Trade Flow Estimation Between Global Economies Using Machine Learning Techniques

[1]PRATEEK RAO, [2] POOSHPAL BAHETI, [3]RAHUL RAMESH, [4]ANAND M S

[1]Department of Computer Science, PES University, Bangalore, India
[2] Department of Computer Science, PES University, Bangalore, India
[3] Department of Computer Science, PES University, Bangalore, India
[4] Adjunct Professor, PES University, Bangalore, India

Email: [1]xrprateek@gmail.com, [2]pooshpalbaheti@gmail.com, [3]ynotrr@gmail.com, [4]anandms@yahoo.com

Abstract: Bilateral trade flow between global economies has been a critical economic indicator for economists and policymakers owing to its potential to significantly influence international trade sanctions and policies, which profoundly impact international relations.

This deep-rooted global interdependence has been historically simulated by the international economy, which engenders substantial mutual benefits for participating companies. The analysis of various economic trends and the generation of accurate future predictions have become indispensable pursuits since the inception of predictive machine learning models. By meticulously examining historical economic data and constructing relevant fine-tuned machine learning models, it is possible to attempt to predict future events and capital flow between countries. The reliable estimation of bilateral flow is of paramount importance, and the use of machine learning techniques for economic forecasting, leveraging the unreasonable effectiveness of data, in lieu of traditional statistical methods, can help in achieving exceptional predictions.

In this study, we endeavor to enhance traditional statistical methods by experimenting with several time-agnostic machine learning models.

Index terms: Artificial intelligence, Data mining, Business analytics, neural networks, gradient boosting, bilateral trade flow

I. INTRODUCTION

In the context of global trade, policymakers collaborate with economists to analyze a variety of economic indicators in order to design strategies, policies, reforms, and roadmaps that are crucial to economic decision-making. Bilateral flow of trade has historically been a significant measure of interdependence between nations, reflecting the absolute exchange of goods and services between the participating countries. As such, it can exert a considerable influence on international trade policies. By studying the bilateral flow of goods and services between global economies, one can draw numerous inferences about the commercial relations between them. In turn, this provides valuable insights into stability, development and growth. These inferences can inform policymakers as they develop effective agendas to achieve their fiscal goals.

A high and continuously increasing inter-country flow rates are indicative of a developing relationship between the participating entities. From the perspective of developing nations, augmenting import activity can pave the way for job creation and enhance competitive markets. Upswings in export activity have been demonstrated to amplify GDP, thereby opening up opportunities for further investments and development. Forecasting bilateral flow stands to furnish policy architects with invaluable insights, enabling perceptive revisions toward maximum efficacy.

Models explored in this paper include Linear Regression, Deep Neural Networks, Support Vector Regression, and gradient boosting models such as LightGBM [9]. We make an attempt at Time-series modeling too, using models such as ARIMA and ARIMAX.

These models were used to estimate the numerical value of the Bilateral Trade Flow. The input to these models was a set of numerical and categorical features spanning over other economic indicators, geographical location, and even alliances between countries. A comprehensive feature selection process will be highlighted in further sections.

II. RELATED WORK

We first focus on the revolutionary work of Walter Isard in 1954, [1] which gave rise to the gravity model of international trade. This particular model estimates the bilateral trade flow between two countries by

considering their economic sizes and the distance between them. In the case of two countries, denoted as i and j, the model is formulated as follows:

$$F_{ij} = G.\frac{M_i^{\beta_1} M_j^{\beta_2}}{D_{ij}^{\beta_3}}$$

Here, the letter G represents a constant, while F refers to the trade flow, D denotes the distance, and M represents the economic characteristics being evaluated for the countries in question. The gravity model has been used throughout history to predict some important events, and it provides us with a suitable baseline for evaluation. Independent calculations have suggested that the gravity model has an adjusted R^2 value between 0.5 and 0.6 . Isaac Wohl and Jim Kennedy [2] used artificial neural networks to analyze international trade data. Making use of data from between United States and its main trade partners, they trained the model on the set from 1986 to 2006, and made predictions from 2007 to 2016.

Jingwen Sun et al. [3] analyzed Bilateral Trade Flow, but with a different target variable in mind. They used yearly import/export data from 217 countries for the 1960-2017 period, and predicted the GDP using several statistical factors. They concluded that RBF Regressor was the better model. S Circlaeys et al. [4] tested several machine learning models to predict pattern of bilateral trade flows. They used a variety of linear regression and neural network models to train on economic and geographic factors. Due to the ability of neural networks in capturing non-linear interaction among features, it performed better than any regression-based model. Machine learning algorithms were employed by Feras Batarseh et al. [5] to identify the most significant economic variables that impact trade for particular commodities. Subsequently, these features were utilized to train models for predicting trade patterns. This type of feature engineering was found to be more accurate than statistical methods in predicting future trend patterns. In their attempt to develop a novel approach to forecasting, Jin-kyu Jung et al. [6] utilized several ML techniques. Specifically, they applied Elastic Net, Super Learner, and Recurrent Neural Network algorithms to analyze macroeconomic data from seven advanced and emerging countries. The outcome of their study revealed that these machine learning models outperformed traditional statistical approaches and have significant potential for enhancing economic forecasting.

III. DATASET

We make use of the "TradHist" dataset by CEPII [7], which provides around 1.9 million bilateral trade observations for 1827-2014, for 200+ different countries and regions. It has both geographical dimension and time dimension. Due to the huge time-frame of this data, there are a lot of missing values due to difference in data collection techniques. We employ a lot of data cleaning and preprocessing steps in order to prepare the data accordingly for our use cases.

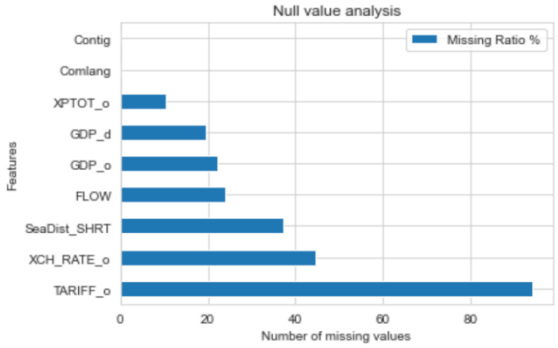

Fig. 1. Null Value Analysis

A. *Data Cleaning*

In order to maintain recency of the data, we only work on data post 2000. In Fig. 1., we can see the null value ratio in most of the important features. These null values can cause a lot of issues during data analysis and training. Hence, we start off by removing data with any null values in our categorical feature columns. After experimenting with several imputation types, we also remove any numeric column with zero values within our selected time-frame, as this can potentially introduce unnecessary noise into our training data. Imputation would not be effective in filling up missing data, as the dataset hasn't been adjusted for inflation and PPP. Detection of outliers was conducted and treatment was done on the dataset likewise. We devise a suitable threshold of 500 Pound Sterling to get rid of trade flow within this limit, as it is insignificant and can cause bias in predictions.

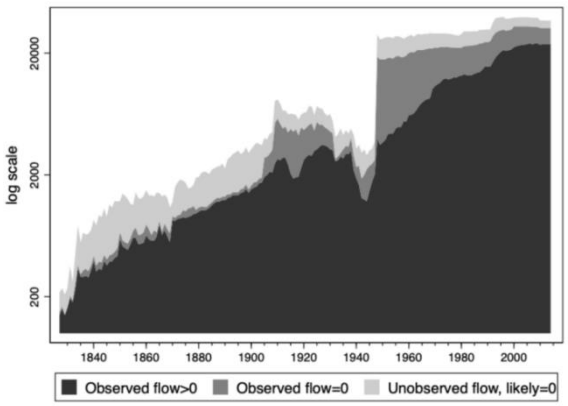

Fig. 2. Number of bilateral trade observations

B. *Exploratory Data Analysis*

In this section, we will explore the data briefly in order to understand trends and uncover some patterns. First, let us examine the number of bilateral trade observations for different values of FLOW, which serves as our target variable.

Fig. 2., which was presented in [7], depicts a noticeable increase in the number of observations per year. This trend can be explained by three factors. Firstly, it is a result of the growing number of countries, which automatically leads to an increase in the potential for international trade flows. Secondly, the increase in observed flows reflects the growing number of country pairs that are actively involved in bilateral trade. Lastly, the rise in availability of primary data sources in recent times could also be a contributing factor, attributed to challenges in accessing historical statistics and conservation issues for older records.

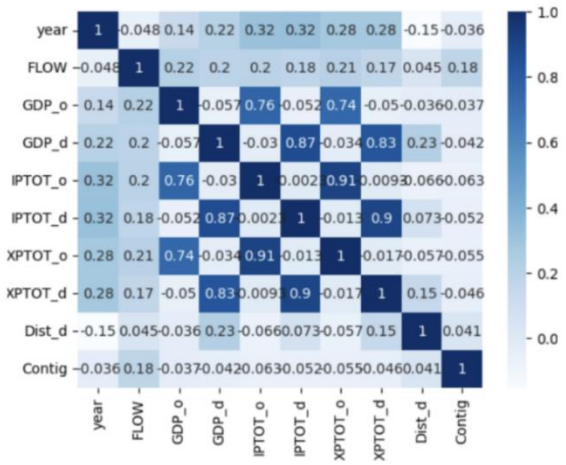

Fig. 3. Correlation Heatmap

On performing Correlation Analysis, we observe that there are some strong linear correlations between our data, as outlined in Fig. 3. We can observe that there is a significant linear positive correlation between the GDP of the origin country, the GDP of the destination country, and the total exports of the origin country, with respect to the target variable FLOW. We will take these relationships into consideration as we move to feature selection.

C. Feature Engineering

In order to further explore relationships between data, we run a preliminary round of LightGBM[9] on our prepared data, and plot the feature importance in predictions, as outlined in Fig. 4.

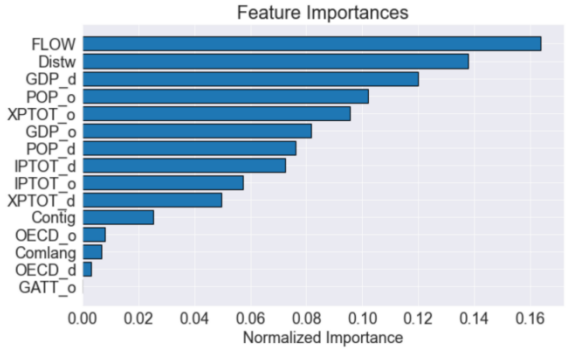

Fig. 4. Normalized feature importance

On further inspection, we don't find an issue of spurious correlation here. Thus, we will incorporate the top 12 features into our further working models, and use these as exogenous features to model our predictions.

Although we realize that we have passed up an opportunity to perform advanced feature engineering by construction of our own features, we aim to populate the feature space with similar variables as used in the Gravity Model for an accurate comparison. As a byproduct, we aim to prove the unreasonable effectiveness of big data [8] by using the basic available features only.

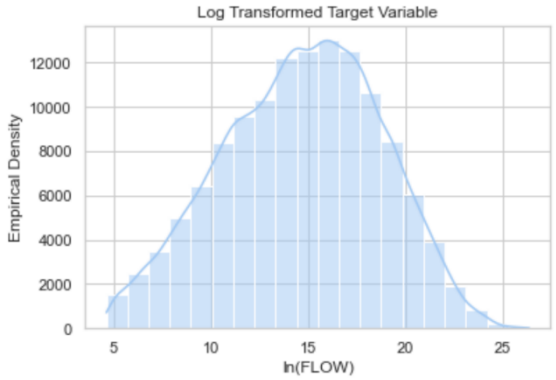

Fig. 5. Log Transformed Distribution

D. Data Transformation

On plotting the histogram for the target variable, we find that the data is highly skewed. In order to reduce the skewness, we use a log transformation on all the numeric attributes of the dataset. In Fig. 5., we can observe that the log transformation is useful in normalizing the distribution of the target variable. This turns out to be valuable in making data patterns more interpretable, and also adhering to the assumptions of inferential statistics.

E. Time Series Transformation

All of the above data transformation methods are used to prepare the data for time-agnostic modelling. However, we also explore time series modeling for this data. Here, we take a pair of countries as our tuple, which acts as the unique identifier. We cut the data down to between 2001 and 2014 only, as we want to explore recent trends. Log transformations are applied to normalize the data.

IV. PROPOSED METHODS

In order to select the best model, we have tried and tested multiple models. These models have been highlighted in the following section. Although most of the models make use of time-agnostic features, we have attempted to transform the given data into time series data and build the models with time-lagged features. The premise of this paper is to highlight how time-agnostic models can generalize and predict a

target that would generally be considered as time series.

We provide some insight into our model selection process in this section. Since we are predicting FLOW, which happens to be a continuous variable, we will focus on implementing regression-based models over classification. On data exploration, it is obvious that most of the interactions between features are non-linear, and hence we will be focusing on models that can best capture non-linear correlations. We have been careful to avoid overfitting, and each model has been tested against completely unseen data. Before training of the model, the data has been split into the usual Train-test validation. We decided not to use K-fold Cross Validation as we have ample number of data instances to test against. The first model to be tried was normal Linear Regression, which accounts for the best fit line. The model regresses the bilateral trade flow data as such:

$$\text{FLOW} = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

However, as expected, the model performed quite miserably as it wasn't able to capture non-linear interactions. We proceeded to try out Log Transformed Linear Regression, with the entire numerical dataset being transformed by Logarithmic function to distribute the data normally.

$$\text{FLOW} = \beta_1 ln(\text{GDP}_o) + \beta_2 ln(\text{GDP}_d) + ... + \beta_n ln(\text{Distw}) + \epsilon$$

This model provided a slightly better result, with a R2 value of 0.65, which provides a good baseline to compare against traditional economic methods. The third model tried was a Support Vector Machine, with logarithmic features. Here, we employ the RBF kernel, to try to model non-linear features.

$$k(x, x') = \exp\left(\frac{-(x - x')^2}{2\gamma^2}\right)$$

We also implement a fully feed-forward artificial neural network, without hyperparameter tuning. The model architecture is shown in Fig. 6.

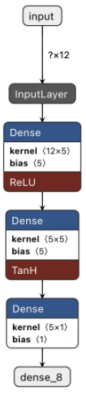

Fig. 6. ANN Model architecture

We make use of logarithmic features and aptly scaled values to train the neural network. We make use of both tanh and ReLU activation functions. The last time agnostic model implemented was LightGBM[9], which is an ensemble gradient boosted model. It makes use of decision trees to regress the predictive variable. The number of estimators chosen was 10000.

We now move toward time series modelling, by making use of the data transformed specifically for time series. We can only utilize the data between two countries. The first time series model implemented was ARIMA(2,1,2). The order was calculated by plotting the ACF and PACF plots. We further implemented ARIMAX(2,1,2) leveraging exogenous variables. We implemented Augmented Dickey-fuller test and discovered that we can obtain a stationary series through first order differentiations. This result was utilized in deciding the differencing factor for ARIMAX.

V. RESULTS

A. Comparison Metric

In order to evaluate and compare the predictive abilities of the models, the coefficient of determination (R^2) was utilized as the metric of comparison. The formula for R^2 is presented below:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Models	R^2
Linear Regression	0.14
ARIMA(2,1,2)	0.38
ARIMAX(2,1,2)	0.61
Log Transformed Linear Regression	0.65
Support Vector Regression	0.67
Artificial Neural Network	0.70
LightGBM	0.94

B. Validation Technique

To eliminate any sort of bias, all the models have been validated in the same format, taking the coefficient of determination into consideration. We have a *30%* holdout of the total data for testing and validation. As mentioned earlier, k-fold cross validation was not considered as we had enough data.

C. Inferences

Here, we can observe that the LightGBM Model is performing the best, with an R^2 value of 0.94, which is state-of-the-art result for this task. We infer this is because Gradient Boosting models are exceptional with tabular data, and making use of the feature analysis, we are able to capture most of the non-linear interactions between the independent variables.

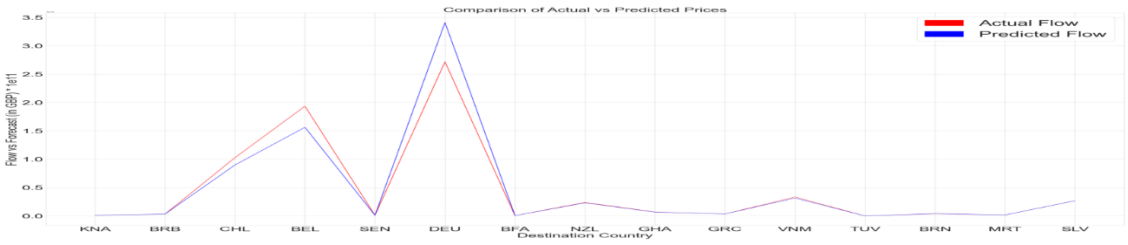

Fig. 7. Predicted Flow with Source Country USA for 2013

Although the Artificial Neural Network model seems to be underperforming, there is a lot of scope for hyperparameter tuning, to make the model perform better.

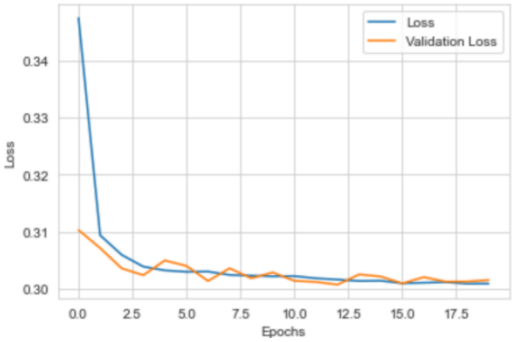

Fig. 8. ANN Training

Fig. 7. shows the predicted vs. actual values for Flow, with USA as source country, using the LightGBM model. As observed, the predictions are pretty spot-on in most cases. Special care was taken to make sure that there was no data leakage.

VI. CONCLUSION

Our results show that using a stochastic gradient boosting model such as LightGBM shows promising results in modeling tabular data and capturing non-linear relationships. In theory, neural networks should also provide similar results due to its advancements in modeling non-linear relationships, and it is an option worth being explored further.

Time-series models provided disappointing results due to the fact that there did not exist enough data to accurately model between two countries. Utilizing deep neural networks for time series data would've resulted in overfitting, and hence was skipped.

VII. FUTURE WORK

As denoted by our results, gradient boosting with hyperparameter tuning is worth exploring further. More advanced neural network models such as CNNs can also be explored. Continuing with time series modeling would be useful if more data per country could be collected. Access to more granular data, such as monthly data would be beneficial for time series modeling. Another approach could be to utilize graph representation to apply graph neural networks, to capture country-wise interaction as a link-prediction model.

REFERENCES

[1] Isard, Walter. "Location theory and trade theory: short-run analysis." The Quarterly Journal of Economics 68.2 (1954): 305-320.

[2] Wohl, Isaac, and Jim Kennedy. "Neural network analysis of international trade." US International Trade Commission: Washington, DC, USA (2018).

[3] Sun, Jingwen, et al. "Analysis of bilateral trade flow and machine learning algorithms for GDP forecasting." Engineering, Technology & Applied Science Research 8.5 (2018): 3432-3438.

[4] S. Circlaeys, C. Kanitkar, and D. Kumazawa, "Bilateral trade flow prediction," Unpublished manuscript, available for download at http://cs229.stanford.edu/proj2017/final-reports/5240224. pdf, 2017.

[5] Batarseh, Feras, et al. "Application of machine learning in forecasting international trade trends." arXiv preprint arXiv:1910.03112 (2019)..

[6] Jung, Jin-Kyu, Manasa Patnam, and Anna Ter-Martirosyan. An algorithmic crystal ball: Forecasts-based on machine learning. International Monetary Fund, 2018.

[7] Fouquin, Michel, and Jules Hugot. "Two centuries of bilateral trade and gravity data: 1827-2014." (2016).

[8] A. Halevy, P. Norvig and F. Pereira, "The Unreasonable Effectiveness of Data," in IEEE Intelligent Systems, vol. 24, no. 2, pp. 8-12, March-April 2009, doi: 10.1109/MIS.2009.36.

[9] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017).