Missing No More: Estimating Predictors of Missingness using Algorithmic Modelling & Multiverse Analysis

Albert Cort Banke (7590423)

Thesis Project

Utrecht University

MSc. Applied Data Science

Supervisor: Kyle M. Lang

Date of submission: June 29th, 2023

Referencing system: APA (sixth edition)

Second examiner: Maarten Cruyff

Number of pages: 31

Number of characters: 69.592

Interactive result dashboard: https://albertbanke-thesis-missing-data-2023-app-ou02mg.streamlit.app/

Code and data: https://github.com/albertbanke/thesis_missing_data_2023

Abstract

Missing data often presents challenges for researchers and professionals. The condition of Missing at Random (MAR) is often based on informed assumptions due to the absence of a comprehensive registry of predictors of nonresponse. This paper investigates missingness in the Human Freedom Index using multiverse analysis and algorithmic modeling with Random Forest, XGBoost, LightGBM, and Neural Networks. The findings highlight the best overall performance for LightGBM and XGBoost, with high Macro F1 scores and Matthew's Correlation Coefficient scores. The most important predictors for these models are year, pf_movement, ef_gender, and the spatial x- and y coordinates, highlighting geographical, societal, and temporal influences on missingness. The study underscores the significance of understanding missingness mechanisms in global datasets and encourages similar research in other contexts.

Table of Contents

1. Introduction

Missing data is often a reason for headaches for researchers and professionals when it comes to analysis and machine learning (Baraldi & Enders, 2010). Various methods exist to handle missing data, yet the decision to assume a Missing at Random (MAR) mechanism is largely based on educated guesswork (Little, Jorgensen, Lang, & Moore, 2014). This is partly due to the lack of a universal registry that identifies variables that tend to correlate with the propensity of a response being missing. This means the missingness is often assumed to depend only on observed data, a condition that may not always hold in real-world scenarios (van Ginkel, Linting, & Rippe, 2020).

To help remedy this conceptual headache, the project aims at identifying predictors of missingness in the Human Freedom Index (HFI). The HFI is a spatial-temporal public data set from 2020 with 141 attributes, of which 112 have missing data (Fraser Institute, Cato Institute, 2023). In this scenario, predictors refer to features that exhibit strong influence, as determined by feature importance measures in XGBoost, Random Forest, LightGBM, and Neural Network coefficient analysis, on the binary classification of missing data. Through algorithmic modeling and a multiverse analysis, the most important and consistent predictors of missingness will be identified. These are predictors that classify missingness for the best performing models and that hold consistency across three varied data frame samples with increased degrees of preprocessing and feature engineering. This leads to the following two research questions: 1) Which models yield the best performance for classifying missingness in the Human Freedom Index? and 2) For these top-performing models, which predictors contribute the most to feature importance for classifications of missingness in the Human Freedom Index data*?*

Four key variables, each representing a different common pattern of missingness in the HFI data, are jointly analyzed through a multiverse analysis. This involves applying multiple data processing approaches from pure, imputed, and spatial to test various scenarios within the modeling matrix (Bell, Kampman, Dodge, & Lawrence, 2022). Either a time series approach or stratified sampling is employed for splitting the data into training and testing sets (Bergmeir & Benítez, 2012). The same approach follows in the implementation of cross-validation methods, further ensuring the robustness of the predictive models across the variables. The motivation behind adopting a multiverse approach is to reinforce the internal validity of the project through cross-agreement among various models (Bell, Kampman, Dodge, & Lawrence, 2022). In addition, the project intends to introduce an approach to visualizing multiverse modeling in a user-friendly way. This will be accomplished through a matrix model visualization, implemented via an interactive dashboard using the program Streamlit.

2. Literature review

Missing data has been a persistent challenge in the field of data analysis, significantly affecting the reliability of statistical inferences (Van Buuren, 2023). Various authors, such as Rubin, Schafer, and Graham have contributed significantly to the development of techniques for handling missing data (Rubin, 1974) & (Schafer & Graham, 2002). Despite these advancements, especially the identification of predictors for the Missing at Random (MAR) conditions remains a gap in current research. This literature review is two-fold. The first part synthesizes relevant literature on the MAR assumption. The second part conceptualizes multiverse analysis and algorithmic modeling.

2.1 Missing at Random (MAR)

The MAR condition, a central assumption in the missing data mechanism, requires that the missingness of a specific variable is unrelated to its value when the values of other variables are controlled (Davison, 2003). This is formalized as:

$$Probability\ (Y = Missing \mid X, Y) = Probability\ (Y = Missing \mid X) \qquad (1)$$

Whether a data point in variable Y is missing is not related to the actual value of Y itself once one has accounted for X (Davison, 2003). A core assumption of the MAR condition is that all predictors of missingness must be accounted for in the statistical analysis (Little T. D., Jorgensen, Lang, & Moore, 2014). In other words, all factors which might influence whether incorporated data is missing need to be included in the analysis. The formalization of the MAR condition traces back to Rubin's research from 1974.

Rubin's research on the causal effects of treatments in both randomized and nonrandomized studies helped set the stage with formalized frameworks for missing data (Rubin, 1974). Rubin highlights the necessity of comprehending the mechanism behind missing data, in addition to employing a suitable model for such data. Doing so, by putting forth three categorizations for missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). According to Rubin, the missing data mechanism significantly affects the way it should be treated in statistical analysis, subsequently guiding the selection of an appropriate missing data (Rubin, 1974). The introduction of these concepts; the mechanisms and models for treatment, has had profound effects on the field of missing data and has also greatly influenced statistical inference. However, despite Rubin's technical and practical formalization of MAR, the focus of this project and review, the complexity of verifying the MAR assumption in practical research situations has been a consistent issue.

Schafer and Graham's 2002 paper "Missing Data: Our View of the State of the Art" highlights the difficulty of proving or disproving the MAR assumption in practice. They argue the MAR assumption is often untestable and may not hold in many real-world scenarios. For example, if participants drop out of a longitudinal study because of their worsening health status, which is not measured or observed, then the missing data are not MAR (Schafer & Graham, 2002). The authors also clear up common misunderstandings regarding the MAR condition and challenge Rubin's MAR findings. Specifically, they argue that Rubin's formalization of MAR as per Equation 1 should hold for all possible values of Y, not only for the Y that appeared in the sample as Rubin denoted (Schafer & Graham, 2002). Finally, they argue that methods able to handle non-MAR data, such as pattern-mixture models, selection models, and sensitivity analysis should be considered. They suggest that researchers should not rely on MAR as a default assumption, but rather explore other possibilities and evaluate their implications for their results (Schafer & Graham, 2002).

Bridging the gap between Rubin, Schafer and Graham's views on the MAR assumptions is Seaman et al (2013). According to them, of the two distinct definitions of MAR, one is stronger than the other (Seaman, Galati, Jackson, & Carlin, 2013). The stronger definition, which they call MAR1, states that the probability of missing data depends only on the observed data and not on the unobserved data, for all possible values of Y. The weaker definition, which they call MAR2, states that the probability of missing data depends only on the observed data and not on the unobserved data, for the realized values of Y in each sample (Seaman, Galati, Jackson, & Carlin, 2013). They illustrate that Rubin's (1976) original definition of MAR corresponds to MAR2, while Schafer and Graham's align with MAR1. Building on the above definitions they expand the arena of MAR and show that MAR1 is needed for direct-likelihood inference and Bayesian inference to be valid, while MAR2 is sufficient for frequentist inference using the likelihood function to be valid (Seaman, Galati, Jackson, & Carlin, 2013). The authors conclude that the choice of inference framework and missingness model should be guided by the research question and available data.

Over the past decade, literature on 'missing at random' has significantly increased. Data from Scopus show a 201% increase for the term 'missing at random' in article titles, abstracts, and keywords comparing 2010 to 2022 (Scopus, 2023). Meanwhile, during this popularization, researchers have advanced the arena of MAR. A brief look at the contributions of van Buuren & Little, Jorgensen, Lang, and Moore helps to assert the gap in the literature that the project aims to contribute to addressing.

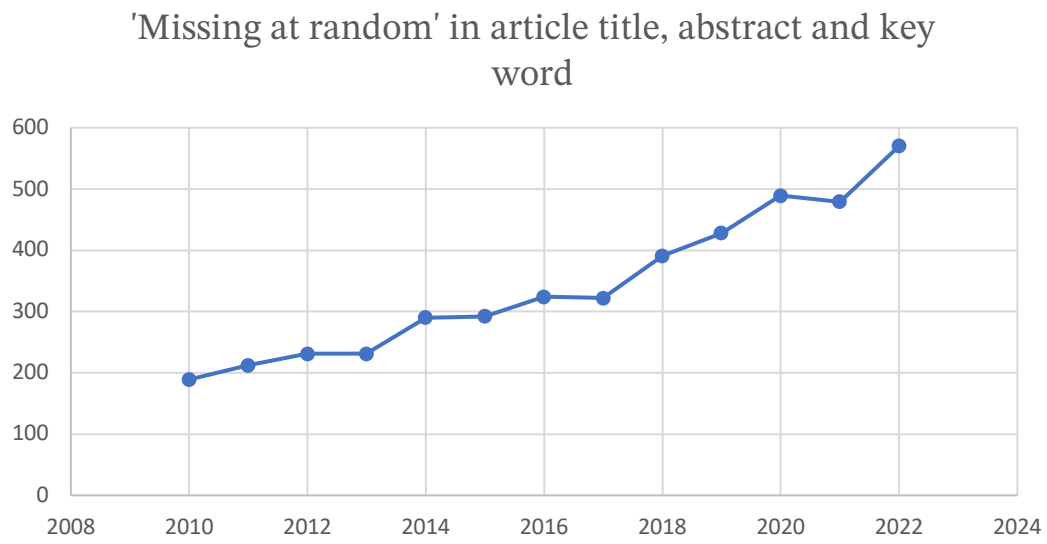

Figure 1 – Line plot of 'missing at random' in academia over time

Van Buuren's contributions helped advance open science and source for data imputation with the 'mice'-package for R (van Buuren, 2023). This statistical package is accompanied by his book 'Flexible Imputation of Missing Data'. Although straying a bit from the MAR condition, the contribution is arguably essential to mention for the field of missing data. The statistical package sets the scene for how contributions to the field could be delivered. Relating this to the MAR condition, in the 'Flexible Imputation of Missing Data' van Buuren argues that a remedy to fulfill the MAR assumption is to 'expand the data in the imputation model in the hope of making the missing data mechanism closer to MAR' (van Buuren, 2023).

Similarly, Little et al. (2014) provide a conceptual introduction to the mechanisms of missing data and the modern methods of handling them in pediatric research. Focusing on the MAR condition they argue that proactively accounting for likely causes of missingness in the estimated model can adjust parameters to accurately reflect the original population values (2014). For example, including the measure of religiosity as an auxiliary variable in the survey instrument can remedy this. However, this approach also addresses limitations and challenges, such as identifying the relevant auxiliary variables and assessing their validity (2014).

While the inclusion of auxiliary variables has been proposed as a method to bring the data mechanism closer to MAR, this often remains a 'hopeful' approach based on educated guesswork rather than an analyzed framework. This project aims to contribute to this gap, by applying rigorous techniques of multiverse analysis and algorithmic modeling to identify predictors of missingness. This shift, moving from hopeful inferences to more definitive findings, represents a significant step forward in the study of MAR conditions.

2.2 Multiverse Analysis

To further improve the understanding and strengthen the findings of predictors of missingness for the data, the project draws on the principles of multiverse analysis. Multiverse analysis is a scientific method that runs a set of plausible alternative models for a single hypothesis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). It provides a solution to the issue of researcher degrees of freedom, where multiple decision points may introduce variability in research outcomes. The multiverse method was developed in response to the credibility and replication crisis in science, since it can diagnose p-hacking and provide insight into how different model specifications impact results for the same hypothesis, thereby guiding researchers towards better theory or causal models (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

The arbitrary decisions made during data processing create a multiverse of reasonable data sets, each leading to different potential statistical results. Any arbitrariness in data construction is therefore inherited by the statistical outcome (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). Selective reporting that privileges one arbitrary data set from the range of possible data sets effectively ignores the multiverse of statistical results. To mitigate this, the project will employ a sequential data processing process, where the multiple possible data sets are all utilized for modelling. A major strength of this approach is that results have increased internal validity as varied processing frameworks and data frames are tested and synthesized to find a consistent conclusion (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016).

2.3 Algorithmic Modelling

Algorithmic modeling represents a data-driven approach that can employ a range of algorithms with the primary objective of optimizing a specific parameter or metric (Courtney, Thacker, & Clark, 1997). This approach contrasts with traditional, top-down, theory-driven statistical modeling, which often starts with a theoretical hypothesis about the data and then applies statistical methods to test this hypothesis. Instead, algorithmic modeling often starts from the bottom up, beginning with the raw data and leveraging computational power to explore and optimize patterns within the data itself (Kolkman, 2020).

Algorithmic modeling can excel at identifying complex interactions and nonlinear relationships within large data sets, making it well-suited for discovering strong predictors (Kolkman, 2020). This approach offers greater flexibility and adaptability than traditional statistical modeling, which relies on preconceived hypotheses and assumptions. Consider the challenge of enhancing the accuracy in diagnosing heart diseases. In a theory-driven

approach, a researcher might start with a hypothesis based on established medical knowledge. For example, they may hypothesize that specific demographic factors and genetic factors contribute to heart disease. The researcher would then set up a statistical model to test this hypothesis, such as a logistic regression model, with heart disease as the outcome and the hypothesized factors as the predictors. In contrast, an algorithmic model would begin with the raw data. Using machine learning algorithms, the model would analyze the available data to identify patterns and relationships for attributes that best predict heart disease, irrespective of any preconceived hypotheses (Courtney, Thacker, & Clark, 1997).

Despite the power of algorithmic modeling to extract patterns from large and complex data sets, it is not without challenges, such as the bias-variance trade-off. If not addressed, these challenges may lead the models to overfit or underfit the underlying data, resulting in an inaccurate representation of the true population (Courtney, Thacker, & Clark, 1997). However, algorithmic modeling's strength lies in its ability to utilize large and complex data sets to extract subtle patterns and relationships that might not be evident using traditional statistical methods. This doesn't render traditional methods obsolete. The choice between a theory-driven or data-driven approach relies on the specific research question, the available data, and the context (Kolkman, 2020).

To conclude, while significant progress has been made in understanding and conceptualization of the MAR condition, gaps still exist. Specifically, there is a need to transition from educated guesswork to a streamlined framework when including predictors of missingness. The project aims to analyze features importance from models which perform the best for the classification of missingness. The importance of model performance in this context lies in its reflection of feature relevance. Poor model performance typically indicates a weak association between the predictor and the target variable. Thus, it serves as a key measure of the effectiveness of the selected predictors in modeling. Predictors of missingness will be identified and considered for the body of open-source knowledge of the MAR condition. These predictors are features that consistently classify missingness across different models, cross validations, and data domains in the HFI data set. This leads to the two-fold research question of this study:

2.4 Research questions

1. Which models yield the best performance for classifying missingness in the Human Freedom Index?
2. For these top-performing models, which predictors contribute the most to feature importance for classifications of missingness in the Human Freedom Index data?

3. Data

For this project, the study area is set on a global scale encompassing 98.1 percent of the world's population by countries in a longitudinal study setting (Fraser Institute, Cato Institute, 2023). By having both a temporal and spatial aspect of the data, settings such as spatial distributions and temporal patterns can be investigated in both the data processing parts as well as the modeling. Analyzing at the global level provides a comprehensive macro-perspective of the HFI data and enables the project to account for different spatial patterns around the world (Li, Sun, & Fang, 2018). Countries are constantly in development and changes occur which can be monitored and analyzed, making this global setting of analysis important for the investigation of missingness mechanisms and their patterns across regions and in different time contexts (Li, Sun, & Fang, 2018). The four target attributes used for classification are ef_score, ef_government_tax, pf_expression_bti, and pf_rol_civil. These were chosen for their general representation of different patterns of missingness.

3.1 Human Freedom Index

The primary data source for this project is the Human Freedom Index (HFI) data set, jointly produced by the Cato Institute and Fraser Institute. This spatial-temporal public data set spans two decades, from 2000 to 2020, providing a comprehensive longitudinal perspective (Fraser Institute, Cato Institute, 2023). The HFI data set is expansive, consisting of 141 individual attributes. Out of these attributes, 112 contain some degree of missing data, requiring consideration during data preprocessing and imputation (Fraser Institute, Cato Institute, 2023).

The HFI comprises both continuous and categorical variables. Continuous variables include various indices and scores measuring different aspects of human freedom. These scores, which are generated based on numerous parameters and data sources, provide a nuanced view of the state of human freedom in each country (Fraser Institute, Cato Institute, 2023). Categorical variables include country names and region identifications, offering valuable contextual information. These variables are critical in understanding the variation and evolution of human freedom in different geographical and cultural contexts.

This data set was specifically chosen due to several important aspects. Firstly, the spatial and temporal dimensions of the data enable a more universal and holistic modeling approach that can be investigated across other data sets. Secondly, significant proportions of missing data are present, which is a necessity for classifying missingness. Additionally, the data set's value is underscored by the meaningful economic, personal, and human freedom scores it provides yearly (Fraser Institute, Cato Institute, 2023). There are also

some limitations of the data set. While the temporal aspect is explicit in the data, with a year attribute, the spatial aspect is implicit as countries are not marked with geometries or coordinates. To remedy this, the project adds other data sources, to the latter parts of the data preprocessing in the multiverse matrix, to enable spatial modeling.

3.2 Geometries

Complementing the HFI data set, the project also incorporates data from OpenDataSoft's GeoJSON geometries. This secondary data set presents an important addition of data where each country, represented by its ISO-3 code, has an associated geometry (OpenDataSoft, 2023). Adding the data provides a robust geographical context to the human freedom data, potentially revealing spatial patterns and variations in the data that might otherwise go unnoticed.

An important aspect of this data set is that it is licensed under the Open Government Licence v3.0, demonstrating a commitment to data transparency and open access. As such, it can be freely utilized and shared, connecting with this project's spirit of open-sourced research (OpenDataSoft, 2023). The integration of these geographical data points with the HFI data will enrich the data analysis and visualization, helping to draw more holistic and context-aware conclusions. It serves to provide a spatial dimension to the economic, personal, and human freedom scores recorded annually in the HFI data, broadening the scope and depth of the analysis (OpenDataSoft, 2023).

3.3 Global North and South

An additional layer of geographical context is introduced for the analysis through the integration of data from the World Population Review's ranking of the global north and south countries. This information is combined with the existing data frames to include a combined economic and spatial attribute (World Population Review, 2023). The World Population Review's ranking helps understand the global divide in economic development and wealth, which is often referred to in terms of 'Global North' and 'Global South'. The data offers an interesting perspective on the differences in human, personal, and economic freedom between developed and developing nations (World Population Review, 2023). However, it is essential to note that there is no universally accepted metric for classifying global north and south countries. At best, this is a proxy for information regarding the combination of economic and geographical conditions.

4. Methods

To best describe and propose the methodology section when working with a multiverse approach consider the visualization below. The visualization demonstrates the multiverse matrix approach which is sequential and iterative in nature. 3 different data frames are utilized and preprocessed. These 3 are layered on top of each other, in the sense that the imputed becomes an expanded version of the pure with imputation. Meanwhile, the spatial is an expansion of the imputed with spatial feature engineering. This sequential approach makes the methodology section a bit less linear, as iterations across multiple multicollinearity checks and VIF analyses are performed. To best understand where the different data frames appear, the 'checkpoint' subtitle will be used when a variant of the 3 data frames is created and extracted for modeling and analysis.

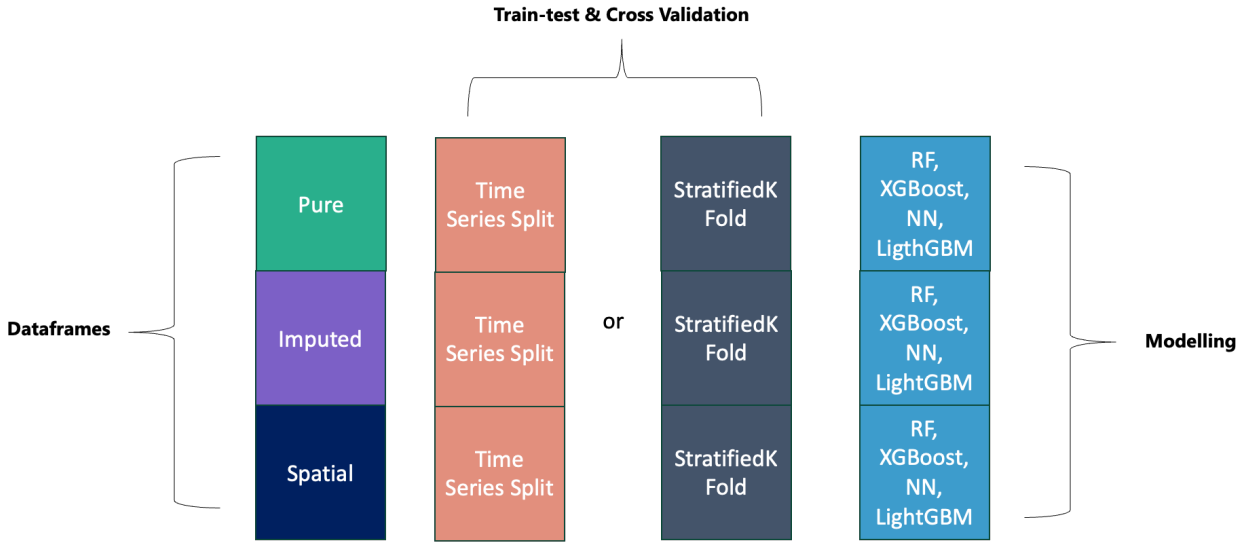

Figure 2 - Overview of methodology for multiverse matrix approach

For this project, a GitHub repository and YAML file ('msngdata') have been created for environment control, configuration, and reproducibility. The project uses Python 3.9.16 along with the needed packages.

4.1 Preprocessing: Addressing Multicollinearity

Multicollinearity is an essential step to investigate during the preprocessing phase. Multicollinearity, as a concept, occurs when there exists a high correlation among two or more independent variables. This makes it a challenge to decipher the distinct association each independent variable exerts on the dependent variable (Bhandari, 2023). The first step to address this is creating correlation matrixes which are visualized and computed statistically. Thus, in this phase, potential instances of multicollinearity are detected and selected, to be addressed in the VIF analysis. By addressing this issue, the project aims to

enable more accurate algorithmic modeling and classification with the best subset of the attributes (Bhandari, 2023). Below is a snapshot of a part of the original data frame with a heatmap visualization of the correlation.

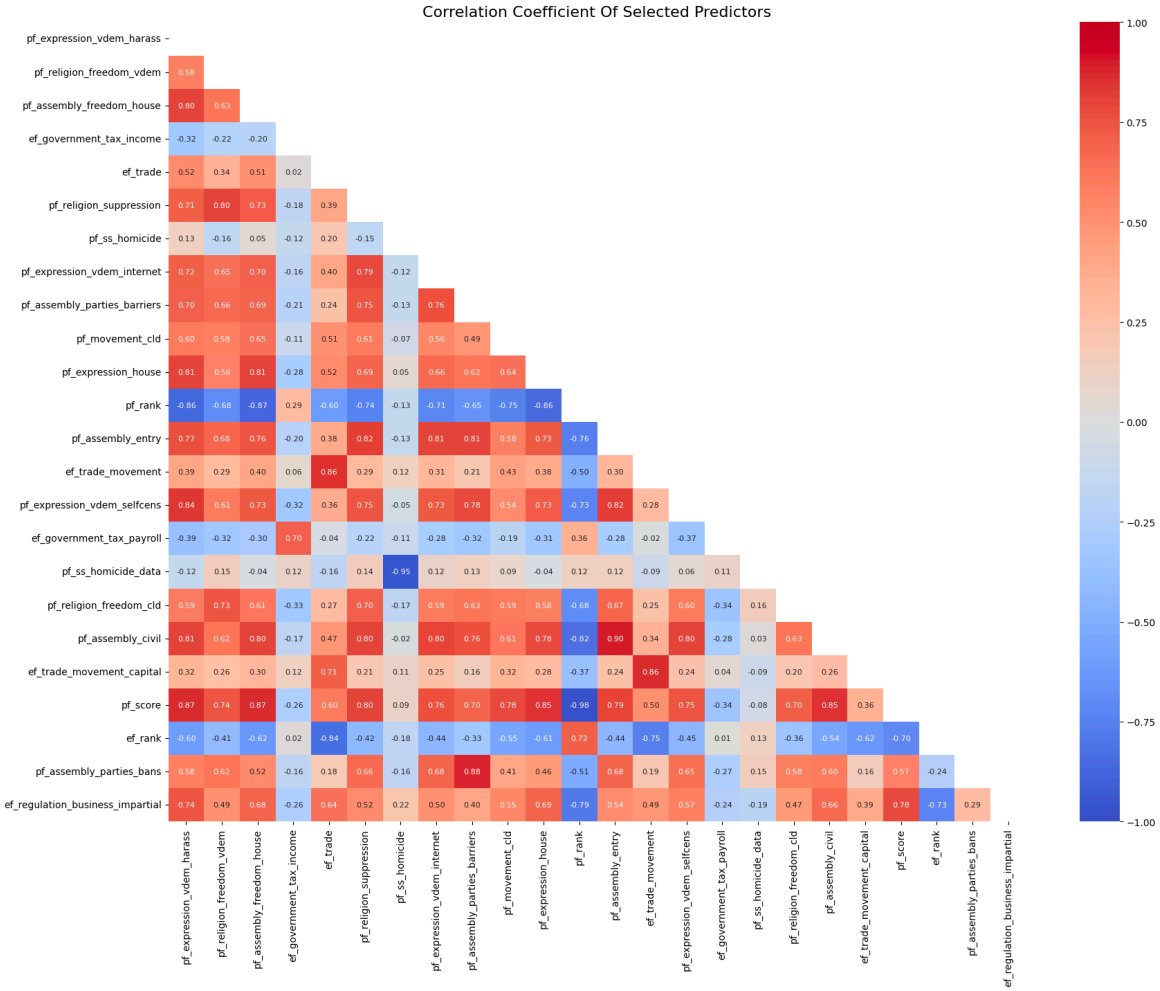

Figure 3 - Heatmap for correlation amongst attributes in the HFI data frame

The visualization suggests that there is a mix of positive correlation (colored red) and negative correlation (colored blue) in the data. The cutoff for the correlation threshold is set to 0.85. There is no correct number for this setting, and it is a matter of context and purpose of the analysis (Bhandari, 2023).

4.2 Variance Inflation Factor (VIF) analysis and target selection

After identifying predictors with high absolute correlation values, the next stage involves conducting a Variance Inflation Factor (VIF) analysis. This analysis helps by identifying and removing those variables that could potentially inflate the results of the subsequent modeling due to their high multicollinearity with other independent variables (Potters, 2023). In the context of this project, the VIF analysis is conducted iteratively using a function that calculates the VIF for each variable by implicitly fitting a linear regression

model. The VIF quantifies the degree of increased variance in the estimated regression coefficients due to multicollinearity (Potters, 2023).

A threshold for VIF is set at 10, a common practice in the field. Variables with VIF above this threshold are considered to have high multicollinearity. They are sequentially removed from the data set, starting with the variable having the highest VIF. This process continues until no variable in the data set has a VIF above the threshold. Through this process, the VIF analysis ensures the integrity of the modeling results by mitigating the potential impact of multicollinearity. This approach contributes to the robustness and reliability of the attribute subset, thus facilitating more accurate and meaningful modeling.

Below is a visualization of the overall missingness of the data frame across rows and columns. The white spaces represent missingness while the blue represents completeness. The four target attributes selected for classification are ef_score, ef_government_tax, pf_expression_bti, and pf_rol_civil.

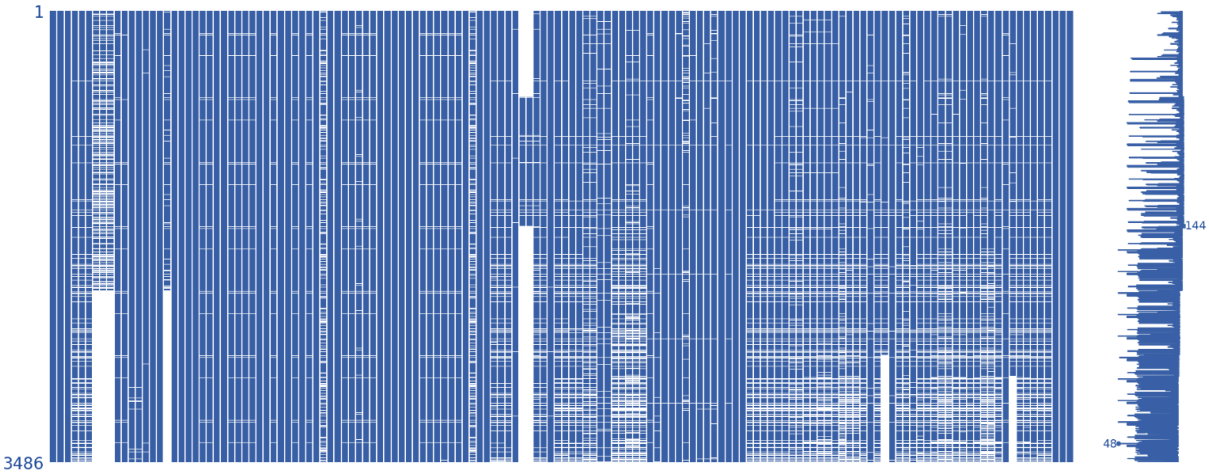

Figure 4 - Missingness matrix of HFI data frame

What becomes evident from the above visualization, is the distinctive different patterns of missingness. The four target attributes each represent a general pattern of missingness as seen in the above plot. By employing four distinct targets, the project aims to check for any overlap between attributes associated positively with predicting and classifying the above target's binary missingness. These four distinctive patterns were located with a histogram of the percentage of missingness for the different columns.

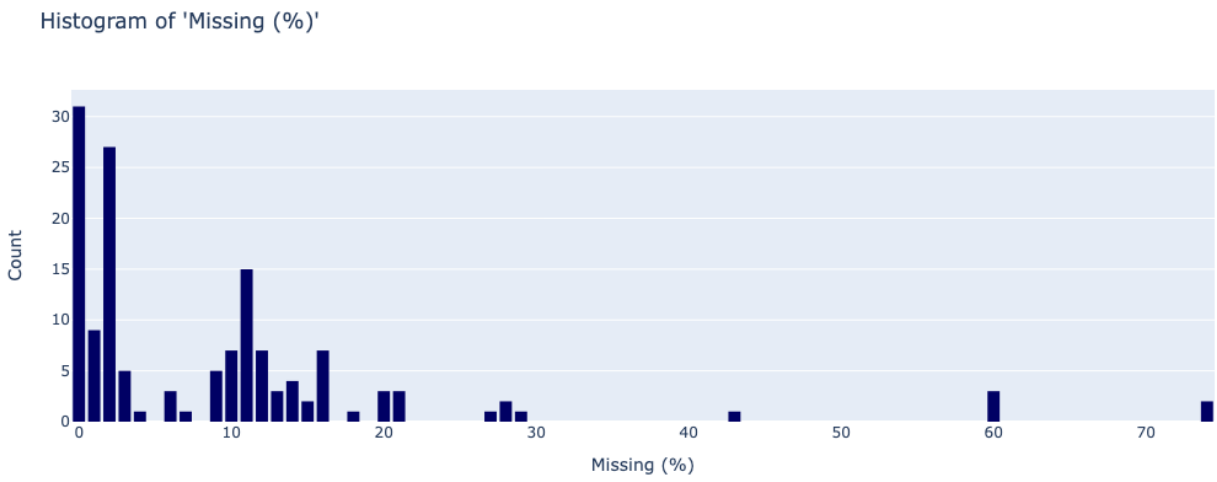

Figure 5 - Histogram of missing data in percentages

4.3 Checkpoint 1: Pure data frame

After performing the above preprocessing steps of correlation checks and VIF analysis the pure version of the data frame was constructed. This data frame consists only of the HFI data (no spatial explicit data mergers) while removing all columns with missingness except for the four targets. This leaves the data frame with 23 attributes for modeling. This version is the least processed which will be modelled on. The idea is to give statistical perspectives on how performance changes for classification modeling when deeper preprocessing steps are computed.

4.4 Imputation

Next to the heart of this project lies the models to impute missing data. Two of these are utilized for the second data frame 'imputed'. It is important to note here, that the data is rolled back to its original nature, thus not removing columns with missing values. The adopted approach for addressing this involves a two-step process, utilizing interpolation and KNN (k-Nearest Neighbors) imputation.

Processing: To ensure compatibility with the imputation methods, preprocessing was required. The data set contained both numerical and non-numerical data. For this data frame, non-numerical columns, 'region' and 'ISO-code', were identified and extracted while the 'countries' column was used for grouping, leaving only the numerical data subject to the imputation process.

Interpolation: The first step in the imputation process was linear interpolation. Before linear interpolation, there were 22262 missing values. Given the temporal nature of the

data, linear interpolation provided a suitable method to fill in gaps in the data sequence. This method was applied on a per-country basis to account for the spatial attribute of the data (Li & Parker, 2008). The effectiveness of this step was then evaluated by calculating the remaining missing values. After, there were 2163 missing values.

KNN Imputation: After performing interpolation to fill out missing data, some gaps remained. To further address this, the KNN imputation method was implemented. KNN was chosen due to its effectiveness in handling spatial-temporal data (Li & Parker, 2008). The algorithm was set to consider the six nearest neighbors and used uniform weights for the distance calculation. The KNN imputed data set was then evaluated to ensure no missing data remained. There were no missing values after this step.

4.5 Checkpoint 2: Imputed data frame

After performing the above imputation preprocessing steps, the imputed version of the data frame was extracted. This version is more processed than the prior pure version. An important distinction here is the extra imputed attributes not available in the pure data frame. This version has 30 attributes.

4.6 Feature engineering: Data Merging and One-Hot Encoding

This section focused on enhancing the models' ability to account for spatial factors. This enhancement involves the construction of the third and final data frame, created specifically to include spatially explicit attributes. Initially, the 'imputed' data frame merges with two additional data sets. These include the geometries and the global north & south attributes, previously identified. This merging process relies on ISO-3 codes, which are present across all three data frames. Following the merger, a secondary feature engineering phase begins. To enable analysis at global and semi-global levels in terms of spatial factors, the 'region' attribute, originally categorical, undergoes one-hot encoding. This transformation is an integral part of the feature engineering process. Additionally, the centroids of the geometries are extracted, along with their respective x- and y-coordinates. These metrics serve as solid analytical decision boundaries within the spatial dimension.

4.7 Checkpoint 3: Spatial data frame

The final variation for the multiverse analysis is completed after the feature engineering. This enables a sequential stepwise analysis of data frames from pure, to imputed to spatial. This version has 42 attributes. A look at the next part of the methodology helps address

the temporal and imbalanced aspects of the data with train-test splitting and standardize and optimize the three variations of the data frames for modeling.

4.8 Time series & stratified train-test splitting, standardization

Two distinctive approaches for train-test splitting are employed to account for the different characteristics of the data frames. These are the time-series split and the stratified split. The time-series split acknowledges the temporal dimension in the data frames. By respecting the sequence of the data, this method ensures minimal data leakage concerning the time aspect (Bergmeir & Benítez, 2012). This approach aligns well with real-world scenarios where future predictions are based on past data. However, this method can lead to more imbalanced classes, which is a trade-off one must consider (Bergmeir & Benítez, 2012). Conversely, the stratified split is applied because of the class imbalance present in three of the four targets. This approach attempts to preserve the proportion of classes in each split, ensuring a more balanced representation of each category (Martinez & Zeng, 2000). This is vital in ensuring the trained model does not become biased toward dominant classes. Nevertheless, it also has its drawbacks. While stratification does well at handling class imbalance, it might inadvertently introduce minor data leakage between temporal aspects (Martinez & Zeng, 2000).

Thus, the choice between these two methods is not straightforward, as each offers its unique advantages while posing potential challenges. However, as this is a multiverse analysis approach, both will be employed and compared respectively. Now having split the data into train and test sets, the following step of feature scaling can be introduced.

4.9 Features: Standardization & Selection

To ensure optimal conditions for algorithmic modeling the code creates standardization of numerical attributes. This process is applied to the numerical data, excluding binary one-hot encoded regions, coordinates, and year. This ensures a uniform scale across the data set, making it more digestible for algorithms to process (Mohamad & Usman, 2013). The rationale for this step is as follows: Standardization removes the units of measurement from the data by rescaling the data to have a mean of 0 and a standard deviation of 1. This allows different variables to be compared on equal terms. In addition, many machine learning algorithms require this as they do not perform well when the input numerical attributes have very different scales (Mohamad & Usman, 2013). By ensuring consistent ranges of data, the methodology aims to avoid biasing the model towards features with higher magnitudes. This helps the model to learn more effectively and improves overall predictive accuracy (Mohamad & Usman, 2013).

The most important predictors are selected in two ways. For Random Forest, XGBoost, and LightGBM: After each model is trained and optimized, the .feature_importances_ attribute is accessed. This attribute provides the feature importance scores directly as these models inherently compute feature importance during training. The higher the importance score, the more the feature has contributed to the models' decision-making process. For the Neural Network: Feature importance is approximated differently because neural networks do not inherently provide feature importance. Here, the code computes the mean of the absolute values of the weights (.coefs_) connected to the input layer of the network. This indicates how much each feature influences the models' predictions. Lastly, the top 3 predictors in terms of highest importance for each model were extracted and then aggregated in total across all data frames to find the most important feature predictors when classifying missingness in the HFI data.

Overfitting was an issue with more than 20 predictors for each fold and an increasing number of predictors for the imputed and spatial data frames. To address this, a sparse application of the Sequential Feature Selection (SFS) technique was carried out. This was carried out for each target, for each cross-validation folds for each data frame. Thus, ensuring a multiverse representation of the attributes. Essentially, SFS was used to trim the set slightly, to prevent overfitting and producing more robust and generalizable results (Aggrawal & Pal, 2020). Although SFS employs a specific model, its goal is to identify a set of predictors that, together, are useful for predicting the target variable. The retrieved predictors provide valuable information that can be leveraged by all the selected types of models, not just the one used in the SFS process (Aggrawal & Pal, 2020).

4.10 Algorithmic Modeling

This analysis uses four distinct models: Random Forest, XGBoost, Neural Network, and LightGBM. These models were selected for their individual strengths and unique contributions to the field of machine learning (Mahesh, 2020).

Random Forest, an ensemble learning model, uses a collection of decision trees, each trained on distinct data subsets. This model excels due to its simplicity, clarity, and resistance to overfitting. Its mechanism reduces the impact of outliers and noise, but care must be taken to avoid underfitting by adjusting the complexity of the forest (Schonlau & Zou, 2020). XGBoost, another ensemble learning method, works on the principle of correcting the preceding predictor's residual errors. Known for its optimization for speed and performance, XGBoost easily handles missing values and resists overfitting through an added regularization term (Sagi & Rokach, 2021). However, model configuration needs careful attention to prevent underfitting or overfitting.

Neural Networks process algorithms that identify patterns in data. These networks excel at learning intricate non-linear relationships and autonomously perform feature extraction (Samek, Montavon, Lapuschkin, Anders, & Müller, 2021). However, they may overfit training data if the network's complexity is not properly managed. Mitigation techniques such as dropout, early stopping, or L1/L2 regularization can help control overfitting. LightGBM, a gradient boosting framework, grows trees vertically, making it faster and reducing memory usage. Like XGBoost, LightGBM has built-in features to prevent overfitting.

In summary, these models stand out for their speed, performance, and their ability to model complex relationships. They also offer measures against overfitting (Mahesh, 2020). However, the risk of overfitting or underfitting always exists, necessitating mindful model selection, hyperparameter tuning, and validation.

4.11 Grid Search Hyperparameter Optimization

Grid search hyperparameter optimization was utilized to enhance each model's performance. This involves examining combinations of parameter tunes and using cross-validation to identify the optimal tune for the best performance. For example, the Random Forest model tested variations of 'n_estimators', 'max_depth', and 'min_samples_split'. XGBoost model optimization tested 'n_estimators', 'learning_rate', 'max_depth', and 'gamma'. The Neural Network model adjusted 'hidden_layer_sizes', 'learning_rate_init', 'max_iter', and 'alpha'. For the LightGBM model, optimized parameters were 'n_estimators', 'learning_rate', and 'max_depth'.

In the Random Forest model, 'n_estimators' varied from 5 to 20, impacting the number of trees in the forest and computational complexity. The XGBoost model had learning rates of 0.01, 0.05, and 0.1, influencing model learning speed and convergence. For the Neural Network model, 'hidden_layer_sizes' with values 50, 100, and 150 affected the network's capacity to learn complex patterns. In the LightGBM model, 'max_depth' varied between 2 and 4, affecting the complexity and risk of overfitting the (Bergstra & Bengio, 2012). Varying these parameters helps explore the models' behavior, pinpointing the optimal configuration for the classification of missingness. The settings can have a strong influence on the models' learning and generalization capabilities (Bergstra & Bengio, 2012). Grid search optimization was crucial in fine-tuning the models for best performance. All the different grids searched are available in the GitHub code (modeling notebooks).

4.12 Model Performance Assessment

The performance of each model on the four distinct targets was evaluated using several key performance indicators. Specifically, the macro F1 score, Matthew's correlation coefficient, and balanced accuracy were employed to assess the models' effectiveness.

The macro F1 score serves as a reliable metric for measuring the accuracy of the model on a binary classification task, especially when class distribution is imbalanced (Lipton, Elkan, & Narayanaswamy, 2014). The F1 score is a harmonic mean of precision and recall, providing a single metric that balances the trade-off between these two crucial aspects, making it particularly useful for evaluating the performance of models in tasks with an imbalanced class distribution (Lipton, Elkan, & Narayanaswamy, 2014). It calculates the F1 score independently for each class and then takes the average, treating all classes equally.

$$F1\ score = 2 \cdot \frac{precision\ \cdot\ recall}{precision + recall} \quad (2)$$

$$Macro\ F1\ score = \frac{\sum_{i=1}^{n} F1\ Score_i}{n} \quad (3)$$

The balanced accuracy metric was employed to overcome the drawback of standard accuracy which might be misleading in the case of imbalanced class distributions. Balanced accuracy gives equal weight to the positive and negative classes, preventing a bias towards the majority class.

$$Balanced\ accuracy = \frac{1}{2} \cdot (recall + sensitivity) \quad (4)$$

Lastly, Matthew's correlation coefficient (MCC) was utilized (Chicco & Jurman, 2020). This metric provides a more complete view of the model's performance as it takes into consideration true and false positives and negatives. It is particularly useful in binary classification tasks and yields a value between -1 and 1. A coefficient of +1 represents a perfect prediction while 0 indicates a random prediction and -1 implies total disagreement between the prediction and observation (Chicco & Jurman, 2020).

$$MCC = \frac{TN \cdot TP - FN \cdot FP}{\sqrt{(TP \cdot FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

Using these three performance indicators allows for a comprehensive evaluation of each model, taking into consideration not just the number of correct classifications, but the nature of the errors made, and the balance between sensitivity and specificity Fields

(Chicco & Jurman, 2020). This way, a holistic understanding of the models' capabilities and areas of improvement can be obtained.

5. Results and analysis

The Results and Analysis section is structured in two parts. The first is the initial exploratory data analysis. This is the shortest section and aims to introduce the statistical nature of the data for the temporal and spatial components. Following this is the selected algorithmic modeling results, which answer the research questions. The aim here is to showcase the multitude of interesting results that follow a multiverse approach.

5.1 Selected data exploration results

Below is a plot of the four targets and their binary balance. It is intended to illustrate the degree of balance, or lack thereof, among the classes. From this snapshot the attributes' completeness shows a distinct bias of missingness, indicating a class imbalance.

Figure 6 - Overview of target variables

An examination of the temporal aspect of the data reveals interesting changes in the line plots for the freedom of expression, money inflation, and personal freedom score across the world. From 2000 to 2020, there has been a steady decrease in the mean values for the freedom of expression and personal freedom scores from 6.76 to 6.28 and 7.6 to 6.87, respectively. Meanwhile, there have been fluctuations in the standard deviation for money inflation.

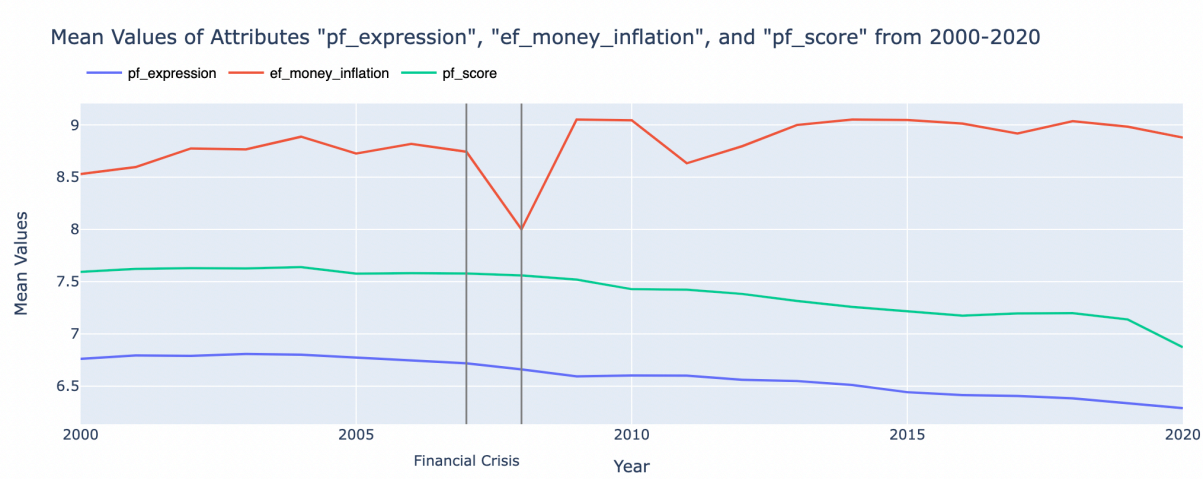

Figure 7 – Line plot of attributes from 2000 to. 2020

Personal freedom and expression move similarly across time in parallel. Intuitively, this seems in accordance as they could be associated. According to the Freedom in the World report by Freedom House, one of the suppliers of data for the HFI, the global freedom score has been declining since 2006 (Freedom House, 2023). The report also states that the decline in freedom is due to the rise of authoritarianism and populism (Freedom House, 2023). Meanwhile, the standard deviation for money inflation has fluctuated, especially during the financial crisis as marked on the plot. Turning to the spatial aspect shows two new attributes of 'government' and 'minimum wage' as well as the money inflation attribute again, grouped by region.

The boxplot showcases how spatial factors appear to influence these attributes. Take, for example, money inflation (represented by the green boxplots), and observe how North America, Western Europe, and Oceania all have relatively low minimum values (9.23, 7.59, and 6.8 respectively). This is a positive sign for this attribute, where high scores signify low inflation. Conversely, Eastern Europe and Sub-Saharan Africa have much lower minimums, some even at 0. The boxplot reveals underlying geographical differences. Thus, it is important to consider both the temporal component seen in the point plots and the spatial component seen here in the box plots. Incorporating these factors into the algorithmic modeling will aim to produce the most robust results for predicting missingness in the four target attributes.

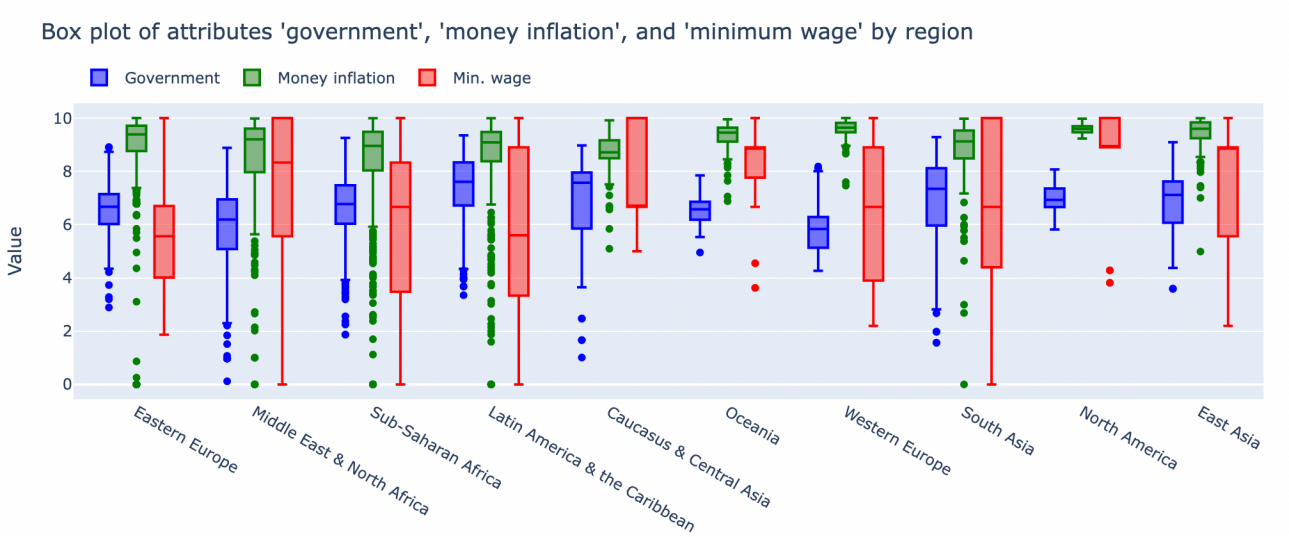

Figure 8 - Boxplot of attributes grouped by region

One of HFI's core attributes, the economic freedom score, shows what countries are at the top and bottom according to the Cato and Fraser Institute's data analysis. The data is grouped from 2018 to 2020 to give a recent snapshot of the freedom across these different segments.

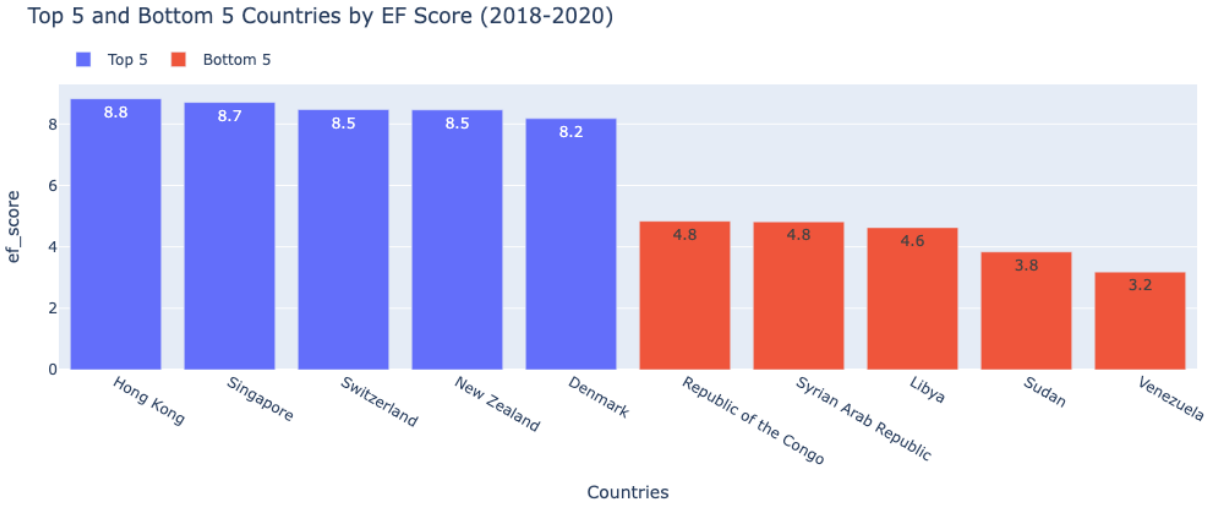

Figure 9 - Top 5 and Bottom 5 countries for economic freedom score (2018-2020)

Comparing the economic freedom scores, Hong Kong and Singapore top the list, while Sudan and Venezuela are at the bottom. Hong Kong and Singapore's economies are known for free trade via low taxes and almost tariff-free ports (Richards, 1993). Their robust legal systems, flexible labor markets, and entrepreneurial-friendly regulatory environments contribute to their high economic freedom scores as per the Cato and Fraser Institutes' rankings (Richards, 1993). In contrast, Sudan and Venezuela's low scores result from varied issues. Sudan's economy has long been hindered by conflicts, political instability limited access to capital, and weak property rights (Abbadi & Ahmed, 2006). Likewise, Venezuela has long been in turbulent economic weather with multiple occurrences of hyperinflation and widespread corruption, which contributes to its low score (Abbadi & Ahmed, 2006).

This comparison highlights the correlation between a country's economic freedom level and its economy's overall health (Abbadi & Ahmed, 2006). Countries with higher scores generally foster environments conducive to business and individual growth, leading to prosperity. Conversely, countries scoring lower often face significant economic issues due to systemic problems like corruption, inflation, and substantial state interference. The final part of the exploratory data analysis is a look at the grouped completeness across all attributes and time for each country. In the map below, the lighter-yellowish the color, the more complete the data is for a country.

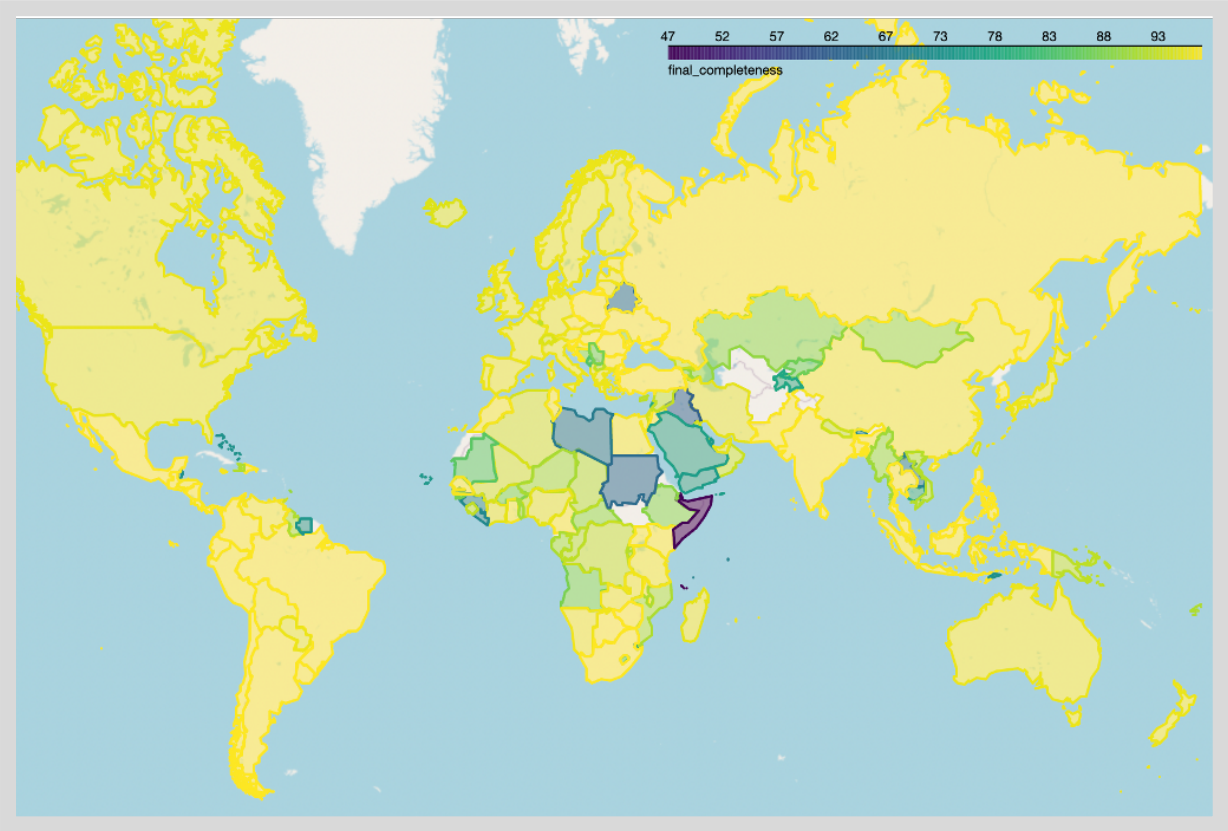

Figure 10 - Global map of completeness per country for all attributes

Overall, there is high data availability and the missingness does not fall below 47% for any country. The highest data availability is around 97% for Hong Kong and Sweden. What becomes evident is that there might be certain spatial patterns to this missingness. A look at the map reveals that potential missingness might arise in areas of the global south, for example in Sudan and Iraq which have 64% and 62% completeness respectively. This analysis reveals a pattern that is prominent outside of the HFI data sets as well. Overall, the global south tends to be underrepresented in data collection in many fields of study like this one. This lack of data can be attributed to a variety of reasons, including limited resources for data collection, issues of conflict and instability, and a lack of technological infrastructure (Brown, Saxena, & Wall, 2023). Therefore, it is critical to address these gaps in data collection to ensure a more accurate and equitable representation of global trends,

providing a more comprehensive picture of global conditions. The HFI aims to mitigate this by triangulating its data sources across different cultural and political institutions so that there are multiple sources for information on the freedom (Fraser Institute, Cato Institute, 2023).

In summary, the selected data exploration analysis has highlighted the differences across time and space for the attributes. For time, changes were evident in personal freedom and fluctuations in money inflation between 2000 and 2020 on a global scale. In other words, the temporal aspects of the data should be considered in the modeling. For space, the region-based boxplots revealed distinct differences in money inflation amongst other attributes. Finally, for the economic freedom score and completeness with respect to countries, the global south and north divide became clear.

5.2 Selected algorithmic modeling results

To answer the questions: Which models yield the best performance for classifying missingness in the Human Freedom Index? and 2) For these top-performing models, which predictors contribute the most to feature importance for classifications of missingness in the Human Freedom Index data? An overview of the algorithmic modeling results is presented. This is done in a manner that utilizes the multiverse approach. Specifically, the results will be analyzed and compared across variations of data frames and cross-validation methods to provide the most coherent and robust argumentation for the research question.

RQ1: Models. Analyzing all the results, the models yielding the best and most consistent classifications of missingness are LightGBM and XGBoost, with average macro F1 scores of 0.754 and 0.751 across all four target attributes, data frames, and cross validations, respectively.

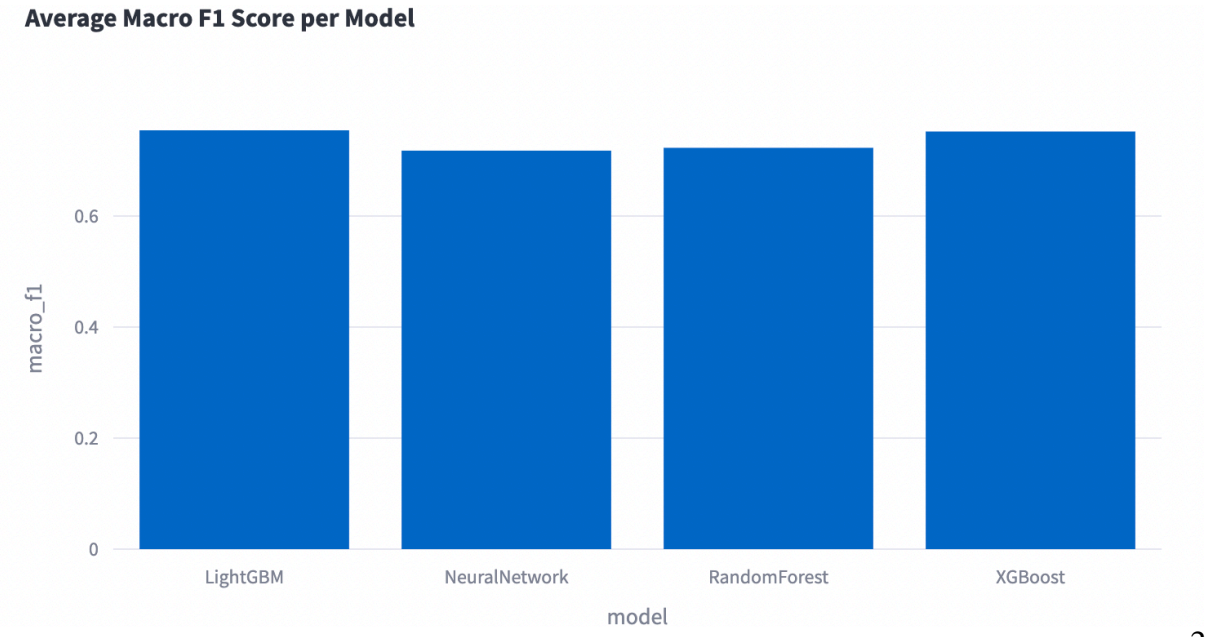

Figure 11 - Average Macro F1 Per Model (all multiverses)

As evident from the visualization above, Neural Networks and Random Forest perform worse than LightGBM and XGBoost. However, it is important to compare the performance of the models through other metrics. Looking at the violin plot for the average Matthew's correlation, the argument remains positive for XGBoost and LightGBM as the best performing models, but the analysis highlights how Neural Networks have a higher ceiling, as well as a lower floor for performance. Neural Networks have a median value of 0.5, while XGBoost has a median value of 0.515. However, the Neural Networks' third quartile (Q3) is 0.72, while XGBoost's Q3 is 0.64. This signifies that although the medians are relatively close, Neural Networks have a higher potential for correctly classifying and achieving a higher Matthew's correlation score.

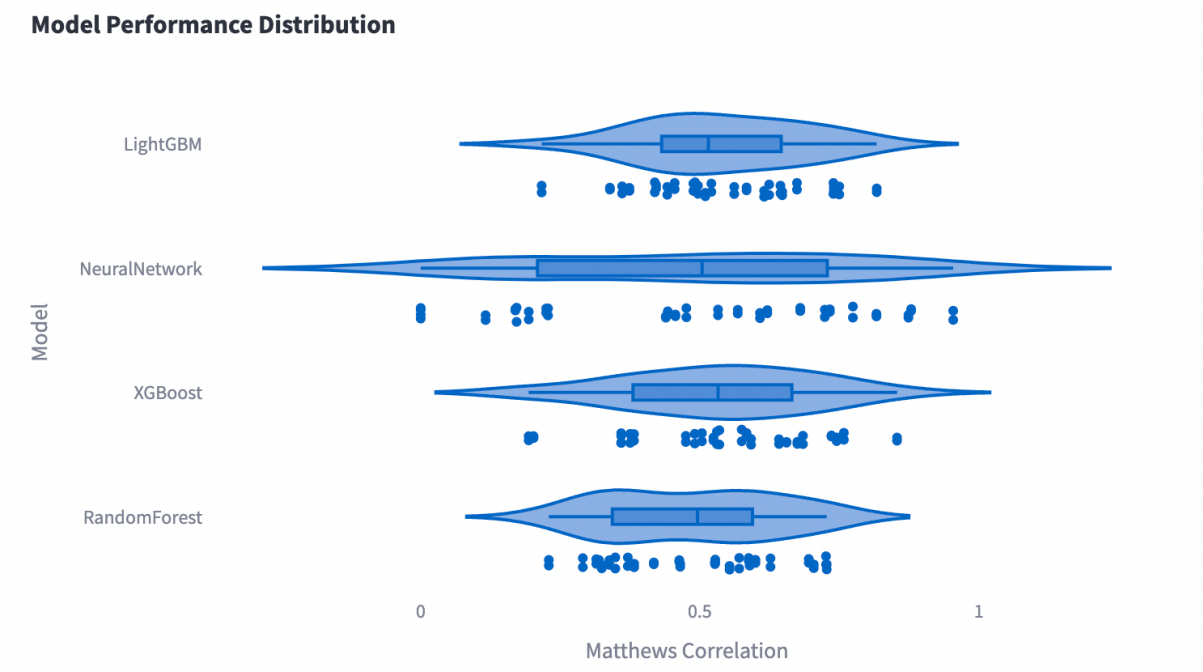

Figure 12 - Model Performance Distribution, Matthew's correlation (all multiverses)

The models' performance for the macro F1 metric and Matthew's correlation suggest that XGBoost and LightGBM perform the best overall, on average, for classifying missingness, while Neural Networks have the highest performance potential, as seen through Matthew's correlation metric. By shifting to another data dimension through a multiverse portal, more specifically the spatial data variation, the models' hypotheses are tested. Examining the same metrics, the macro F1 score, and Matthew's correlation but now solely for the spatial data frame variation, the best performing model emerges as the Neural Network. By showcasing another bar chart, the analysis reveals how changing small choices in the multiverse approach can have significant changes in the results.

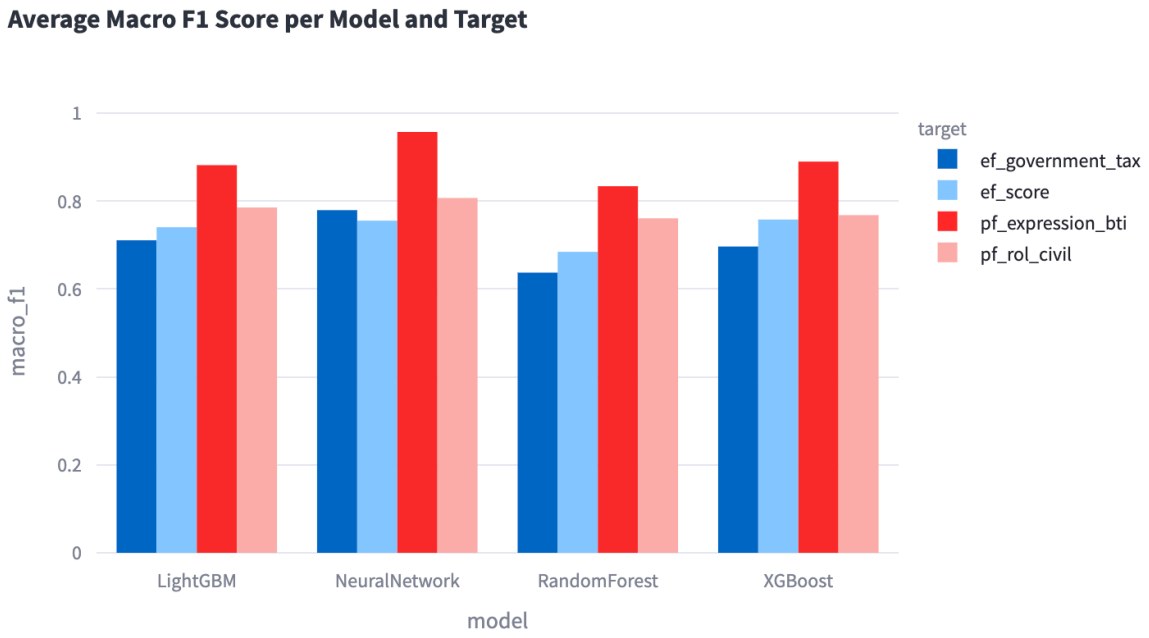

Figure 13 - Average Macro F1 per model and target for spatial data frame

For this multiverse variant, across all targets, and all cross-validations, but now limited to the spatial data frame, the Neural Network stands out as the best-performing model with a robust macro F1 score of 0.82. Moreover, the Neural Network has a Matthew's correlation median value of 0.74 while XGBoost has a median value of 0.55. While both values have increased compared to all data variations, the Neural Network has increased significantly more. The approximate reason for this change becomes evident when also considering the attributes and their importance. However, before addressing these, the argument, as stated above, remains that the best models for robust and consistent classifications of missingness continue to be LightGBM and XGBoost, given their superior performance across all multiverse variations. To extract the best performing models a cut-off is used for Matthew's correlation coefficient at 0.60. Essentially, the models with the best performance of all data frames and cross validations for all targets are extracted.

RQ2: Features. By looking at the best performing models, as those with Matthew's correlation coefficient above 0.60, the predictors with the most feature importance and consistency for classifying missingness are year, pf_movement, ef_gender, and the spatially explicit x- and y-coordinate features. These most important features for this multiverse view of the best performing models illuminate the significance of spatiality in explaining the marked improvement in Neural Networks' performance metrics. Incorporation of the x- and y-coordinates into the modeling enhances the complexity and enables the pinpointing of key countries across time. This suggests the potential influence of geographical factors on the data, indicating that patterns of missingness might not be randomly distributed but rather associated with specific locations.

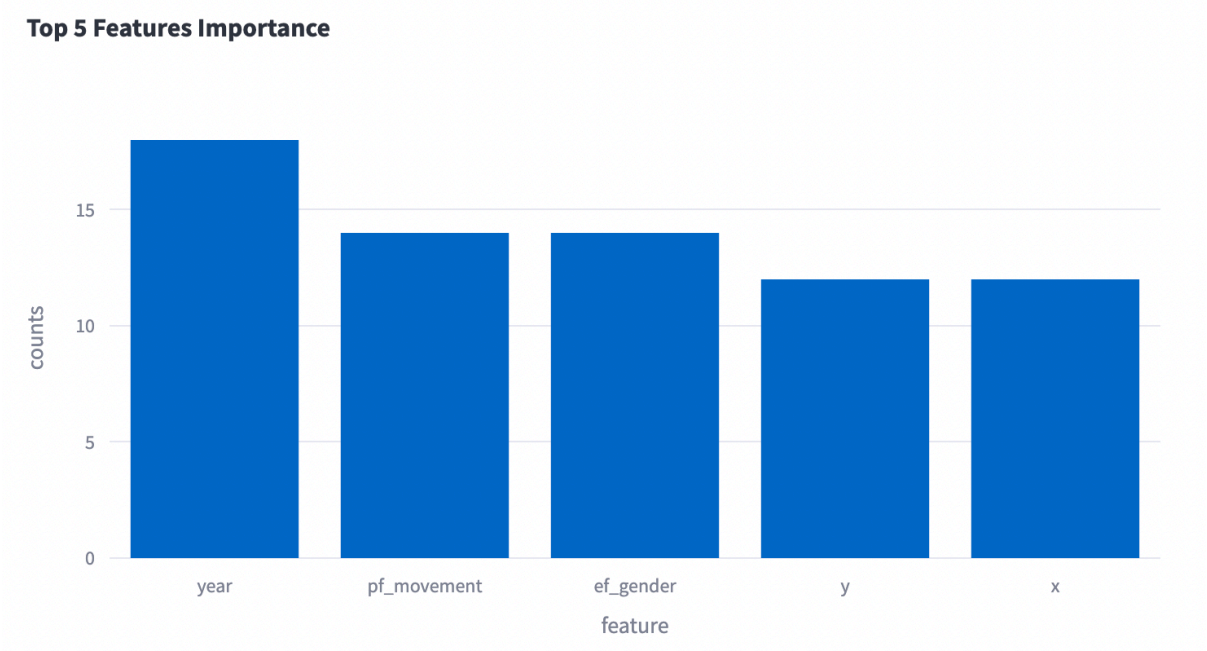

Figure 14 - Top 5 Feature Importance for best performing models (MCC >= 0.6)

Moreover, ef_gender and year emerge as consistently important attributes for classifying missingness in all data variations for the best performing models. This could be because these variables capture key societal and temporal dynamics that may be associated with data missingness.

As showcased above, is important to include ef_gender as it is a consistent predictor of missingness. Arguably because it reflects not just biological differences but also societal inequalities and legal barriers that can differentially impact data availability. The Gender Disparity Index (GDI), used in the Economic Freedom of the World Index, highlights the significant influence of legal and regulatory barriers on women's economic activities in certain countries (Cater, Yoon, Lowell, & Campbell, 2018). This means that in these regions, women may face additional challenges in accessing and engaging with resources, services, and opportunities. The value of this attribute has a strong association with data missingness in the target attributes. The project argues that the ef_gender attribute is critical as it encapsulates a complex interplay of (in)equalities of gender on the spatial scale, making the interaction effect amongst these important to consider. It is arguably an overlooked, but important, predictor of missingness.

Separately, the year attribute captures temporal trends or events that lead to periods of increased or decreased data missingness. This attribute can be vital in identifying shifts in data patterns over time, reflecting evolving societal norms, policy changes, technological advancements, or even wars and conflicts. These temporal factors can influence data availability and should be accounted for to ensure a robust understanding of the data's availability and inherent missingness mechanisms.

The graph for all variations in the multiverse analysis no matter the model performance, unsurprisingly, does not include the x and y attributes. However, ef_ gender, year, and pf_movement feature again, reinforcing their consistent importance in classifying missingness. Generally, the most crucial attributes for classifying missingness are year, pf_movement, and ef_gender. Yet, the multiverse approach suggests incorporating x- and y-coordinates as well, given that the spatial data variation boasts the highest macro F1 score of all three approaches as seen below with the highest score of all with 0.77.

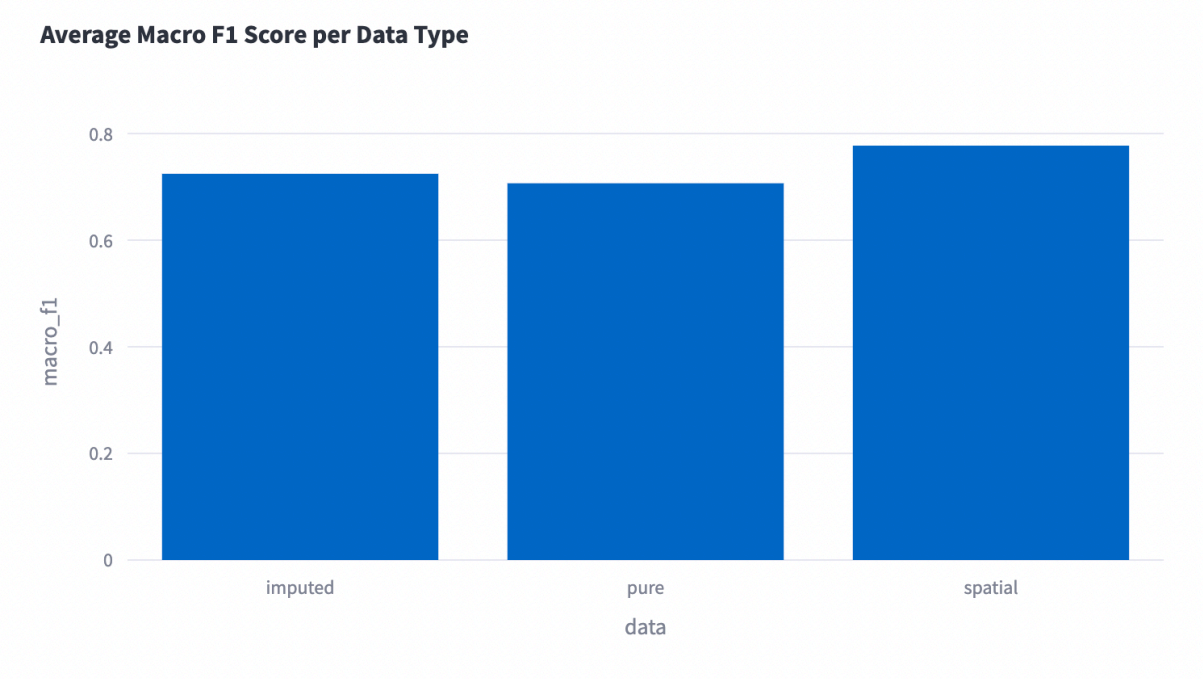

Figure 15 - Average Macro F1 Score per Data Type

In analyzing the pf_movement variable, which signifies 'Freedom of Movement', it's apparent that it could serve as a proxy for numerous larger societal and political contexts. As this attribute encapsulates citizens' rights to travel, settle, depart from, and return to their country without restrictions, it reflects a nation's governance quality and political climate. For instance, countries with high 'Freedom of Movement' scores might indicate democratic, open, and transparent governance, while low scores could suggest autocratic regimes or political instability.

Moreover, the 'Freedom of Movement' status can shed light on the overall human rights situation and prevalent societal norms within a country. It's a fundamental human right and a core civil liberty, and its presence or absence can provide insights into various aspects such as freedom of speech, the rule of law, and societal openness. In regions where freedom of movement is severely restricted, data collection may pose significant challenges. Citizens' limited ability to choose their place of residence or to travel freely might lead to incomplete or uneven data representation. On the other hand, in areas with

unrestricted freedom of movement, data could potentially be more complete, considering the ease of information flow and broader representation.

The recurring importance of pf_movement across various models and multiverse variations emphasizes the influence that societal freedoms have on data availability. Therefore, when constructing models to predict or account for missingness, incorporating democratic variables like pf_movement, along with other significant features such as year, ef_gender, and spatial coordinates, could result in more accurate reflections of the missingness mechanisms. A look at the cross-validation techniques casts a technical light on the model performance, feature importance, and underfitting of the algorithmic models through multiverse choices.

The graph below showcases the average macro F1 score per model and CV method for all data frames. The data suggest that StratifiedSplit cross-validation performs significantly better than TimeSeriesSplit across all models. For example, in the Neural Network, the macro F1 is 0.64 for TimeSeriesSplit cross-validation and 0.79 for the StratiefiedSplit.

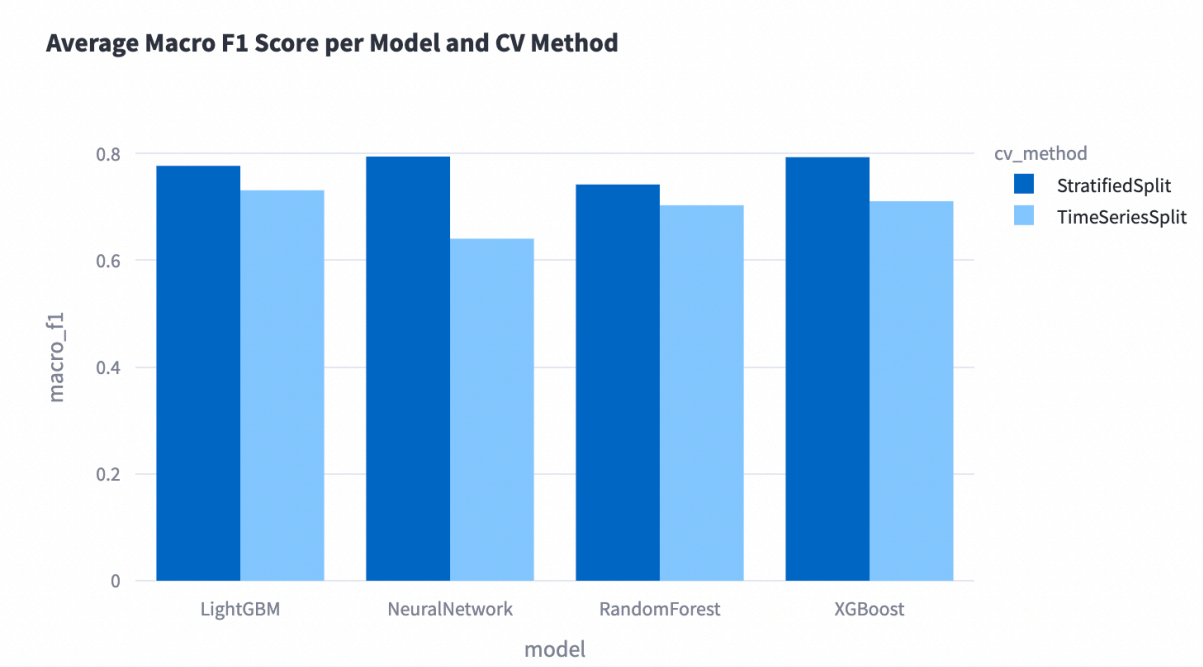

Figure 16 - Average Macro F1 Score per Model and CV Method

StratifiedSplit's better performance when classifying missingness could be attributed to its ability to maintain the proportion of samples for each class in both training and testing data sets, despite the significant class imbalance. This approach ensures that minority classes are adequately represented during model training, which is critical for the development of a well-generalized model. On the other hand, TimeSeriesSplit might not effectively capture the minority class, especially if occurrences of this class are sporadically

distributed over time. As a result, there could be instances where TimeSeriesSplit creates train-test splits with minimal representation or even the absence of the minority class in the training set. This circumstance can lead to underfitting of the minority class, where the model fails to capture important distinctions between classes and performs poorly.

To finalize the selected algorithmic results, a heatmap of the targets and the models' respective performances is presented. This heatmap, scoring with balanced accuracy, helps cement the argumentation of which models are the best and which features they used while also showing how the performances varied across the targets.

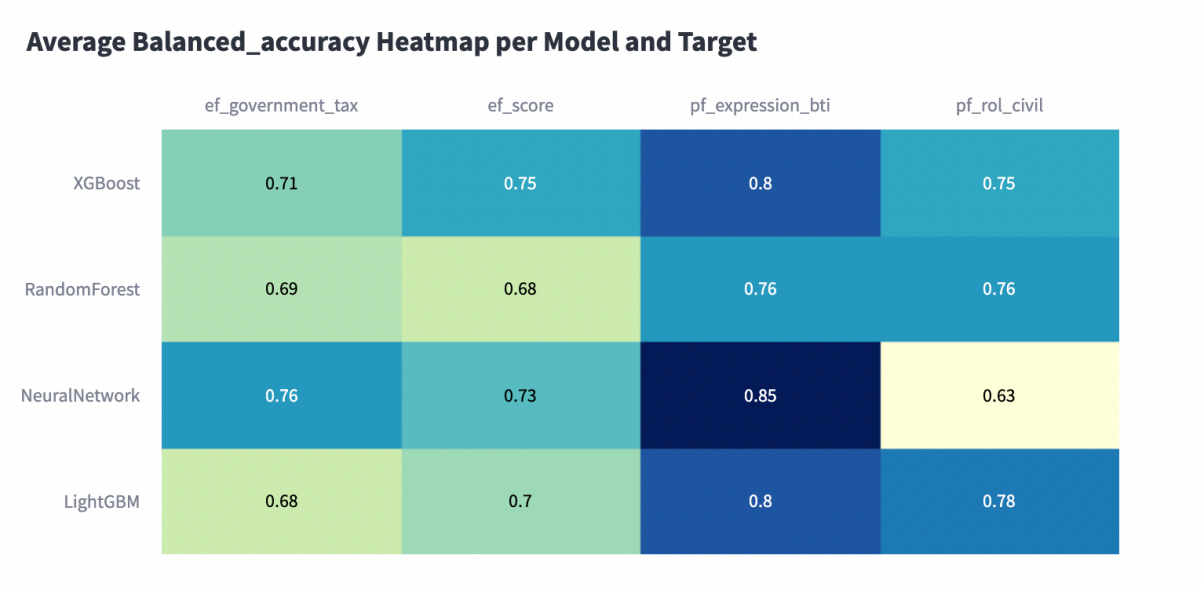

Figure 17 - Average balanced accuracy Heatmap per Model and Target

Viewing the results through the lens of balanced accuracy, both XGBoost and Neural Network emerge as the top-performing models. However, it's worth noting a certain variability in their performances across different targets. For instance, when applied to the pf_rol_civil target, the Neural Network model exhibits the least impressive performance among all four models, yielding a balanced accuracy of just 0.63, while the next lowest score stands significantly higher at 0.75. Overall, the balanced accuracy confirms that XGBoost and LightGBM perform the best, while also confirming that Neural Networks have the widest performance range, with high highs and low lows. Finally, the heatmap can be altered to the spatial version, which interestingly showcases that Neural Networks change from the worst to the best performing model for the pf_rol_civil target. Again, this reemphasizes the spatial feature's importance in predictions of missingness.

To summarize the analysis, a comprehensive exploration of various models, including XGBoost, LightGBM, Neural Networks, and Random Forest, was conducted, specifically focusing on their abilities to classify missingness robustly and consistently within the Human Freedom Index data. Based on a variety of metrics and tests across different data frames and cross-validation methods, LightGBM and XGBoost emerged as the most reliable models for this task. However, the analysis revealed that the choice of data frame significantly impacts model performance, as exemplified by the improved performance of Neural Networks when spatial data variation was considered. This variation allowed for a more complex model and emphasized the importance of spatially explicit x- and y-coordinate features, along with ef_gender, year, and pf_movement attributes.

Moreover, the study revealed the most importance of the year, pf_movement, ef_gender, and x- and y coordinate attributes in classifying missingness, for the best performing models in all data frames and cross validations. Finally, the analysis underscored the importance of employing multiple evaluation metrics in model selection. Focusing solely on one measure could lead to misleading conclusions and suboptimal outcomes. The results obtained provide an insightful foundation for understanding the mechanisms of data missingness and demonstrate the necessity of incorporating important variables to achieve more accurate predictions. A discussion of the limitations of the findings and the ethical data science implications is presented to address how future work can improve upon this project.

6. Discussion

Discussing the implications and limitations of the project helps understand how future research can contribute to the field.

Ethical implications and considerations

One implication of the results that greater model performance and feature importance are achieved by adding spatially explicit attributes, such as x- and y- coordinates, is the ethical consideration of treating the global south as an auxiliary coordinate variable. On one hand, adding spatial attributes can help fulfill the MAR assumption and improve the imputation of missing data. It allows for a more accurate representation of the data and can potentially uncover patterns or relationships that are associated with missingness. From the perspective of a data scientist, this may seem a reasonable and correct way to address the issue of the MAR mechanism. However, on the other hand, it is crucial to acknowledge the potential bias that may arise from treating the global south as an auxiliary variable. This approach assumes that missingness is related to the spatial location, which may overlook the possibilities of improving socioeconomic, political, or cultural factors that

contribute to data availability issues (Schatz, Angotti, Madhavan, & Sennott, 2015). In other words, there is a potential for treating the symptom rather than the disease.

To address this ethical consideration by considering the disease rather than the symptom, it is important to be aware and explicit about the limitations and potential biases associated with using spatially explicit attributes. Researchers should also consider alternative strategies to improve data availability in underrepresented regions, such as investing in data collection efforts and promoting inclusive data practices (Schatz, Angotti, Madhavan, & Sennott, 2015). Balancing the need for strong imputation modeling and fulfilling the MAR mechanism with the goal of equitable data representation is crucial in advancing ethical data science practices.

Multiverse approach

While the multiverse approach has certain advantages such as addressing the researcher's degree of freedom one must be cautious to avoid the potential pitfall of 'boosting' results. This occurs when running multiple, slightly similar, repetitions, potentially inflating the results (Rijnhart, Twisk, Deeg, & Heymans, 2021). Such inflation can bias results, resulting in a synthetic representation that does not accurately reflect the true population. To mitigate this limitation, future research could compare these variations in more detail. Specifically, it would be worthwhile to investigate the degree of similarity across variations of data frames and processing steps. For instance, one could examine the uniqueness of classes across different cross-validation strategies, such as time series and stratified methods used in this project, and identify any instances of overlap.

A second limitation of the multiverse approach in this study is the reliance on a single core data set of HFI. While this approach can enhance the internal validity of the results, it may limit their generalizability. To ensure the broader applicability of the findings, future studies could incorporate and compare additional, but conceptually similar, data sets. This would mean analyzing whether the predictors of missingness perform equally well across different data sets' target attributes. By doing so, one could increase the external generalizability of the identified features and test for natural overlap among the data sets.

Imputation

The project used interpolation and knn imputation in the processing part to fill out the missing values, but this holds a potential for reducing the variability in the data leading to biased estimates (Abma & Kabir, 2005). Interpolation assumes that there is a smooth continuity in the data, which can sometimes lead to an underestimate of the actual variability. This was observed for some parts of the data, while other parts of the data had a fixed scale of 0 to 10, as seen in the exploratory data analysis. Moreover, using knn

imputation can lead to imputations that do not accurately capture the complexity or structure of the data (Sanjar, Bekhzod, Kim, & Paul, 2020).

Other imputation methods can be considered to address these potential limitations of interpolation and knn imputation. Techniques such as multiple imputation or expectation maximization algorithms could provide a more accurate representation of the missing data. While multiple imputation and expectation maximization methods are more advanced than interpolation and knn imputation, they are not one-size fits all solutions. These methods provide better estimates under certain conditions. However, they rely on certain assumptions about why the data is missing. If these assumptions are not met, the methods may not perform well (Sanjar, Bekhzod, Kim, & Paul, 2020). An alternative to changing imputation methods is to show results both with and without imputation. As demonstrated in this project with the pure data frame, this approach can help understand how the imputation process adds to the results.

Model and feature importance

The type of models used has a strong association with the features extracted and their inherent importance measurements. In this study, the employment of tree-based models such as Random Forest, XGBoost, and LightGBM may unintentionally contribute to the previously discussed 'boosting' problem highlighted in the multiverse discussion. Using different types of models, also spatial in nature, such as Spatial Autoregressive Models and Spatial Error Models, could account for local and global patterns of non-response in the spatial predictors.

Moreover, while the study identifies key predictors of non-response in the HFI data set, the interpretability and causal relationships of these predictors could be extended for further analysis. A way to enhance the paper’s robustness and findings could be to visualize decision patterns for the tree-based models with nodes and splits. Understanding the underlying mechanisms and causal relationships can provide deeper insights into the missingness patterns. Additionally, feature importance measures based on algorithmic models, for example, the Neural Network used, might not always align with human interpretability, which could limit the practical understanding of the findings.

7. Conclusion

In conclusion, the project contributes to the understanding of missing data, a common cause of headaches in many fields of research, including psychology, social science, and machine learning. Through a comprehensive investigation of missingness in the Human Freedom Index (HFI) employing four key variables representative of common patterns of missingness in the data, this work identified key predictors and established the relative performance of several predictive models. The multiverse analysis approach allowed a thorough evaluation of a range of models, including XGBoost, LightGBM, Neural Networks, and Random Forest, across different data frames and cross-validation methods.

This analysis shows the best overall performance from the LightGBM and XGBoost models in classifying missingness, as demonstrated by macro F1, balanced accuracy, and Matthew's Correlation Coefficient measures. Simultaneously, year, pf_movement, ef_gender, and the spatial coordinates x and y are consistently key predictors across multiverse variations with the most feature importance. This emphasizes the necessity of considering explicit spatial attributes in data sets with spatial characteristics. The analysis further revealed that model performance can be significantly influenced by the choice of the data frame, as evidenced by the Neural Networks' improved performance when spatial data variation was considered.

The project also acknowledged potential limitations which could be addressed with further research. For example, for the multiverse 'boosting' of results and limited generalizability of the features and their external strength in other data sets. Future work should delve deeper into multiverse variations and include additional data sets to improve generalizability and move closer to fulfilling the MAR condition. By continuing to strive towards more holistic and reliable ways of addressing missing data, this research contributes to the ongoing effort to mitigate the headache of educated guesswork for fulfilling the MAR assumption often associated with this issue in data analysis and machine learning. For now, it is fair to say that a thoughtful inclusion of temporal, social, gender, and spatial factors can help parts of this headache when working with missing data patterns.

Bibliography

Abbadi, K. A., & Ahmed, A. E. (2006). Brief overview of Sudan's economy and future prospects for agricultural development. Khartoum Food Aid Forum (pp. 6-8).

Abma, R., & Kabir, N. (2005). Comparisons of interpolation methods. The Leading Edge, 24(10), 984-989.

Aggrawal, R., & Pal, S. (2020). Sequential feature selection and machine learning algorithm-based patient death events prediction and diagnosis in heart disease. SN Computer Science, 1(6), 344.

Allison, P. D. (2001). Missing Data. Thousand Oaks: Sage Publications, Inc. .

Bergmeir, & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Bhandari, A. (2023, May 10). Multicollinearity - Causes, Effect and Detection. From https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/#:~:text=Multicollinearity%20occurs%20when%20two%20or%20more%20independent%20variables%20have%20a,variable%20on%20the%20dependent%20variable.

Brown, S., Saxena, D., & Wall, P. J. (2023). Data collection in the global south: practical, methodological, and philosophical considerations. Information Technology for Development, 1-21.

Cater, S. W., Yoon, S. C., Lowell, D. A., & Campbell, J. C. (2018). Bridging the gap: identifying global trends in gender disparity among the radiology physician workforce. Academic radiology, 25(8), 1052-1061.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21, 1-13.

Courtney, P., Thacker, N., & Clark, A. F. (1997). Algorithmic modelling for performance evaluation. Machine Vision and Applications 9(5), 219-228.

Davison, A. C. (2003). Statistical Models. Cambridge University Press.

Freedom House. (2023, May 20). Freedom in the World. From The Data: https://freedomhouse.org/report/freedom-world

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200-1205).

Kolkman, D. (2020). The usefulness of algorithmic models in policy making. Government Information Quarterly.

Lee, K. J., & Carlin, J. B. (2012). Recovery of information from multiple imputation: a simulation study. Emerg Themes Epidemiol, 3-9.

Li, G., Sun, S., & Fang, C. (2018). The varying driving forces of urban expansion in China: Insights from a spatial-temporal analysis. Landscape and Urban Planning, 174, 63-77.

Li, Y., & Parker, L. E. (2008). A spatial-temporal imputation technique for classification with missing data in a wireless sensor network. IEEE/RSJ International

Conference on Intelligent Robots and Systems (pp. 3272-3279). IEEE.
Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. arXiv preprint arXiv:1402.1892.
Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.
Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.
Martinez, T. R., & Zeng, X. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. Journal of Experimental & Theoretical Artificial Intelligence, 12(1), 1-12.
Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17), 3299-3303.
OpenDataSoft. (2023, April 30). World Administrative Boundaries - Countries and Territories . From OpenDataSoft - Explore: https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/export/
Potters, C. (2023, April 30). Variance Inflation Factor (VIF). From Investopedia: https://www.investopedia.com/terms/v/variance-inflation-factor.asp
Python Software Foundation. (2023, May 24). Documentation. From Python: https://docs.python.org/3/
Richards, A. (1993). Hong Kong, Singapore, Malaysia, and the fruits of free trade. Organisation for Economic Cooperation and Development. The OECD Observer, (185), 29.
Rijnhart, J. J., Twisk, J. W., Deeg, D. J., & Heymans, M. W. (2021). Assessing the robustness of mediation analysis results using multiverse analysis. Prevention Science, 1-11.
Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. journal of Educational Psychology 66 (5), 688-701.
Sagi, O., & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. Information sciences, 572, 522-542.
Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE, 109(3), 247-278.
Sanjar, K., Bekhzod, O., Kim, J., & Paul, A. (2020). Missing data imputation for geolocation-based price prediction using KNN–MCF method. ISPRS International Journal of Geo-Information, 9(4), 227.
Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147.
Schatz, E., Angotti, N., Madhavan, S., & Sennott, C. (2015). Working with teams of “insiders” qualitative approaches to data collection in the global south. Demographic Research.
Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29.
scikit-learn. (2023, May 15). From https://scikit-learn.org/stable/
Scopus. (2023, May 15). Scopus Search. From

https://www.scopus.com/results/results.uri?sort=plf-f&src=s&st1=missing+data&sid=301f8b60505f1a736fc644c0679584f7&sot=b&sdt=cl&sl=27&s=TITLE-ABS-KEY%28missing+data%29&origin=resultslist&editSaveSearch=&yearFrom=2013&yearTo=2023&featureToggles=FEATURE_DOCU

Seaman, S., Galati, J., Jackson, D., & Carlin, J. (2013). What is meant by “missing at random”? Institute of Mathematical Sciences.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science.

Van Buuren, S. (2023, April 28). Flexible Imputation of Missing Data. From Stefan Van Buuren: https://stefvanbuuren.name/fimd/sec-problem.html

World Population Review. (2023, April 28). Global North Countries. From https://worldpopulationreview.com/country-rankings/global-north-countries

Appendix

Visualizations from the analysis can be found and interacted with in the in the result dashboard below.

- Interactive result dashboard: https://albertbanke-thesis-missing-data-2023-app-ou02mg.streamlit.app/

All the code and data frames used for this project can be found in the GitHub repository below.

- Code and data: https://github.com/albertbanke/thesis_missing_data_2023