US China Alignment in Year 2023

Enyu Li

September 7, 2025

1 Introduction

In today's rapidly shifting international landscape, questions of political and economic alignment have become increasingly critical. Global tensions between major powers such as the United States and China have reshaped diplomatic relations, trade networks, and security alliances. Countries often find themselves navigating a complex spectrum of alignment, ranging from pro-China, to pro-USA, to maintaining a neutral stance. Understanding these patterns is essential not only for policymakers, but also for businesses, researchers, and international organizations that must adapt to a multipolar world order.

This project seeks to provide a data-driven approach to mapping global alignments by leveraging machine learning and semi-supervised classification methods. Using a comprehensive dataset that integrates multiple dimensions—such as military cooperation, economic dependencies, democratic indices, sanctions, and geopolitical disputes—the model predicts each country's alignment category. Specifically, the classification labels are defined as:

- Pro-China (countries whose policies, economics, or security ties lean toward China),
- Neutral (countries balancing or avoiding alignment with either superpower), and
- Pro-USA (countries closely aligned with the United States).

To accomplish this, the project applies a semi-supervised learning pipeline that combines expert-labeled data with a self-training classifier, enabling the model to generalize alignment predictions for countries with limited prior labels. Feature importance and SHAP (Shapley Additive Explanations) analysis are used to interpret the contribution of individual covariates, offering transparency into which factors—such as trade dependencies, alliance scores, or regime type—drive the predictions.

Finally, the results are visualized in both static and interactive formats. A global choropleth map provides a high-level view of alignment predictions, while interactive Tableau dashboards allow for deeper exploration of country-level probabilities and covariate effects. Together, these tools form a comprehensive framework for analyzing and communicating the evolving dynamics of global alignment.

2 Research Objectives

2.1 Construct a comprehensive dataset

- Integrate diverse sources of information on political, economic, and military factors, including UN voting patterns, trade dependencies, democratic indices, alliance scores, sanctions by the USA, and military exercises, gdp per capita, economic structure, and geological region.
- Standardize and preprocess the dataset to ensure comparability across countries and time.

2.2 Classify countries into alignment categories

- Apply a semi-supervised learning approach using XGBoost with self-training to categorize countries as Pro-China, Neutral, or Pro-USA.

- Leverage both expert-labeled data and model-driven predictions to maximize coverage of countries.

2.3 Interpret model predictions

- Use feature importance of XGBoost and SHAP (Shapley Additive Explanations) analysis to identify the most influential covariates and their directions.
- Provide insights into how specific variables (e.g., economic dependence on China vs. the U.S.) shape alignment outcomes.

2.4 Visualize global alignment patterns

- Use feature importance of XGBoost and SHAP (Shapley Additive Explanations) analysis to identify the most influential covariates and their directions.
- Provide insights into how specific variables (e.g., economic dependence on China vs. the U.S.) shape alignment outcomes.

2.5 Contribute to broader understanding of geopolitical trends

- Offer a transparent, replicable methodology that can be updated as new data becomes available.
- Provide policymakers, researchers, and analysts with tools to assess how global alignments may evolve over the next decade.

2.6 Future Exploration

- Use the same method to predict the labels over the years from 2000 and find the global alignment pattern.
- Use a time series SARIMA to predict the labels for each country in the year of 2024 and 2025.

3 Construct the Dataset

The dataset is constructed from multiple data sources. Some of them are directly downloaded from the websites for free, and some of the dataset is created after extracting textual information from official documents. Below is the full list of all of the covariates used for analysis.

3.1 Variable Explanations

The dataset integrates multiple dimensions of political, economic, and security indicators. Below is a description of each variable:

- **country_clean**: Standardized country name for consistent merging across datasets.
- **year**: The observation year of the data point.
- **agree_China**, **agree_USA**: Proportion of agreement with China or the United States in UN General Assembly voting, derived from the UN Votes dataset.
- **Military_Imports_USA**, **Military_Imports_China**: Value of major arms imports from the U.S. or China, based on SIPRI Arms Transfers data.
- **US_military_aid**: U.S. military aid flows to a country, measured in constant dollars, from USAID / Foreign Assistance databases.
- **v2x_libdem**, **v2x_polyarchy**: Liberal democracy and electoral polyarchy indices from the Varieties of Democracy (V-Dem) project.

- **military_drills_China**: Count of joint military exercises with China in a given year, compiled from U.S. Department of Defense reports.
- **export_China_dependency**, **export_USA_dependency**: Share of a country's total exports destined for China or the U.S., from UN Comtrade / IMF DOTS.
- **import_China_dependency**, **import_USA_dependency**: Share of a country's total imports originating from China or the U.S., from UN Comtrade / IMF DOTS.
- **IdealPointsDistance_USA**, **IdealPointsDistance_China**: Ideological distance measures based on ideal point estimation from UN General Assembly voting similarity.
- **GDP_per_capita**: Economic development indicator, measured in constant dollars, from World Bank World Development Indicators.
- **subregion**: Regional classification of each country (e.g., Balkans, East Asia), based on UN geoscheme.
- **US_alliance_score**: Coded measure of U.S. alliance relationships, based on the Correlates of War (COW) Alliance dataset.
- **sanction_scores**: Index of sanctions imposed on the country, created based on the Global Sanctions Database.
- **sco_brics_score**: Membership or participation score in SCO and BRICS, capturing institutional ties to China-led blocs.
- **labels**: Expert-provided classification of alignment (Pro-China, Neutral, Pro-USA).
- **pred_label**: Predicted alignment label from the machine learning model.
- **p_proChina**, **p_neutral**, **p_proUSA**: Model-predicted probabilities for each alignment category.
- **total_label**: Final consolidated label after integrating expert and model predictions.
- **terr_dispute_China**: Indicator for whether the country has territorial disputes with China, from COW and conflict reports. The number 2 indicates a severe territorial or oceanic dispute with China. The number 1 indicates a minor dispute with China.
- **economic_structure**: Economic composition of the country (e.g., low-end manufacturing, mid-end manufacturing, services/finance, resources/agriculture).

3.2 Data Wrangling

To construct the final dataset, I integrated information from multiple international sources, including UN voting records, SIPRI arms transfers, World Bank economic indicators, V-Dem democracy indices, and various geopolitical datasets. Since these sources differ significantly in structure, coverage, and formatting, a substantial amount of data wrangling was required.

For this process, I relied heavily on R's dplyr package and SQL queries. Using dplyr, I performed operations such as data cleaning, variable renaming, filtering, and the creation of derived indicators (e.g., export/import dependency ratios, alliance scores). The dplyr syntax allowed me to efficiently chain together transformations and maintain readable, reproducible code.

In addition, I used SQL for merging large tables across different datasets. SQL joins were especially useful for linking records by country and year across heterogeneous sources, ensuring that the dataset preserved relational integrity. SQL also facilitated handling missing values, filtering historical ranges, and constructing subqueries for intermediate aggregates (such as average arms imports per country-year).

Together, these tools enabled me to standardize disparate datasets into a coherent panel structure with consistent country-year observations. The resulting integrated dataset provides the foundation for all subsequent modeling and visualization.

3.3 Data Preprocessing

To prepare the dataset for machine learning, all continuous numerical features were standardized using the `StandardScaler` from the `sklearn.preprocessing` module. This transformation ensured that variables measured on different scales, such as *GDP per capita*, *US alliance scores*, and *IdealPointsDistance*, were rescaled to have mean zero and unit variance, making them directly comparable in the modeling process.

For categorical features, two different strategies were employed. First, the `OneHotEncoder` was applied to the *subregion* variable, which created binary indicator columns such as `subregion_East Asia`, `subregion_Balkans`, and `subregion_North America`. This allowed the model to capture region-specific effects without imposing any artificial order.

Second, for the *economic structure* variable, which could take on multiple categories simultaneously (e.g., a country characterized by both *Mid-end Manufacturing* and *Services/Finance*), the `MultiLabelBinarizer` was used. This approach generated separate binary columns such as `Mid-end Manufacturing = 1` or `Services/Finance = 1`, correctly representing overlapping economic classifications.

Together, these preprocessing steps ensured that both numerical and categorical features were standardized and encoded in a format suitable for the semi-supervised XGBoost model.

4 Model Building

To classify each country into one of three alignment categories (*Pro-China*, *Neutral*, or *Pro-USA*), a semi-supervised learning framework was developed. This framework integrates expert-labeled cases with model-driven predictions, making it possible to extend reliable classifications even to countries where alignment is ambiguous or uncertain. The use of semi-supervised learning is particularly advantageous in international relations datasets, where reliable labels are scarce and many observations fall into a gray zone between competing spheres of influence.

The model-building process was carried out in three main stages, as detailed below:

1. **Choosing Covariates.**
   Before finalizing the covariates for modeling, I conducted diagnostic checks to avoid multicollinearity issues among predictors. Specifically, I examined the correlation matrix to identify highly correlated pairs of variables, ensuring that redundant information would not bias the estimation process. In addition, I computed the Variance Inflation Factor (VIF) for each covariate, using a threshold value of 5 as a guideline to flag problematic predictors. Variables with excessive VIF values were either removed or combined to reduce collinearity. These steps helped ensure that the selected covariates contributed unique, independent information to the model and improved both interpretability and stability of the results.

2. **Initial supervised training with XGBoost.**
   The first stage involved training an XGBoost classifier on a subset of countries with clear expert-assigned labels. XGBoost (Extreme Gradient Boosting) is an ensemble tree-based algorithm that combines the outputs of many weak learners (decision trees) into a strong predictive model through boosting. It was chosen for its robustness, efficiency, and ability to handle heterogeneous, tabular features with nonlinear relationships. The objective was set to `multi:softprob`, enabling the model to output probability distributions over the three classes rather than a single hard label. Hyperparameters such as the learning rate (0.08), maximum depth of trees (3), subsample ratio (0.8), column subsampling ratio (0.8), and regularization term ($\lambda = 1.0$) were tuned to balance accuracy and prevent overfitting. This produced an initial model that captured alignment patterns from expert-labeled data.

3. **Self-training with pseudo-labeling.**
   Since the majority of countries did not have ground-truth alignment labels, the base XGBoost classifier was incorporated into a `SelfTrainingClassifier`. This wrapper implements a pseudo-labeling strategy: the base model is first trained on labeled data, then used to predict labels for the unlabeled data. Predictions above a confidence threshold of 0.90 were treated as reliable pseudo-labels and added to the labeled training set. The classifier was then retrained on this expanded dataset. This process iterated up to 20 times, gradually enlarging the pool of labeled cases and allowing the model to propagate knowledge from confidently labeled examples into less certain cases. Self-training thus leveraged both expert knowledge and model inference, ensuring more complete global coverage.

4. **Final predictions and probability outputs.**
   After convergence, the final model generated predictions for every country-year in the dataset. In addition to discrete class assignments (`pred_label`), the model produced probability estimates for each class (`p_proChina`, `p_neutral`, `p_proUSA`). These probability distributions provide richer information than single categorical labels. For example, a country classified as *Neutral* with a probability of 0.55 but also having a 0.40 probability of being *Pro-China* reflects an ambiguous stance between neutrality and alignment with China. Such soft predictions are particularly valuable in international relations, where alignment is rarely absolute. The final dataset therefore contains both hard predictions for categorical classification and continuous probabilities for nuanced interpretation.

In summary, the semi-supervised framework combined **domain expertise** with **machine learning inference**. Expert labels anchored the model in ground-truth cases, while the self-training process expanded classification coverage by exploiting the structure of the feature space. The result is a comprehensive, interpretable mapping of global alignments that balances reliability, generalizability, and nuance. As we can see from the graph below:

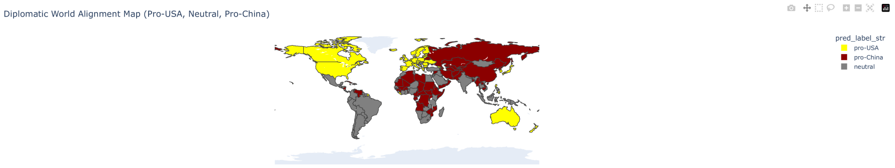
Diplomatic World Alignment Map (Pro-USA, Neutral, Pro-China)
pred_label_str
pro-USA
pro-China
neutral

Figure 1: Diplomatic world alignment map (Pro-China, Neutral, Pro-USA).

5 Key Factors Driving Country Alignment

The following variables were identified as the most important covariates in predicting whether a country aligns as Pro-China, Neutral, or Pro-USA:

> The following variables were identified as the most important covariates in predicting whether a country aligns as Pro-China, Neutral, or Pro-USA. Their relative importance scores are listed in parentheses:
>
> - **sanction_scores (0.155)**: Measures the extent to which a country is subject to U.S. or international sanctions. High sanctions often push countries away from the U.S. and toward China, while low or no sanctions leave countries more open to U.S. or Western partnerships.
> - **IdealPointsDistance_USA (0.143)**: Captures ideological and policy similarity with the United States, typically based on UN voting behavior. Smaller distances imply closer alignment with U.S. values and interests, while larger distances suggest divergence toward neutrality or China.

- **v2x_libdem (0.138)**: An index of liberal democracy (from V-Dem). Higher scores indicate democratic governance, which strongly correlates with pro-U.S. alignment. Lower scores increase the likelihood of neutrality or pro-China leanings.
- **IdealPointsDistance_China (0.115)**: Policy distance to China's positions. Smaller distances suggest closer alignment with China, while larger distances indicate divergence, often correlating with pro-U.S. positions.
- **subregion_West Africa (0.070)**: A geographic dummy variable capturing regional dynamics. Many West African countries balance between Chinese infrastructure investments and Western aid, making the region predictive of neutral or pro-China alignments.
- **subregion_Central Africa (0.063)**: Central Africa shows strong dependence on Chinese investment and resource extraction, explaining its predictive power toward pro-China alignments.
- **US_alliance_score (0.058)**: Reflects formal and informal defense or security cooperation with the United States. High values strongly predict pro-U.S. alignment (e.g., NATO members), while low scores allow for neutrality or pro-China drift.
- **GDP_per_capita (0.051)**: A measure of economic development. Higher income countries are more likely to be pro-U.S., while lower income countries are often more influenced by Chinese aid and infrastructure projects.
- **subregion_Middle East (0.037)**: The Middle East balances U.S. security ties with growing Chinese energy partnerships, often resulting in neutral alignments.
- **terr_dispute_China (0.029)**: Indicates whether a country has territorial disputes with China. Countries with disputes (e.g., India, Vietnam, Philippines) tend to lean pro-U.S., while those without such conflicts may remain neutral or pro-China.
- **subregion_Southeast Asia (0.023)**: Captures ASEAN geography. The region is strategically contested: countries with disputes often lean pro-U.S., while others (e.g., Cambodia, Laos) lean pro-China.
- **[Mid-end Manufacturing] (0.022)**: Economic structure variable identifying countries specialized in mid-end manufacturing. These economies rely heavily on global supply chains and their industrial orientation influences whether they align with U.S.- or China-led blocs.

Taken together, these results show that both *structural factors* (e.g., GDP, economic structure, region) and *political factors* (e.g., democracy, sanctions, alliances, territorial disputes) interact in shaping a country's geopolitical alignment.

Top Factors by Alignment Category

To better interpret the SHAP results, we summarize the top five most important factors for each alignment class (Pro-China, Neutral, Pro-USA). Importance is based on the average absolute SHAP value.

2.5 Contribution to Broader Understanding of Geopolitical Trends

This study contributes to the broader understanding of geopolitical alignments by introducing a transparent and replicable methodology that integrates expert knowledge with machine learning–driven inference. The framework allows researchers and policymakers to trace how structural covariates—such as military imports, alliance scores, ideological distances, and democracy indices—shape a country's orientation toward China or the United States. Because the process is both data-driven and interpretable, it provides a robust foundation that can be updated as new datasets become available, ensuring that the analysis remains relevant and adaptable to the rapidly changing global landscape. In doing so, the project equips analysts, researchers, and decision-makers with tools to assess how global alignments may evolve over the next decade and to identify regions where political shifts are most likely to occur.

Table 1: Top 5 Factors for Pro-China Alignment

Factor	**Relative Impact**	**Explanation**
v2x_libdem	High	Low democracy scores (authoritarian regimes) strongly increase pro-China alignment.
IdealPointsDistance_China	High	Countries voting similarly to China in UN assemblies lean pro-China.
IdealPointsDistance_USA	High	Large ideological distance from the US drives countries toward pro-China alignment.
Sanction Scores	Medium	Countries under Western sanctions (e.g., Russia, Iran) tend to align pro-China.
Subregion: Africa (West/Central)	Medium	Chinese infrastructure projects and loans in Africa push these regions toward pro-China.

Table 2: Top 5 Factors for Neutral Alignment

Factor	**Relative Impact**	**Explanation**
v2x_libdem	High	Neutral countries often have mid-level democracy scores, avoiding extreme alignment.
IdealPointsDistance_USA	High	Countries neither very close nor very far from US ideology tend to remain neutral.
IdealPointsDistance_China	High	Similarly, countries with moderate ideological distance from China often remain neutral.
GDP_per_capita	Medium	Mid-income countries (not too rich, not too poor) frequently adopt neutrality.
Subregion: Southeast Asia	Medium	Many Southeast Asian states balance China–US rivalry, staying neutral.

Table 3: Top 5 Factors for Pro-USA Alignment

Factor	**Relative Impact**	**Explanation**
v2x_libdem	Very High	Strong democracies align pro-USA, reflecting shared governance values.
IdealPointsDistance_USA	Very High	Countries voting closely with the US at the UN align pro-USA.
US_Alliance_Score	High	Existing formal defense treaties and alliances strengthen pro-USA alignment.
Military_Imports_USA	Medium	States importing arms from the US show stronger alignment with Washington.
US_Military_Aid	Medium	Aid recipients (e.g., Israel, Ukraine) strongly lean pro-USA.

2.6 Future Exploration

Looking forward, several avenues for extension can deepen the scope of this analysis. One promising direction is to apply the same methodology retroactively to data spanning from the year 2000 onward, allowing us to uncover historical trajectories of alignment and the emergence of patterns over the past two decades. This temporal extension would shed light on how pivotal events—such as the global financial crisis, the rise of China's Belt and Road Initiative, and Russia's geopolitical maneuvers—have influenced shifts in alignment categories. Another avenue is the incorporation of time series forecasting techniques such as SARIMA models to project alignment labels into the near future. By predicting the expected alignment of countries in 2024 and 2025, the framework would provide a forward-looking perspective on the stability or volatility of global blocs. These explorations not only enhance the explanatory power of the current model but also reinforce its utility as a predictive and policy-relevant tool.

References