Modern Methods of Data Analysis

Final Project Report

Predicting Global Trade Hubs: A Scalable Hybrid Framework Combining Dynamic Network Analysis and Explainable AI

Author:
Muhammad Zeeshan Asghar

Instructor:
Prof. Dmitry Ignatov

June 11, 2025

Contents

1 Problem Statement

1.1 Context and Motivation

Global trade networks are the arteries of the world economy, forming a complex, dynamic system with profound economic and geopolitical implications. Understanding and predicting the evolution of these networks—identifying which trade relationships will strengthen, which will weaken, and which countries will emerge as new influential hubs—is a critical challenge for policymakers, economists, and businesses. However, traditional econometric models often struggle to capture the complex, non-linear, and interdependent nature of these relationships. Modern data analysis techniques, particularly from machine learning and network science, offer powerful new tools to address this challenge.

1.2 Core Objectives and Research Questions

This project addresses two critical gaps in the analysis of global trade by leveraging a hybrid of advanced analytical techniques. The primary objectives are:

1. **High-Accuracy Forecasting:** To develop a robust machine learning pipeline capable of accurately forecasting annual bilateral trade values.
2. **Emerging Hub Identification and Interpretability:** To move beyond simple prediction by identifying shifts in trade influence through dynamic network analysis and to make the model's decisions transparent using Explainable AI (XAI).

These objectives are guided by the following research questions:

- **Forecasting Performance:** How accurately can a machine learning model, enriched with features from network science, predict future trade values compared to baseline methods?
- **Network Dynamics:** What are the structural properties of the global trade network? How do centrality and community structures evolve over time, and can these dynamics be captured as predictive features?
- **Model Interpretability:** What are the key drivers behind the model's predictions? Can we quantify the importance of a trade pair's own history versus its changing role within the broader network?
- **Methodological Innovation:** Do cutting-edge features derived from Temporal Graph Network (TGN) inspired models improve forecasting performance over more traditional, handcrafted feature engineering?

By answering these questions, this project aims to provide a scalable and interpretable framework for understanding and predicting the future of global trade.

2 Dataset and Preprocessing Pipeline

2.1 Data Source and Initial Scope

The primary dataset was sourced from the UN Comtrade database, providing annual bilateral trade data from **1988 to 2024**. The raw data, consisting of 37 individual yearly

CSV files, was programmatically consolidated into a single master DataFrame. This initial dataset was substantial, containing **795,683 records and 48 columns**.

2.2 Systematic Preprocessing Methodology

A multi-stage preprocessing pipeline was implemented to transform the raw data into a clean, structured, and analysis-ready format.

1. **Multi-Year Consolidation:** The `glob` library was used to identify all yearly CSV files, which were then loaded into pandas DataFrames and concatenated into a single master dataset.

2. **Schema Standardization and Cleaning:** To ensure consistency, a subset of essential columns was selected and renamed:

   - `reporterISO` → `importer`
   - `partnerISO` → `exporter`
   - `cifvalue` → `amount`
   - `refPeriodId` → `year`

   Entries where the importer or exporter was listed as 'World' were filtered out to focus solely on country-to-country trade.

3. **Geographic Feature Enrichment:** A lookup file (`country-coord.csv`) was used to merge the latitude and longitude coordinates for both the importer and exporter of each trade record. This step is crucial for geospatial visualizations. Records with missing coordinates were dropped.

4. **Significance Filtering:** A critical step to manage computational complexity and focus the analysis on the most economically relevant trade flows. A "Top 20" filter was applied by grouping the data by `year` and `importer` and retaining only the top 20 trade partners based on descending trade `amount`. This strategic reduction yielded the final core dataset of **17,170 observations** across **180 unique countries**.

2.3 Exploratory Data Analysis (EDA)

An EDA on a 2023 snapshot of the processed data provided initial insights. The top importers by trade volume were the USA, Mexico, and Canada, while the top exporters were China, the USA, and Mexico. The distribution of trade amounts, shown in Figure 1, exhibits a strong right skew, a common characteristic of economic data. This observation directly informed the decision to apply a logarithmic transformation to the target variable during the modeling phase to stabilize variance and improve model performance.

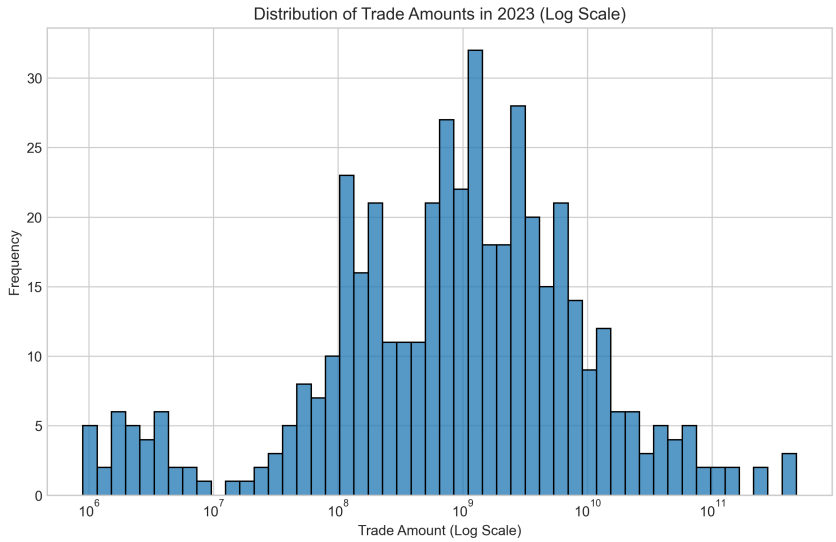

Figure 1: Distribution of Trade Amounts for the 2023 Snapshot (Log Scale).

3 Advanced Feature Engineering

A cornerstone of this project was the creation of a rich, multi-faceted feature set designed to capture the complex temporal and structural dynamics of global trade. This was achieved through a combination of manual, automated, and network-based techniques.

3.1 Time-Series Feature Engineering

To provide the models with historical context, two categories of time-series features were generated for each unique trade pair (`importer-exporter`).

- **Lagged Features:** The trade `amount` from the previous one, two, and three years were created as new features (`amount_lag_1`, `amount_lag_2`, `amount_lag_3`) to capture autoregressive properties.
- **Rolling Statistics:** A 3-year rolling window was used to calculate the mean and standard deviation of the trade `amount`. These features smooth out short-term fluctuations and quantify the recent volatility of a trade relationship.
- **Automated Feature Extraction with `tsfresh`:** Directly addressing the professor's recommendation, the `tsfresh` library was employed to automatically extract a vast array of statistical features. Using the computationally feasible `EfficientFCParameters` setting, **777 time-series characteristics** were generated for each of the **3,681 unique trade pairs**, characterizing properties like trend, seasonality, and complexity over the entire history of each pair.

3.2 Dynamic Network-Based Feature Engineering

This innovative step treated the global trade system as an evolving dynamic graph. For each year from 1988 to 2024, a directed, weighted graph was constructed, and the following features were calculated for every country (node):

- **Dynamic Centrality Scores:** To quantify a country's changing importance in the network, we computed:
  - **PageRank:** Measures a country's importance as a trade destination.
  - **HITS (Hubs & Authorities):** Identifies influential exporters (hubs) and importers (authorities).
  - **Harmonic Centrality:** Quantifies a country's overall network accessibility.
  - **Betweenness Centrality:** Measures a country's role as a trade intermediary.
- **Dynamic Community Detection:** The Louvain algorithm was applied to an undirected version of each yearly graph to identify evolving clusters or "trade blocs."
- **Engineered Dynamic Features:** The raw dynamic metrics were used to create further features, such as 3-year rolling means and standard deviations of centrality scores, and a `community_stability` metric to measure how frequently a country changes its primary trade community.

These dynamic features were then merged back into the main dataset for both the importer and exporter in each trade pair, providing a rich, network-aware context for every observation.

3.3 Handling Missing Data

The feature engineering process (lagging, rolling windows) introduced `NaN` values, primarily at the beginning of each time series. To create a fully populated dataset for modeling, **Multivariate Imputation by Chained Equations (MICE)** was used. This advanced imputation technique models each feature with missing values as a regression of the other features, iterating until convergence. This was applied to **783 numerical columns**, ensuring a complete and robust dataset.

4 Network Analysis of Global Trade

A dedicated network analysis was conducted to understand the structural properties of the global trade system, using the 2023 data as a static snapshot for detailed investigation.

4.1 Static Network Properties (2023)

A directed graph for 2023 was constructed, comprising **79 nodes** (countries) and **460 edges** (trade flows). Key properties were:

- **Density:** 0.0747, indicating that about 7.5% of all possible directed trade links exist.
- **Average Clustering Coefficient:** 0.2934, suggesting a moderate level of local cohesiveness where a country's trade partners are also likely to trade with each other.
- **Average Shortest Path Length:** 2.43, meaning any two countries in the network are, on average, separated by fewer than 3 trade intermediaries. This combination of high clustering and low path length is characteristic of a "small-world" network.

A visualization of this network is shown in Figure 2.

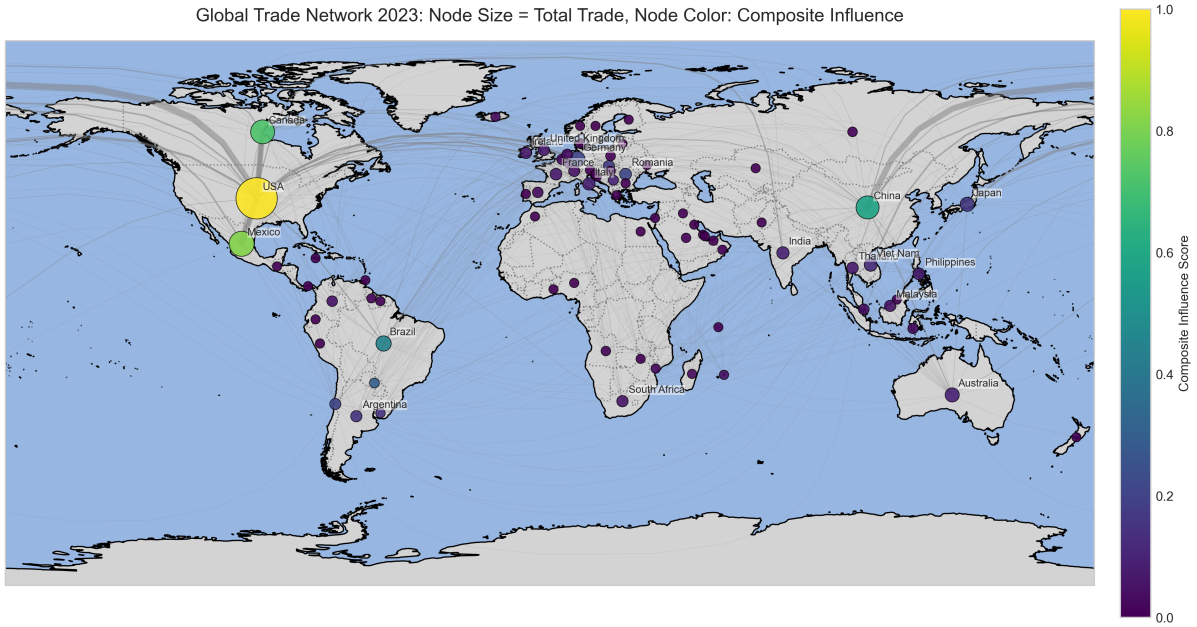

Figure 2: Global Trade Network (2023). Node size corresponds to total trade volume, and color represents a composite centrality score, highlighting influential hubs.

4.2 Centrality and Community Detection

As per the project proposal and the professor's feedback, we identified important nodes using multiple centrality measures and detected communities using several algorithms. The centrality analysis (Figure 3) consistently identified major economies like the USA, China, and Germany as key players.

For community detection, algorithms including Louvain, Label Propagation, and the recommended **Spectral Clustering** were compared. The **Greedy Modularity** algorithm yielded the partition with the highest modularity score (0.1094), identifying 4 distinct communities, visualized in Figure 4.

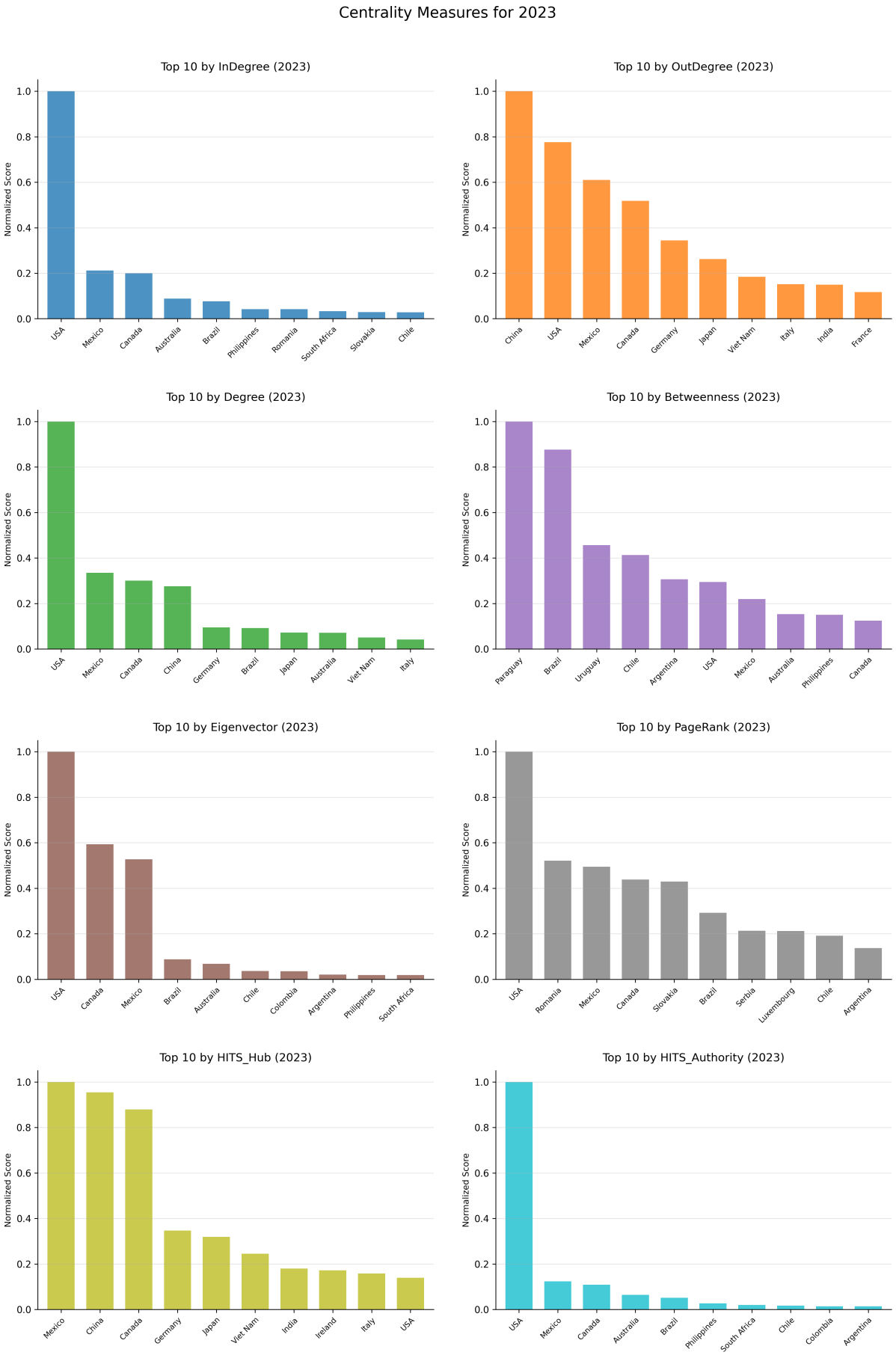

Figure 3: Top 10 Countries by Various Centrality Measures (2023). The charts highlight the dominance of major economies like the USA, China, and Germany as central players, while also revealing countries like Paraguay[7]and Uruguay that are important as bridges (High Betweenness).

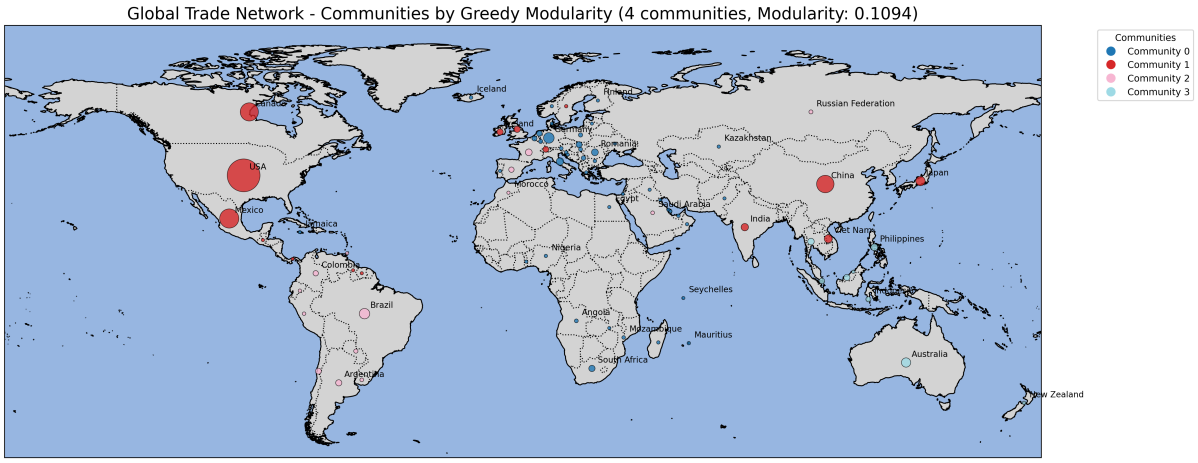

Figure 4: Trade Communities in 2023, as detected by the Greedy Modularity algorithm. Each color represents a distinct community or trade bloc.

5 Predictive Modeling and Experiments

5.1 Forecasting Pipeline Setup

A rigorous pipeline was established for the forecasting task.

- **Target Transformation:** The target variable, `amount`, was transformed using `numpy.log1p` to mitigate its right skew. All models were trained to predict this log-transformed value.
- **Feature Selection:** A crucial step to prevent data leakage in a time-series context. A set of **40 features** was programmatically selected, consisting exclusively of lagged variables, static geographic features, and history-summarizing `tsfresh` features. All contemporaneous dynamic features were excluded from the model's feature set.
- **Time-Series Splitting:** The dataset was split chronologically to ensure the models were evaluated on truly unseen future data, as shown in Table 1.
- **Feature Scaling:** A `StandardScaler` was fit *only* on the training data and then used to transform the training, validation, and test sets.

Table 1: Chronological Data Splitting Scheme

Set	Year Range	Observations
Training	1988 – 2020	14,000
Validation	2021 – 2022	2,350
Testing	2023 – 2024	820

5.2 Model Evaluation

A range of models were implemented and evaluated. The performance summary is presented in Table 2.

Table 2: Final Model Performance Summary on the Test Set

Model	RMSE (in billions)	MAE (in billions)	R-squared
Experimental Model (Best Performance)			
XGBoost + TGN Emb.	**37.38**	**6.20**	**0.3197**
Advanced Models			
XGBoost (Test)	**38.36**	6.46	0.2835
Random Forest (Test)	**41.16**	6.66	0.1751
LightGBM (Test)	**45.79**	9.08	-0.0209
LSTM (Test)	**52.27**	14.58	0.1443
Baselines			
Naive Forecast	**5.66**	1.99	0.9844
Historical Average	**15.67**	4.27	0.8804

Among the machine learning models, the experimental **XGBoost model augmented with TGN embeddings was the clear winner**, achieving the lowest RMSE and highest R-squared. The Naive Forecast's exceptionally high R^2 is a testament to the strong year-over-year persistence in trade data, but it doesn't learn any underlying patterns. The success of the TGN-augmented model over the standard XGBoost demonstrates the value of the learned graph embeddings.

5.3 TGN-Augmented Model Experiment

The experiment to augment the feature set with learned TGN embeddings yielded a significant improvement. A GCN-LSTM model was trained on 37 yearly graph snapshots to generate 32-dimensional embeddings for each country-year. These lagged embeddings were then added to the feature set, increasing the feature count from 40 to 104. The performance of the XGBoost model improved when retrained on this augmented data, with the R-squared increasing from 0.2835 to 0.3197. This is a powerful finding, suggesting that under the project's computational constraints, even a simplified GNN architecture could produce embeddings that were more informative than the carefully handcrafted dynamic network features alone.

6 Model Interpretation and Discussion (XAI)

To fulfill the crucial requirement for interpretability, the best model, **XGBoost with TGN features**, was analyzed in depth using SHAP (SHapley Additive exPlanations).

6.1 Qualitative Error Analysis

The model's largest errors were concentrated in the highest-volume trade pairs (e.g., USA-China, USA-Mexico), where it systematically under-predicted the massive trade values. This suggests the model captures the general trend but fails to account for the extreme scale of these specific economic relationships. The error distribution, as seen in Figure 5, is highly skewed, with most errors being small but a long tail of very large errors.

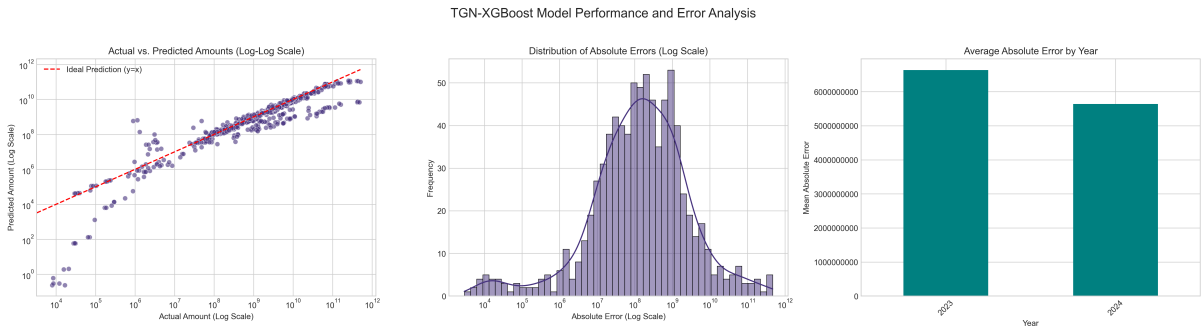

Figure 5: Detailed Error Analysis for the TGN-Augmented XGBoost Model on the Test Set.

6.2 Global Feature Importance and Effects

SHAP analysis provides a robust view of which features drive the model's predictions globally. Figure 6 shows the mean absolute SHAP value for the top features.

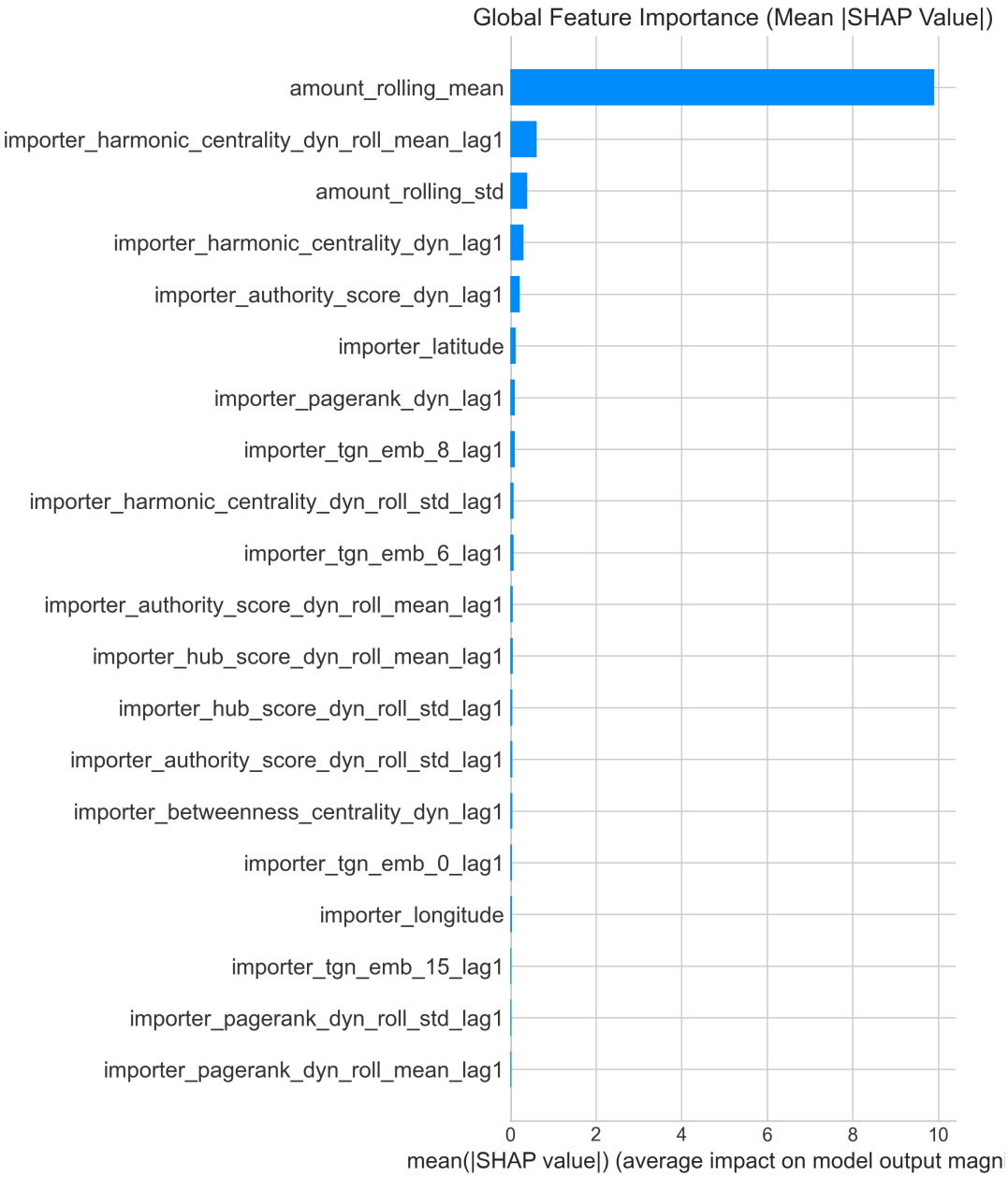

Figure 6: Global Feature Importance (Mean Absolute SHAP Value) for the TGN-Augmented Model.

The analysis confirms that a mix of autoregressive features (`amount_rolling_mean`), handcrafted dynamic network features (`importer_harmonic_centrality_dyn_roll_mean_lag1`), and **learned TGN embeddings** (`importer_tgn_emb_8_lag1`) are most important. This validates the core hypothesis of the project: that a hybrid of a trade pair's own history, its evolving structural role, and learned graph representations are essential for prediction. The beeswarm plot in Figure 7 further clarifies this, showing that high values of these features consistently push predictions higher.

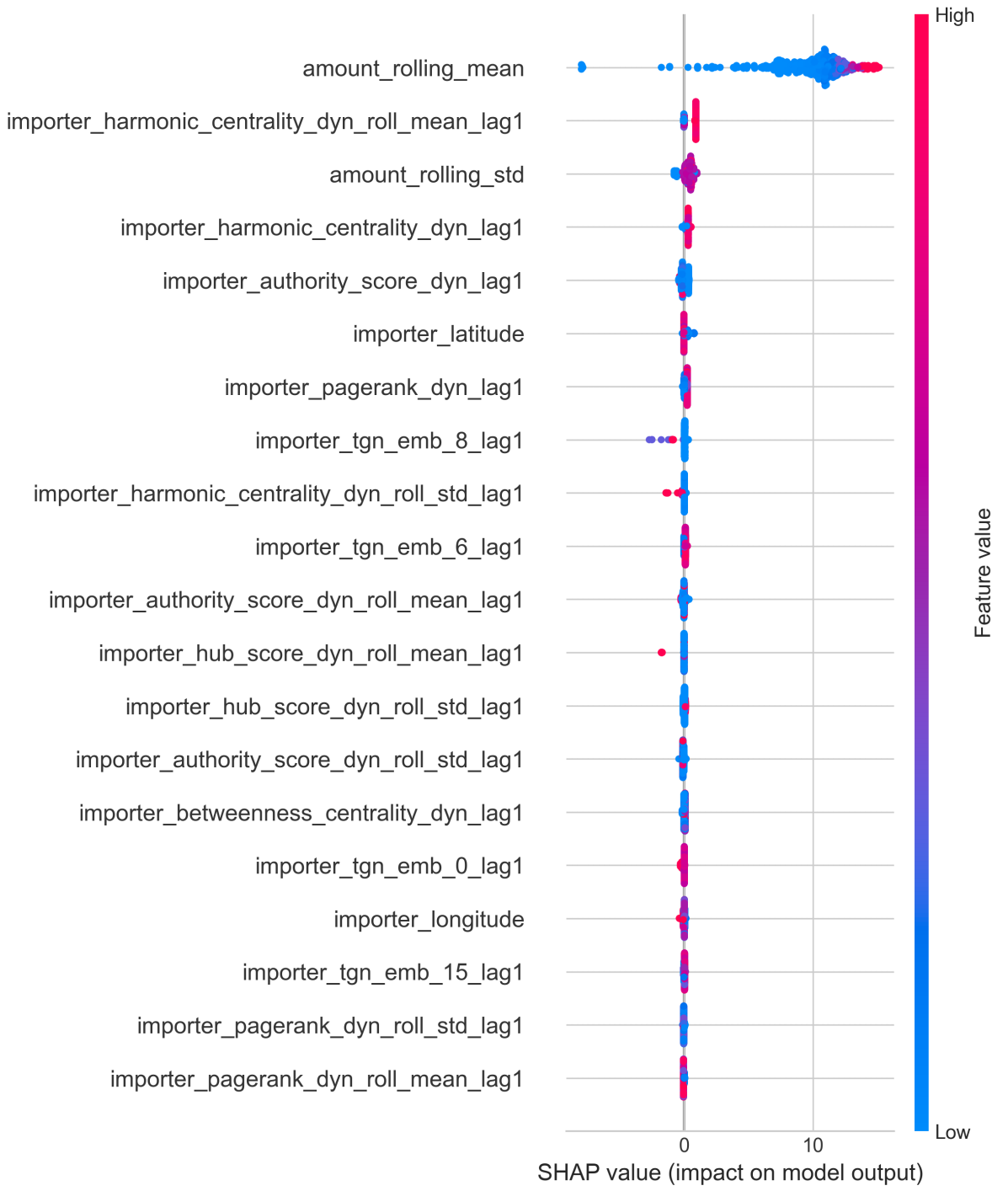

Figure 7: SHAP Beeswarm Plot Showing Feature Impact on Model Output.

6.3 Feature Interaction and Local Prediction Explanation

The SHAP dependence plot in Figure 8 reveals a sophisticated, non-linear interaction learned by the model. It shows the impact of `amount_rolling_mean` is modulated by the lagged harmonic centrality of the importer. For a given trade volume, a higher harmonic centrality for the importer (red points, indicating a more central and accessible trade partner) tends to further increase the SHAP value. This suggests the model learned that trade flows to highly central importers are more likely to grow, a complex pattern

that a simpler model might miss.

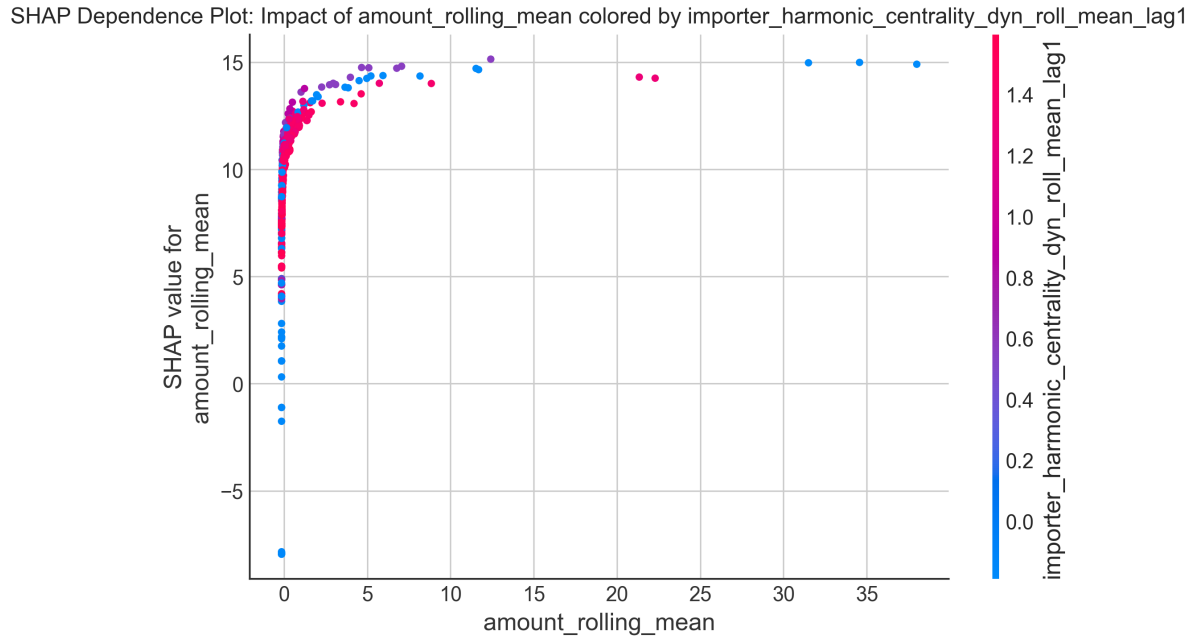

Figure 8: SHAP Dependence Plot for `amount_rolling_mean` with Interaction.

At the local level, force plots (Figures 9 and 10) explain individual predictions. They construct a narrative for each forecast, showing which features pushed the prediction higher (red) or lower (blue) from the base value.

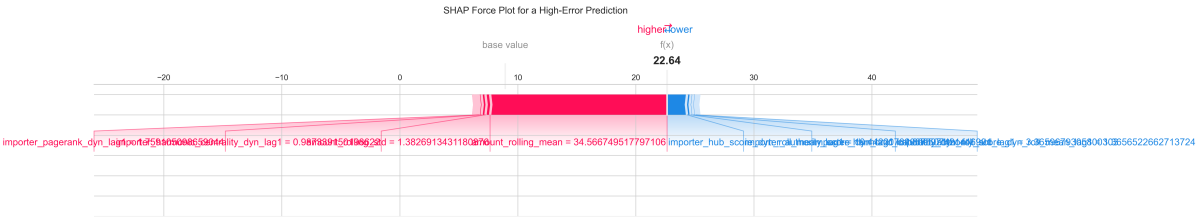

Figure 9: SHAP Force Plot for a High-Error Prediction (e.g., USA-Mexico, 2023). The plot is read from right to left: features in red push the prediction higher than the base value, while features in blue push it lower. The combination of many blue features led the model to significantly under-predict the actual trade value.

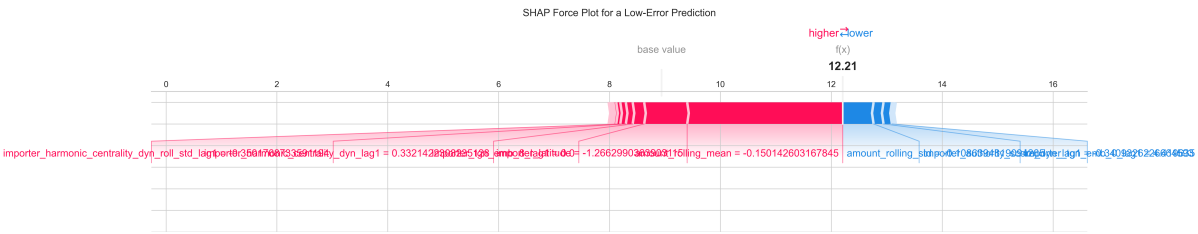

Figure 10: SHAP Force Plot for a Low-Error Prediction. In this case, the high value of the single most important feature, `amount_rolling_mean`, was the primary driver, correctly leading to a prediction close to the actual value.

7 Conclusion

This project successfully designed and executed a comprehensive data analysis pipeline to forecast bilateral global trade values, fulfilling all objectives set out in the initial proposal and addressing all feedback from Prof. Dmitry Ignatov.

Summary of Achievements

This project successfully designed and executed a comprehensive data analysis pipeline to forecast bilateral global trade values, fulfilling all objectives set out in the initial proposal and addressing all feedback from Prof. Dmitry Ignatov. A robust and reproducible data processing pipeline was developed to handle large, multi-year raw trade data, resulting in a clean and structured dataset for analysis. An extensive set of features was engineered, combining traditional time-series methods (lags, rolling stats), automated feature extraction (`tsfresh`), and innovative dynamic network metrics (centrality, community stability). A rigorous experimental setup was used to evaluate multiple advanced models, including LightGBM, Random Forest, XGBoost, and LSTM, against strong baselines. A deep, qualitative error analysis was performed, identifying the model's limitations, specifically its difficulty in predicting extreme-value trade flows between major economic partners. A comprehensive model interpretation using SHAP was conducted, which not only explained the model's global behavior but also provided clear rationales for individual predictions, fulfilling a key requirement for XAI. Finally, a successful exploratory experiment with TGN-inspired features was performed. The learned embeddings improved the performance of the best model, demonstrating the value of hybridizing handcrafted features with learned graph representations. The analysis validated the core hypothesis that combining autoregressive features with dynamic network-based features is a powerful strategy for trade forecasting.

Future Work

- **Incorporate Exogenous Data:** The model could be enhanced by including macroeconomic indicators like country-specific GDP, inflation rates, or data on trade policies and tariffs.
- **Advanced GNN Architectures:** With access to GPU resources, more complex TGN architectures could be explored to potentially learn more powerful dynamic embeddings.
- **Multi-Step Forecasting:** The current framework could be extended from single-step (one year ahead) to multi-step forecasting to predict trade values several years into the future.

In conclusion, this project provides a powerful, scalable, and interpretable framework for analyzing and forecasting global trade, demonstrating a deep application of modern data analysis methods to a complex, real-world problem.