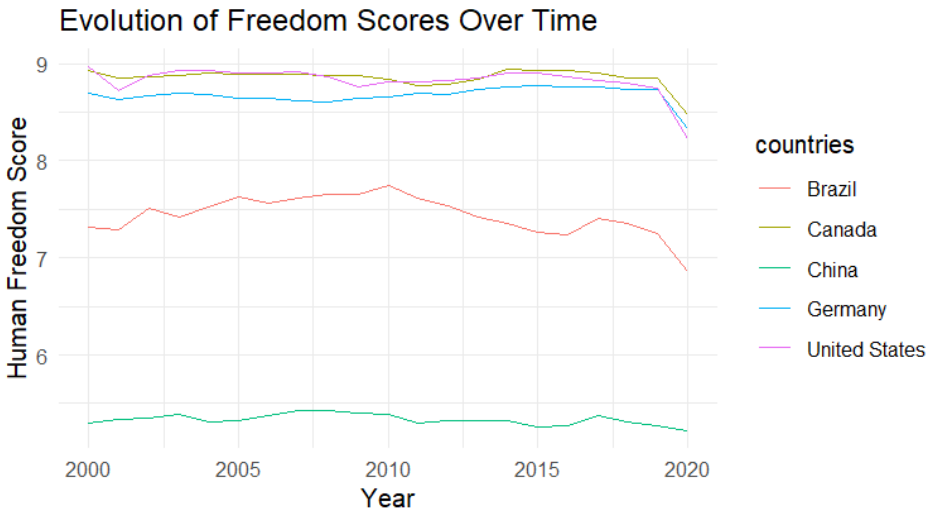
INTRODUCTION

The Human Freedom Index (HFI) is a crucial tool for analysing the differences in individual and collective freedom around the world. The HFI makes it possible to understand these differences through a multitude of indicators ranging from indices relating to freedom of the press, freedom of political bodies or even personal economic or social aspects.

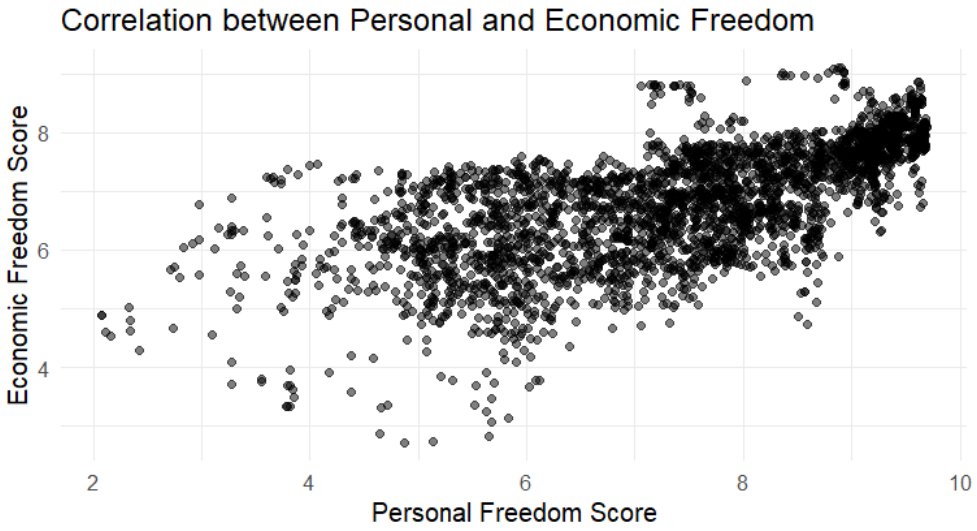
Our project will therefore be based on the most recent edition of the HFI, which covers 165 countries and has 83 distinct indicators from various areas, such as security, freedom of movement, religion, or the size of government.

In this study, I wanted to focus on the impact of religious, security, social, political and economic indices on the index of freedom. By focusing on these specific areas, we aim to uncover the key factors that most significantly influence freedom in different countries.

The study of the determinants of the index of freedom is all the more interesting because it seems rather stable for the countries between the different years, with the exception of 2020 where the scores fell because of the COVID-19 crisis. Thus, understanding what are the determinants of the index of freedom could allow us to understand the orientation of certain policies of governments for example.



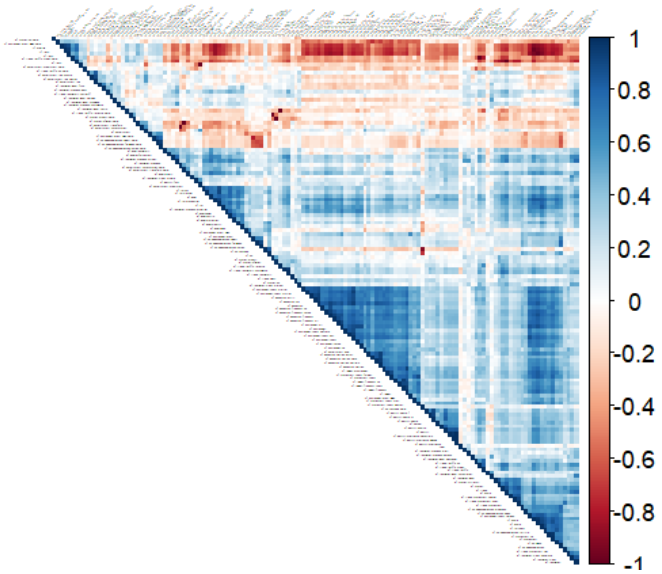
Moreover, there is a strong correlation between the index of human freedom and the index of economic freedom, so it is interesting to know if it is indeed the economic aspect that has the most importance on the determination of the index of freedom.



DATA

In this project, we explore the Human Freedom Index data file produced by Human Freedom Index, which represents a set of indicators on human freedom. The hfi_cc_2020 dataset consists of several columns, each representing a different aspect of freedom or socio-economic indicator. The data are classified by region, country and year (between 2000 and 2020).

A first analysis of the data reveals that the first challenge of data processing is related to the quality of the data, in particular because of the presence of many missing values in several columns. Indeed, in the original database we have 141 columns and 3465 entries, with 44990 missing values. After analyzing the raw data set, certain trends can be observed, in particular, we can see that there are many strong correlations between the different variables in the dataset:



Data cleaning

To clean the data, I performed the following steps:

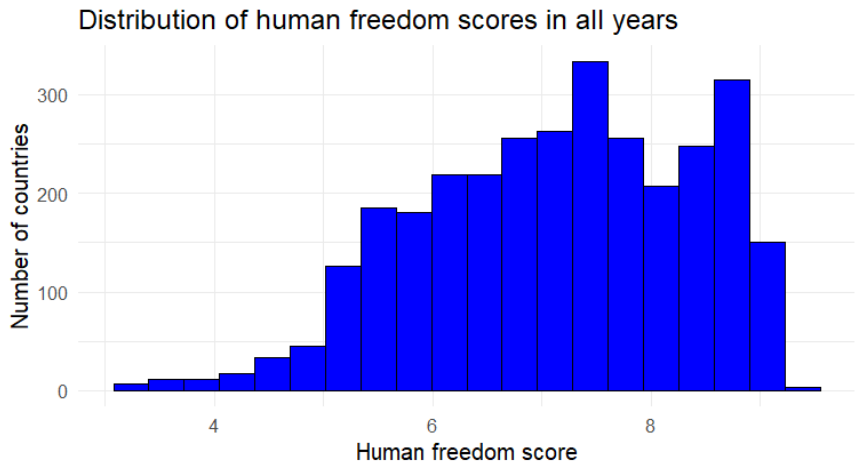
Since we want to focus only on certain values and given the large number of variables present in this dataset, we keep only the following columns (while taking care to select columns that contain few missing values):

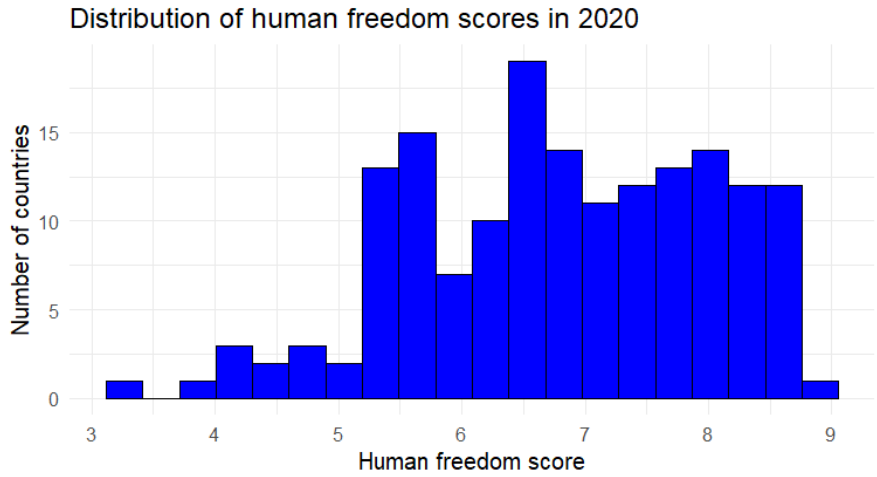
- The index of human freedom (*hf_score* column).
- The religious freedom index to represent the religious dimension (*pf_religion_freedom* column).
- The homicide protection index, for the security dimension (represented in the *pf_ss_homicide* column).
- The index of gender equality with regard to the possibility of divorce, for the social dimension (*pf_identity_divorce* column).
- The freedom of political bodies index, for the political dimension (*pf_assembly* column).
- The inflation index, for the economic dimension (*ef_money_inflation* column).

I have also replaced the few missing data with the country average for previous years on this specific variable. By manually checking the data, I encountered a problem with the divorce index for Suriname, which has no data for this index (in all years), so I decided to take the average of all the other countries over all the years in my dataset. This method was chosen because it preserves the general structure of the data while providing a reasonable estimate for missing values.

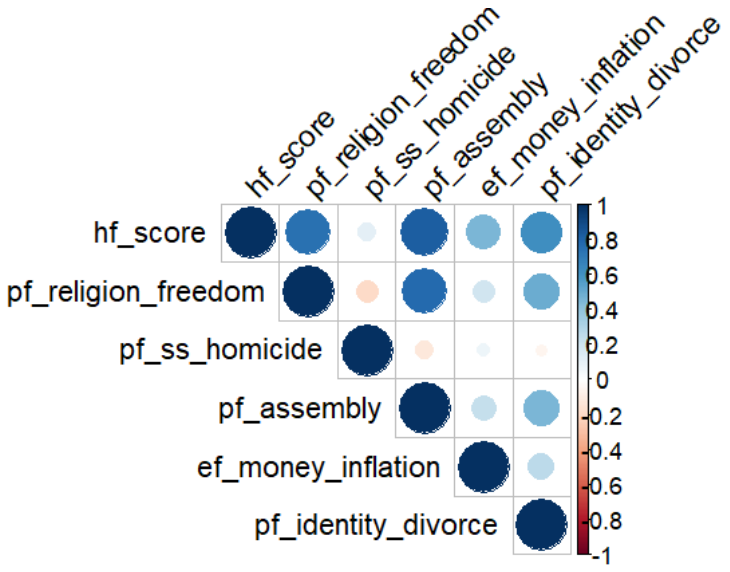
I then decided to keep only the most recent year in order to have a dataset with only one value per variable for each country. With those steps and with the data type, I didn't have to do Feature Engineering.

At the end of this data cleaning process, my cleaned dataset contains 7 variables (including one representing the countries but which will therefore only be useful to be able to correct potential future errors more quickly) with 165 entries (the number of countries). After cleaning up our data set, we can observe differences with the original game, especially with regard to the average index of human freedom, which is not surprising since the latter has fallen over the year 2020, because of the COVID crisis.





In the end, we have this correlation matrix, with therefore a strong correlation between the index of human freedom and the preserved indices of religious and political freedom.



Methodology and Results

Methodology

The main objective of this project is therefore to understand which factors most influence the human freedom score (hf_score).

Our methodology consists of several steps:

1. Model development:

Before the construction of the models, I did an analysis of the data to understand the trends, the possible anomalies and the relationships between the variables. This step confirmed the relevance of the variables that we have selected for the step of modeling.

We applied various statistical models, including GLM, LASSO, Ridge Regression, kNN, Random Forest, and Boosting. Each model has been adjusted to ensure its relevance. This diverse approach allowed us to mitigate the weaknesses inherent in any single model.

I also took care to transform the data as soon as necessary for each model so as not to encounter any problems, in particular to ensure that the latter were at the right scale and were of the right type.

For training and model validation, I divided the dataset into training and test sets, using 70% of the data for training and the remaining 30% for testing.

2. Model Evaluation:

Models were evaluated using standard measurements such as Mean Square Error (RMSE), R-Squared (R^2), and Error Rate. RMSE provided insights into the average model prediction error, R^2 indicated the proportion of variance explained by the model, and the Error Rate offered a straightforward measure of prediction accuracy.

3. Feature Significance Analysis:

For Random Forest and Boosting models, variable importance analysis was conducted. This step was crucial for identifying which predictors most significantly affect the hf_score.

Results

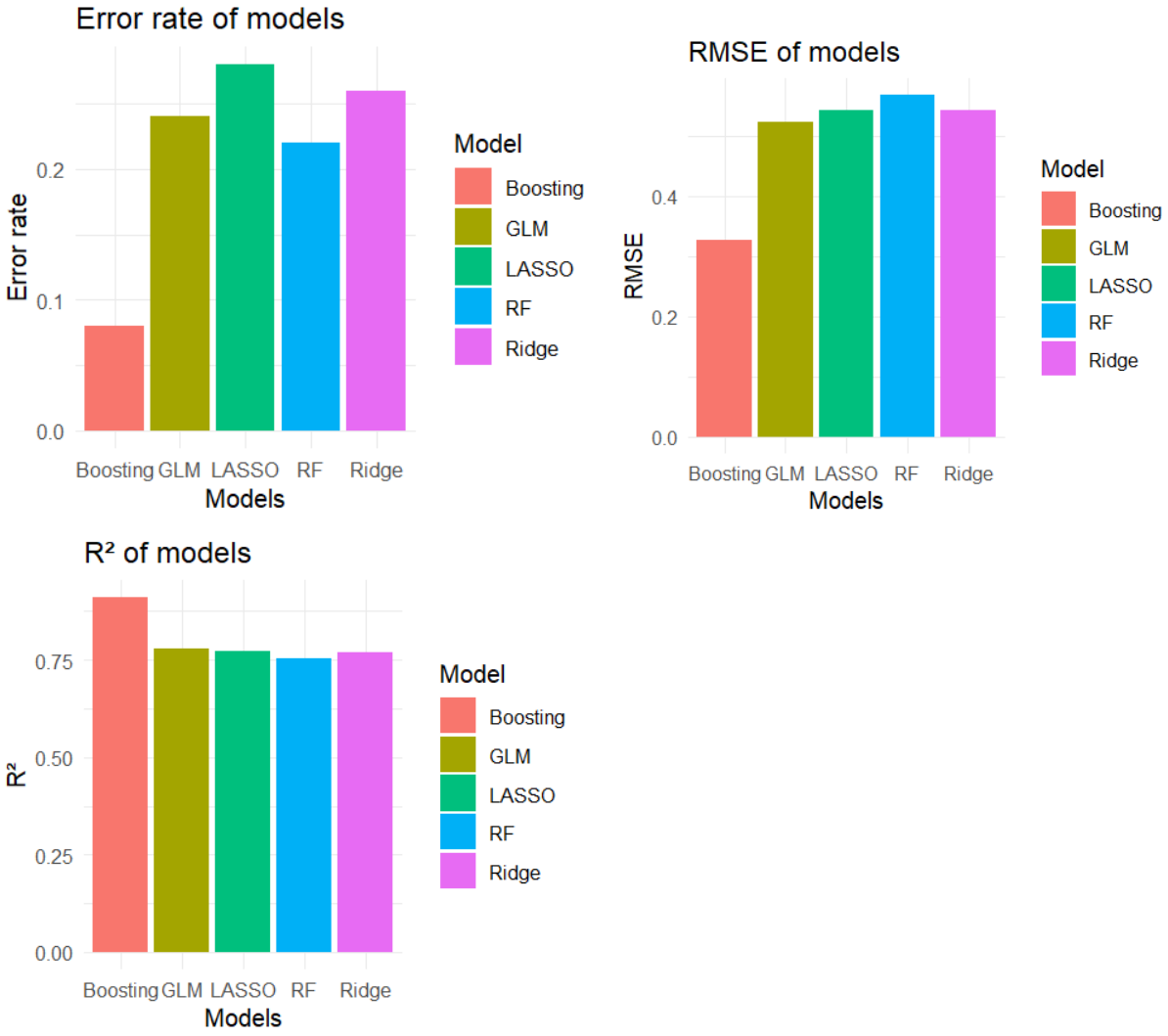
The results of our modelling efforts provided interesting results:

1. Model Performance:

Key influencing factors:

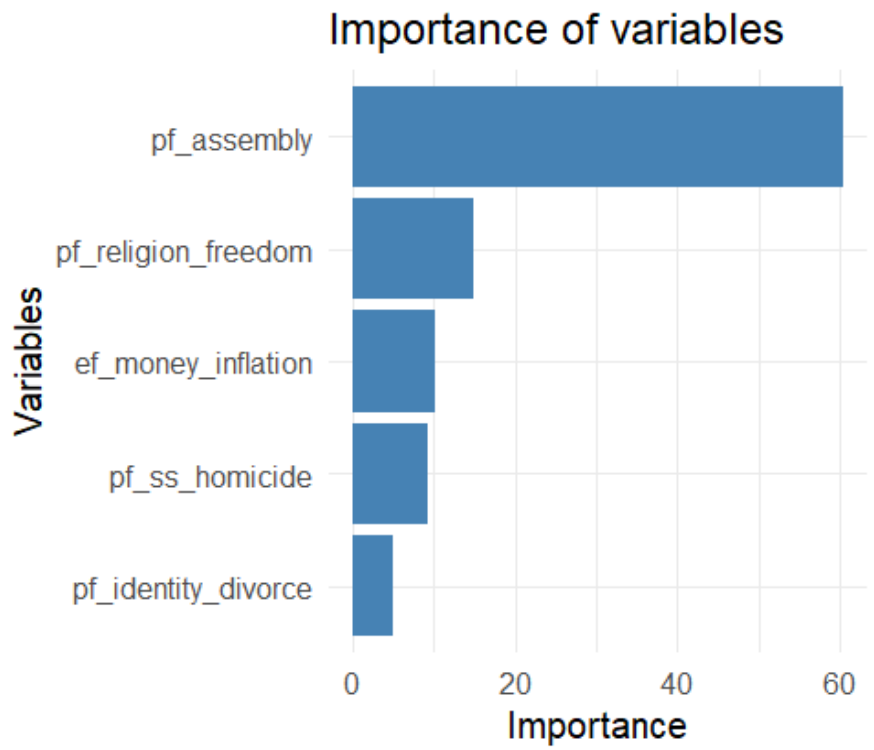
The Boosting model demonstrated the highest level of accuracy, with an RMSE of 0.3278, an R^2 of 0.9104 and an error rate of 0.08, indicating its effectiveness in predicting the index of human freedom.

The other models showed good performance, with reasonable RMSE and R^2 values. However, the kNN model underperformed, probably due to the scale of the features.



2. Key influencing factors:

The analysis of the importance of variables in the Random Forest and Boosting models revealed that pf_assembly was the most influential variable, followed by pf_religion_freedom and ef_money_inflation (as we can see on the following graphs, representing the importance of the variables in Boosting). This conclusion confirms the correlation we saw earlier in the correlation matrix.



In conclusion, our results highlight the significant influence of religious, security and political factors on the index of freedom. Despite this, while our models have demonstrated the predictive power of some indicators, the complexity of human freedom requires ongoing exploration and understanding.