Forecasting US Import Volume from Vietnam Using Supervised and Unsupervised Machine Learning

Le Thanh Thinh, Doan Thi Thuy An, Ong Ich Bao, Nguyen Thanh Quang, Tran Tien Thinh
Department of Artificial Intelligence
FPT University
Danang City, Vietnam

***Abstract*—This paper presents a machine learning approach to forecast the United States' import volume from Vietnam from 1994 to 2024. Leveraging a dataset of macroeconomic indicators such as the Gross Domestic Product (GDP) of both nations, Most-Favored-Nation (MFN) tariff rates, and a policy flag for the US-Vietnam Bilateral Trade Agreement (BTA), this study evaluates several supervised regression models. The models, including Linear, Polynomial, Ridge, Lasso, and Decision Tree Regression, were trained on log-transformed import data to predict future trade volumes. The results indicate that the Ridge Regression model delivered the highest predictive accuracy, achieving an R^2 of 0.8495 on the test set. Feature importance analysis revealed that Vietnam's GDP and the implementation of the BTA were the most significant drivers of import growth. Furthermore, unsupervised learning techniques, specifically Principal Component Analysis (PCA) and K-Means clustering, were employed to analyze trade pattern shifts. The analysis successfully identified two distinct clusters that align almost perfectly with the pre- and post-BTA periods, providing strong quantitative evidence of a structural shift in the bilateral trade relationship. This study demonstrates the efficacy of a hybrid machine learning approach for both forecasting and interpreting complex economic trade dynamics.**

***Index Terms*—Machine Learning, International Trade, Forecasting, Ridge Regression, PCA, K-Means Clustering, Bilateral Trade Agreement (BTA), Vietnam-US Trade.**

I. Introduction

The economic relationship between the United States and Vietnam has undergone a profound transformation over the past three decades. Following the normalization of diplomatic relations in 1995 and the enactment of the US-Vietnam Bilateral Trade Agreement (BTA) in 2001, bilateral trade has surged, making Vietnam one of the fastest-growing sources of imports for the US. Accurately forecasting this trade volume is critical for policymakers, logistics firms, and businesses involved in the transatlantic supply chain.

Traditional econometric methods, such as the gravity model of trade, have long been used to explain trade flows. However, these models may not fully capture the complex, non-linear dynamics introduced by rapid economic development and significant policy shifts. Machine learning (ML) offers a powerful alternative, providing robust tools for prediction and pattern recognition in complex datasets.

This paper applies a comprehensive ML pipeline to forecast US import volumes from Vietnam. The primary objectives are threefold:

1) To develop and evaluate a suite of supervised regression models to predict future import volumes based on key economic and policy variables.
2) To utilize unsupervised learning methods to identify and analyze structural shifts in trade patterns, particularly in relation to the BTA.
3) To determine the key drivers of import volume through feature importance analysis, thereby providing economic context to the model's predictions.

By integrating predictive modeling with unsupervised pattern discovery, this study aims to provide a data-driven validation of the BTA's impact and a reliable model for future trade analysis.

II. Related Work

The analysis of bilateral trade flows has a rich history in economics, dominated by the **gravity model**, which posits that trade is proportional to the economic size of the two countries and inversely proportional to the distance between them [1]. While foundational, traditional gravity models often rely on linear assumptions. Recent studies have begun incorporating machine learning techniques to enhance the predictive power of these models. For instance, research has shown that non-linear models like Random Forests and Gradient Boosting can outperform traditional OLS estimators in trade forecasting [2], [3].

The economic impact of trade agreements like the BTA is another extensively studied area. Many analyses confirm that such agreements significantly boost trade by reducing tariff and non-tariff barriers [4], [5]. This study builds on this literature by using an unsupervised learning approach (K-Means clustering) to provide an independent, data-driven validation of the BTA's structural impact, a method less common in traditional economic analyses.

The application of machine learning in macroeconomics is a rapidly growing field. Techniques like PCA are used to distill complex economic indicators into meaningful indices [6], while regression models like Ridge and Lasso are valued for their ability to handle multicollinearity and prevent overfitting in small datasets [7], [8]—a common challenge with annual macroeconomic data. This project directly applies these established ML methodologies to the specific and economically significant case of US-Vietnam trade [9], [10].

III. Methodology and Data

A. Dataset Description

The study utilizes a time-series dataset spanning 31 years, from 1994 to 2024. The data was compiled from reputable sources including the World Bank and the World Trade Organization. The primary variables are:

- **US_import**: The nominal value of US imports from Vietnam (in USD).
- **Ln(US_import)**: The natural logarithm of the import value, used as the target variable to stabilize variance and model exponential growth as a linear trend.
- **GDP_US_2021 & GDP_VN_2021**: Per capita GDP for the US and Vietnam, adjusted to 2021 price levels.
- **Ln(GDP_US) & Ln(GDP_VN)**: Log-transformed GDP values.
- **MFN_simple**: The simple mean of Most-Favored-Nation tariff rates applied by the US to Vietnamese goods.
- **Distance**: The geographical distance between the two countries in kilometers.
- **BTA**: A binary variable indicating the status of the Bilateral Trade Agreement ('0' for before 2002, '1' for 2002 and after).

B. Data Preprocessing and Feature Engineering

The dataset was first inspected for completeness and quality. No missing values or duplicate entries were found. To enhance the model's predictive power, two additional features were engineered:

1) **GDP_Ratio**: Calculated as `GDP_US_2021 / GDP_VN_2021`, this feature captures the relative economic size of the two nations, a core component of gravity models.
2) **Tariff_Change**: The year-over-year difference in the `MFN_simple` tariff rate, designed to capture the impact of policy adjustments on trade flows.

C. Supervised Learning Models

The core of the forecasting task was addressed using five supervised regression models:

- **Linear Regression**: A baseline model to establish a linear relationship.
- **Polynomial Regression**: To capture potential non-linear trends.
- **Ridge (L2 Regularization)**: To prevent overfitting by penalizing large coefficient values, making it suitable for datasets with multicollinearity.
- **Lasso (L1 Regularization)**: Similar to Ridge but can shrink coefficients to zero, effectively performing feature selection.
- **Decision Tree Regression**: A non-linear model that partitions the data based on feature values.

D. Unsupervised Learning Models

To identify structural patterns in the data, two unsupervised techniques were used:

- **Principal Component Analysis (PCA)**: Used to reduce the dimensionality of the economic features (`Ln(GDP_US)`, `Ln(GDP_VN)`, `MFN_simple`, `GDP_Ratio`) into two principal components for visualization.
- **K-Means Clustering**: Applied to the PCA-transformed data to group the years into distinct clusters based on their economic characteristics. The **Elbow Method** was used to confirm that two clusters (k=2) was the optimal number.

E. Model Evaluation

A chronological train-test split was implemented. Data from **1994 to 2019** served as the training set, while data from **2020 to 2024** was used as the unseen test set for final evaluation. For hyperparameter tuning on the training set, 'GridSearchCV' was combined with a 'TimeSeriesSplit' cross-validation strategy to respect the temporal nature of the data. Model performance was evaluated using three standard metrics: R-squared (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

IV. Results and Analysis

A. Exploratory Data Analysis

The exploratory data analysis, conducted on a dataset of 31 annual observations from 1994 to 2024, reveals several key insights. The time-series plot of `Ln(US_import)`, shown in Figure 1, displays a clear and sustained upward trend over the period. A notable change in this trend appears to coincide with the implementation of the US-Vietnam Bilateral Trade Agreement (BTA) around 2002.

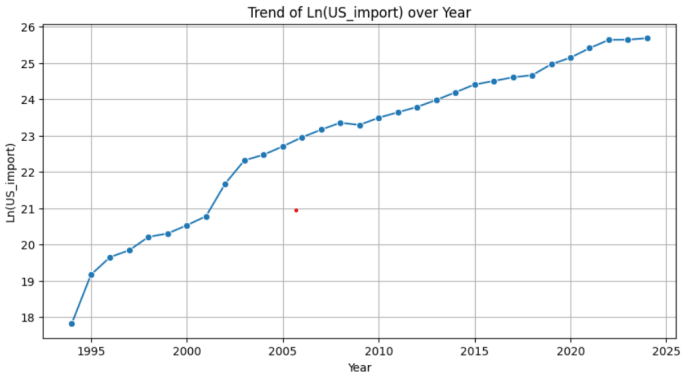

Fig. 1. Trend of the natural log of US imports from Vietnam over time (1994-2024).

A comparison of the pre-BTA (1994-2001) and post-BTA (2002-2024) periods using boxplots (Figure 2) demonstrates a distinct upward shift in the log of US import values and a downward shift in the MFN simple tariff rates after the agreement took effect. A two-sample t-test confirms that the increase in `Ln(US_import)` is statistically significant, yielding a p-value of 0.0000.

The correlation heatmap in Figure 3 highlights strong relationships between the variables. `Ln(US_import)` has a very strong positive correlation with both the log of US GDP,

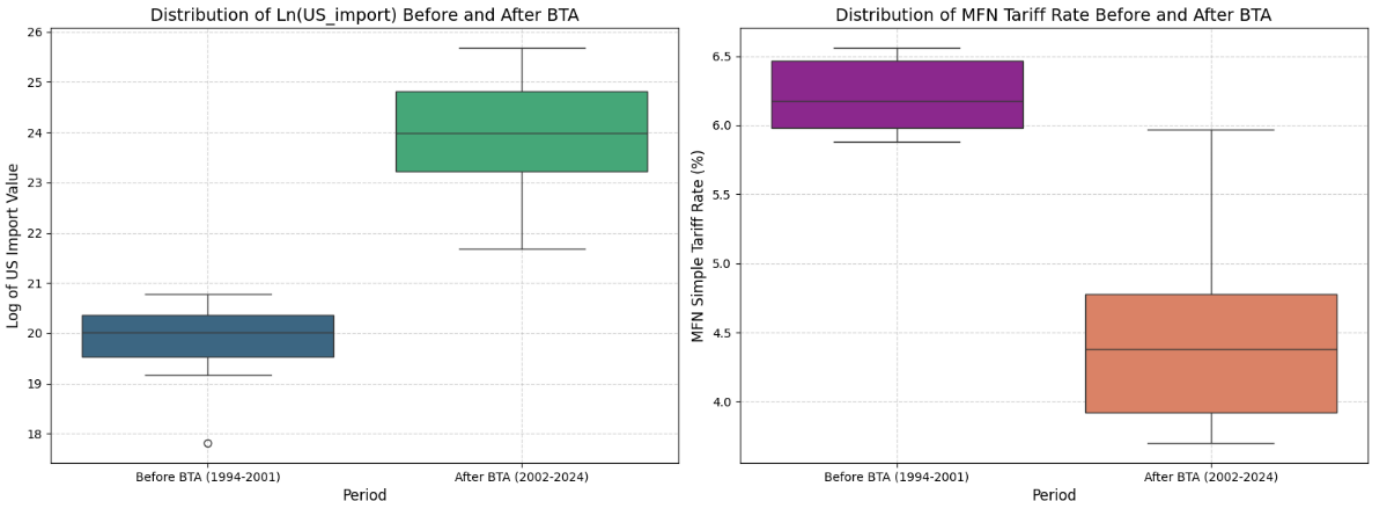

Fig. 2. Distribution of Ln(US_import) and MFN Tariff Rate Before and After the BTA.

`Ln(GDP_US)` ($r = 0.97$), and the log of Vietnam's GDP, `Ln(GDP_VN)` ($r = 0.98$). Conversely, it shows a strong negative correlation with the MFN tariff rate, `MFN_simple` ($r = -0.92$). These relationships suggest that economic growth in both nations and trade liberalization are significant factors associated with the rise in US imports from Vietnam.

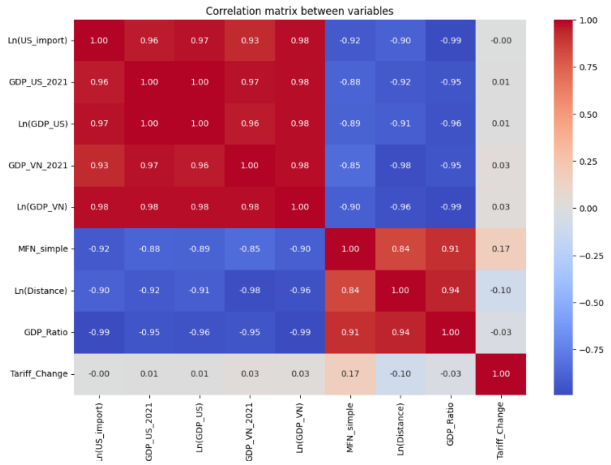

Fig. 3. Correlation heatmap of the primary variables in the model.

B. Supervised Model Performance

After hyperparameter tuning using a grid search methodology, the models were evaluated on the test set (2020-2024). The optimal hyperparameters found for each model are shown in Table I. The final performance results on the test set are summarized in Table II.

The **Ridge Regression model demonstrated superior performance**, achieving a Test R-squared of **0.8495** and the lowest errors (MAE: 0.0745, RMSE: 0.0788). This indicates it explained nearly 85% of the variance in the unseen test data. The Lasso model performed modestly, while the basic Linear, Decision Tree, and Polynomial models performed extremely poorly. Their negative R^2 values suggest severe overfitting on the training data. The $L2$ regularization in the Ridge model proved crucial for generalizing well on this small, collinear time-series dataset.

TABLE I
OPTIMAL HYPERPARAMETERS FOUND VIA GRID SEARCH

Model	Best Hyperparameters
Ridge Regression	`'alpha': 10`
Lasso Regression	`'alpha': 0.1`
Linear Regression	N/A
Decision Tree	`'max_depth': 4`
Polynomial Regression	`'degree': 2`

TABLE II
SUPERVISED MODEL PERFORMANCE ON TEST SET (2020-2024)

Model	Test MAE	Test RMSE	Test R-squared
Ridge Regression	**0.0745**	**0.0788**	**0.8495**
Lasso Regression	0.1401	0.1588	0.3881
Linear Regression	0.6009	0.6062	-7.9174
Decision Tree	0.6231	0.6306	-8.6483
Polynomial Regression	1.3878	1.6467	-64.7957

The strong performance of the final model is visualized in Figure 4, which plots its predictions against the actual data.

C. Feature Importance Analysis

To assess the impact of each variable on US imports, the standardized coefficients from the final Ridge Regression model were analyzed. These coefficients, which serve as a measure of feature importance, are visualized in Figure 5. They reveal the direction and relative strength of each feature's relationship with the target variable.

The key insights from the model's coefficients are as follows:

- **Positive Drivers of Trade**: The four most influential positive features are `Ln(GDP_US)`, the `BTA` dummy

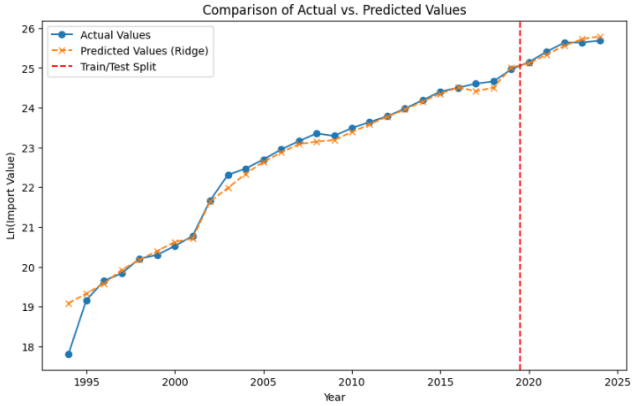

Fig. 4. Comparison of actual log-import values against the predictions from the best-performing Ridge model. The model tracks the trend well, even in the unseen test period.

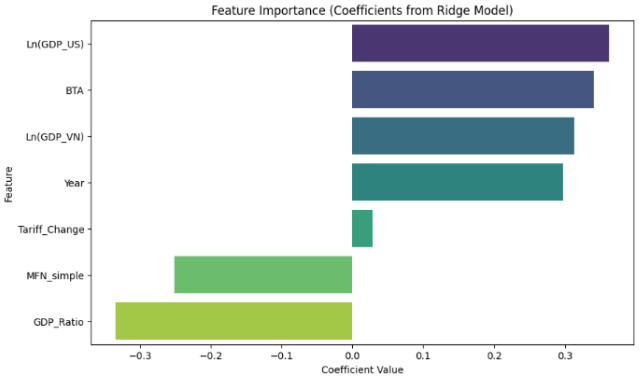

Fig. 5. Feature Importance from Ridge Model Coefficients.

variable, `Ln(GDP_VN)`, and the `Year`. This indicates that the growth in US imports from Vietnam is most strongly driven by the economic growth of the US (the destination market), the trade liberalization effects of the Bilateral Trade Agreement, and the expanding productive capacity of Vietnam's economy.

- **Negative Drivers of Trade**: The features with the largest negative coefficients are `GDP_Ratio` and `MFN_simple`.
  - The negative coefficient for `MFN_simple` is consistent with economic theory: higher tariffs are associated with lower import volumes.
  - The negative coefficient for `GDP_Ratio` (defined as `GDP_US / GDP_VN`) indicates that a larger disparity between the two economies is associated with lower import volumes. As illustrated in Figure 6, this ratio has steadily declined over time, coinciding with a consistent rise in `Ln(US_import)`. The model captures this inverse empirical relationship by assigning a negative weight to `GDP_Ratio`.

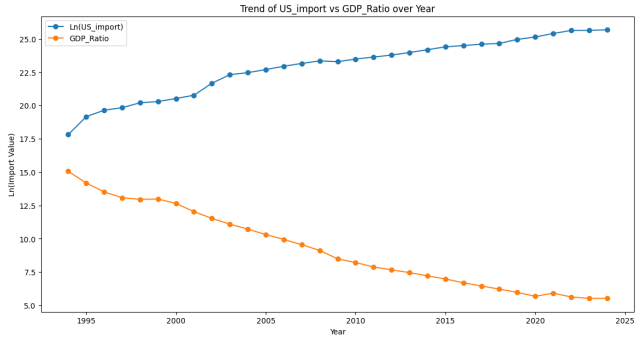

Fig. 6. Trend of US_import vs GDP_Ratio over time (1994-2024)

D. Unsupervised Analysis of Trade Patterns

To investigate the underlying structure of the trade relationship over time, an unsupervised learning approach was employed, combining Principal Component Analysis (PCA) and K-Means clustering.

First, PCA was applied to the four key economic features (`Ln(GDP_US)`, `Ln(GDP_VN)`, `MFN_simple`, `GDP_Ratio`) to reduce dimensionality. As shown in Figure 7, the first two principal components (PCs) successfully captured a remarkable **98.9%** of the total variance in the dataset (PC1: 95.4%, PC2: 3.5%). The component loadings, detailed in Figure 8, reveal the economic meaning of these components:

- **PC1** can be interpreted as an index of **"Economic Growth and Integration."** It is strongly and positively correlated with both US and Vietnamese GDP, and strongly negatively correlated with tariff rates and the GDP ratio.
- **PC2** is almost exclusively driven by the **MFN tariff rate**, effectively isolating the impact of tariff structure from the general growth trend.

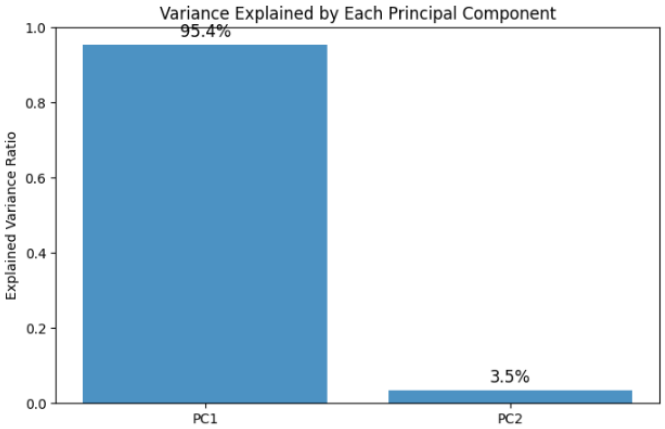

Fig. 7. Variance explained by each principal component. PC1 alone accounts for over 95% of the variance.

Next, K-Means clustering was applied to the PCA-transformed data. The Elbow Method (Figure 9) confirmed that the optimal number of clusters for this dataset is **k=2**.

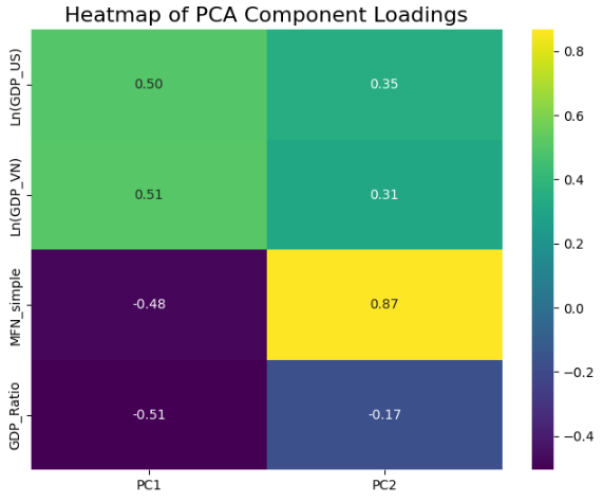

Fig. 8. Heatmap of loadings for the first two principal components.

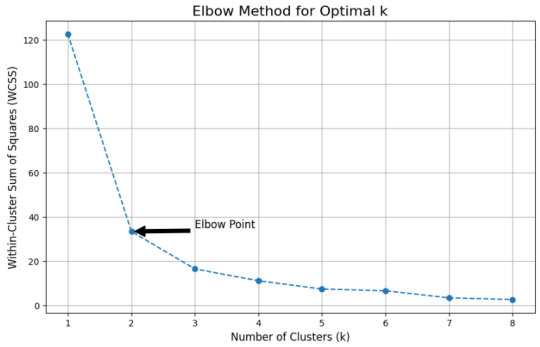

Fig. 9. The Elbow Method showing a distinct "elbow" at k=2, which indicates the optimal number of clusters.

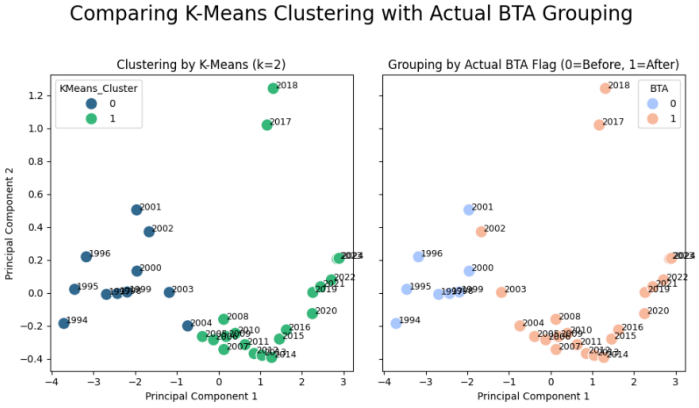

Fig. 10. A side-by-side comparison showing the K-Means clusters (left) align almost perfectly with the actual pre/post-BTA periods (right), validating the BTA's structural impact.

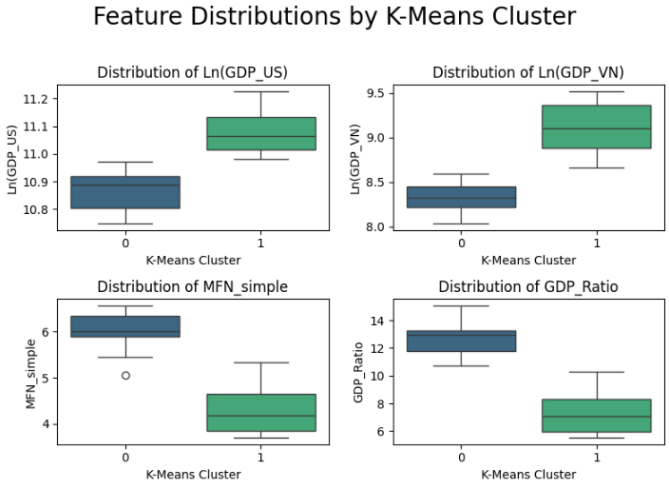

Fig. 11. Distribution of key economic features for each of the two clusters identified by the K-Means algorithm.

The results of the clustering, visualized in Figure 10, are striking. The two clusters identified by the algorithm correspond almost perfectly to the years **before the BTA (Cluster 0)** and **after the BTA (Cluster 1)**. This provides strong, data-driven evidence that the BTA marked a fundamental structural shift in the US-Vietnam trade relationship.

Finally, Figure 11 characterizes the distinct economic profiles of these two clusters. **Cluster 0 (the pre-BTA era)** is defined by lower GDPs, higher tariff rates, and a higher GDP ratio. In contrast, **Cluster 1 (the post-BTA era)** is defined by significantly higher GDPs for both nations, lower tariffs, and a lower GDP ratio. This unsupervised analysis independently validates the BTA as the pivotal event that re-shaped the economic landscape between the two countries.

V. Discussion

The findings of this study highlight the power of machine learning in economic forecasting and analysis. The superior performance of the Ridge regression model underscores the importance of regularization techniques when dealing with macroeconomic time-series data, which is often limited in size and prone to multicollinearity. The model's high accuracy on the test set suggests its potential for reliable near-term forecasting.

The feature importance results provide valuable economic insights, confirming that Vietnam's internal economic development (`Ln(GDP_VN)`) is the most critical driver of its export capacity to the US. This suggests that the US-Vietnam trade relationship is not merely a result of policy but is fundamentally supported by Vietnam's robust economic growth.

Perhaps the most compelling finding is the validation of the BTA's impact through unsupervised learning. K-Means clustering, without any prior knowledge of the BTA's existence, successfully segregated the data into pre- and post-BTA eras. This demonstrates that the BTA was not just a minor policy tweak but a transformative event that fundamentally reshaped the trade dynamics between the two nations.

Limitations of this study include the small dataset size (31 annual data points), which restricts the use of more complex models. The feature set, while informative, could be expanded to include other variables like exchange rates, foreign direct

investment (FDI), and global economic shocks to potentially improve accuracy further.

VI. Conclusion

This paper successfully applied a combination of supervised and unsupervised machine learning techniques to forecast and analyze US import volumes from Vietnam. The Ridge Regression model proved to be a robust and accurate forecasting tool. The analysis identified Vietnam's GDP growth and the US-Vietnam BTA as the most significant factors driving trade. Unsupervised clustering provided compelling evidence of a structural break in the trade relationship coinciding with the BTA. This dual approach not only yields a predictive model but also offers a deeper, data-driven understanding of the economic forces at play. Future work could involve incorporating a wider range of economic indicators and applying more advanced time-series models as more data becomes available.

References

[1] J. Tinbergen, *Shaping the World Economy*. New York: The Twentieth Century Fund, 1962.

[2] J. Anderson, "The gravity model," *Annual Review of Economics*, vol. 3, no. 1, pp. 133–160, 2011.

[3] J. Bergstrand, "The gravity equation in international trade: Some microeconomic foundations and empirical evidence," *The Review of Economics and Statistics*, vol. 67, no. 3, pp. 474–481, 1985.

[4] P. T. H. Anh, "The impact of the vietnam-us bilateral trade agreement on vietnam's trade and investment," *ASEAN Economic Bulletin*, vol. 22, no. 3, pp. 324–342, 2005.

[5] M. J. Ferrantino and M. T. N. Tran, "The US-Vietnam bilateral trade agreement: A quantitative assessment," *Journal of Policy Modeling*, vol. 28, no. 8, pp. 917–935, 2006.

[6] J. H. Stock and M. W. Watson, "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1167–1179, 2002.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[8] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[9] S. L. Pintea, "Machine learning techniques for international trade analysis," *Procedia Economics and Finance*, vol. 32, pp. 696–703, 2015.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.