WS20/21 - Machine Learning: Applications in Economic Research

Predicting Trade Flow

Data Analysis Project

Yangfan Lyu / Andreas Thiele

15.3.2021

Table of Content

1. Research question

We want to investigate whether machine learning methods would improve out-of-sample prediction of bilateral trade flows when we have a high-dimensional dataset. Conventional bilateral trade flow analysis is based on the gravity equation containing economic size of both origin and destination country, geographic distance, price index, and multilateral resistance terms such as tariff, common borders, and variables related to language and political status etc. Additionally, we are also interested in answering the question whether other variables such as greenhouse gas (GHG) emissions and the inflation rate are likewise found to play a part in determining bilateral trade flow. It is expected that GHG emissions would be positively related to trade flows, as when countries trade more with each other, they are going to have more outputs, become richer, and reach more destinations to expand their market size, thus generating more pollution. Cristea et al. (2011) have pointed out that 33% of worldwide trade-related emissions are associated with international transportation of goods. But we do not know how important GHG emissions are in trade and to what degree they are contributing to explain trade flows. The inflation rate, which is approximated by the average price index (API) of a particular country, is included because it is considered to potentially alter both the direction and the size of international trade by changing optimality conditions regarding to labor and capital as production inputs (Stockman, 1985).

2. Data used

The primary dataset we have used is Tradhist from CEPII, a bilateral trade historical series from 1827 to 2014, consisting of five types of variables: (1) bilateral nominal trade flows, (2) country level aggregate nominal exports and imports, (3) nominal GDPs, (4) exchange rates, (5) bilateral factors that could favor or hamper trade, such as geographical distance, common borders, colonial and linguistic links, as well as bilateral tariffs and indicator for political and economic union. Then we extend this dataset to include global greenhouse gas (GHG) emissions from Kaggle and an indicator average price index (API) to represent inflation for a

country pair in a given year. After having our final dataset, our study period is from 1990 to 2014. 90% of the data will be used as training set and the rest will be used as test set.

2.1. GHG Inventory Data

The dataset contains GHG emissions in each country from 1990 to 2014. We merge GHG emissions data with our primary dataset based on country's iso code and year. As GHG emissions are in total value for each country in a given year, we decided to use weighted value of GHG emissions to pin down the emissions induced by bilateral trade. When considering the weights, we need to think about where emissions could come from in this process. Cristea et al. (2011) has mentioned that emissions are associated with output and transport in every origin-destination-product trade flow. GDP in both origin and destination can be used to indicate output, as for transport, we decided to use bilateral trade flow as a proxy for transportation, as the more bilateral trade flow is, the more public transport will be used, then more emissions trade will induce. Therefore, we create a weighted value of GHG emissions in each entry using GDP in both origin and destination and bilateral trade flow as the weights to indicate potential GHG emissions induced in bilateral trade. The following equation shows how we estimate it.

$$GHG_{ijt} = \left(\frac{GDP_{it} + GDP_{jt}}{GDP_t} + \frac{bilateral\ trade\ flow_{ijt}}{total\ trade\ flow_{it}}\right) \times total\ GHG_{it}$$

where GHG_{ijt} indicates the GHG emission induced by trade flow between origin country i and destination j in time t, GDP_{it} and GDP_{jt} can be found in Tradhist, $bilateral\ trade\ flow_{ijt}$ and $total\ trade\ flow_{it}$ can be obtained from Tradhist, $total\ GHG_{it}$ is from GHG inventory data.

2.2. Average price index

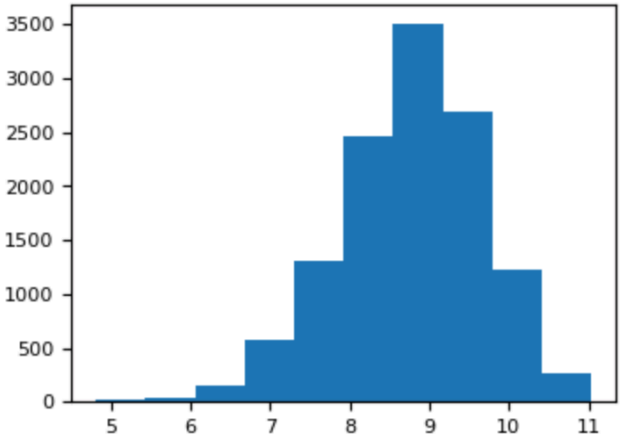
The dataset on average price indices (API) of countries for time period 1960 to 2019 is obtained from the online database of "The World Bank" with the label "Inflation, consumer

prices (annual %). The underlying reasoning for extending our original dataset lies in the economic consideration that inflation is deemed to impact international trade patterns (Stockman, 1985). After downloading the dataset, the given columns involve “Country Name”, “Country Code”, “Indicator Name”, “Indicator Code” and the remaining columns detailing the time periods. As merging requires applying an inner join with a key that identifies entries in both the original and the new dataset, we eliminate all columns except “Country Code” and the time periods. Finally, the structure of the columns for the time periods (horizontally) does not match the vertical layout presented in our original dataset. For this reason, we conduct a transposing measure for all columns except “Country Code” to translate the horizontal axis into a vertical one. After completing this step, we finally apply the inner join and successfully merge our original dataset with our new column “API” detailing the average price index for any given country from time period 1960 to 2019.

2.3. Data cleansing techniques

In order to reserve as much information as possible for further analysis, we have applied several steps in cleaning our data. First, we check the share of NaN in each variable and drop the columns which have a value over 80%. Then we use the reduced primary dataset to do a pre-test of random forest (RF) to check the feature importance and get rid of some variables which are less important and also have a higher share of NaN in their values. Afterwards, we remove the data that have missing values in the remaining columns. Finally, we take log of data to eliminate the long tail in data distribution (see figure 1). After merging it with the other two datasets, we have 12227 observations left.

Figure 1: frequency distribution of log (Flow)



3. Methods

We are going to use linear regression, Least Absolute Shrinkage and Selection Operator (LASSO), Classification and Regression Tree (CART) and RF. Linear regression, including country fixed effects and time fixed effects, will be our benchmark model. We want to check how linear shrinkage model and nonlinear nonparametric model would improve the prediction of bilateral trade flows.

3.1. LASSO

This linear shrinkage model will penalize the coefficients associated with irrelevant variables to zero. Therefore, the consistency for selecting the most relevant variables each time will be achieved only when irrepresentable condition is fulfilled (Zhao and Yu 2006, Meinshausen and Yu 2009), meaning that the selected variables should be as less-correlated as possible to the unselected ones, otherwise the estimators will be biased. This is also referred to as sparsity in the observations. The penalty is given as follows.

$$\sum_{i=1}^{n} p(\beta_{h,i};\ \lambda, \omega_i) \coloneqq \lambda \sum_{i=1}^{n} |\beta_{h,i}|$$

where $p(\beta_{h,i};\ \lambda, \omega_i)$ is a penalty function in which $\beta_{h,i}$ is the coefficient we want to penalize, λ is the penalty parameter and ω_i is the weight. In LASSO case, the penalty term treats every coefficient equally, using the same penalty parameter.

3.2. CART and RF

The RT splits the observations according to the available covariates in a way that when reaching the terminal nodes which are also referred to as regions in Hastie et al. (2001), the outcome variable in each set is as homogenous as possible, meaning that we need to split the observations according to a certain value of covariate k to have the smallest squared error loss. The function below shows the intention.

$$\min \sum_{k \in left} (y_k - \hat{y}_{left})^2 + \sum_{k \in right} (y_k - \hat{y}_{right})^2$$

where y_k is the output values in each terminal node from either left branch or right branch, $\hat{y}_{left}$ and $\hat{y}_{right}$ is the average of the sample y_k. Therefore, RT is a non-parametric model that approximates an unknown nonlinear relation by splitting the covariates continuously until the smallest value in loss function is achieved. The RF model was proposed by Breiman (2001) to reduce the variance of RT by averaging the noise out among randomly constructed RTs.

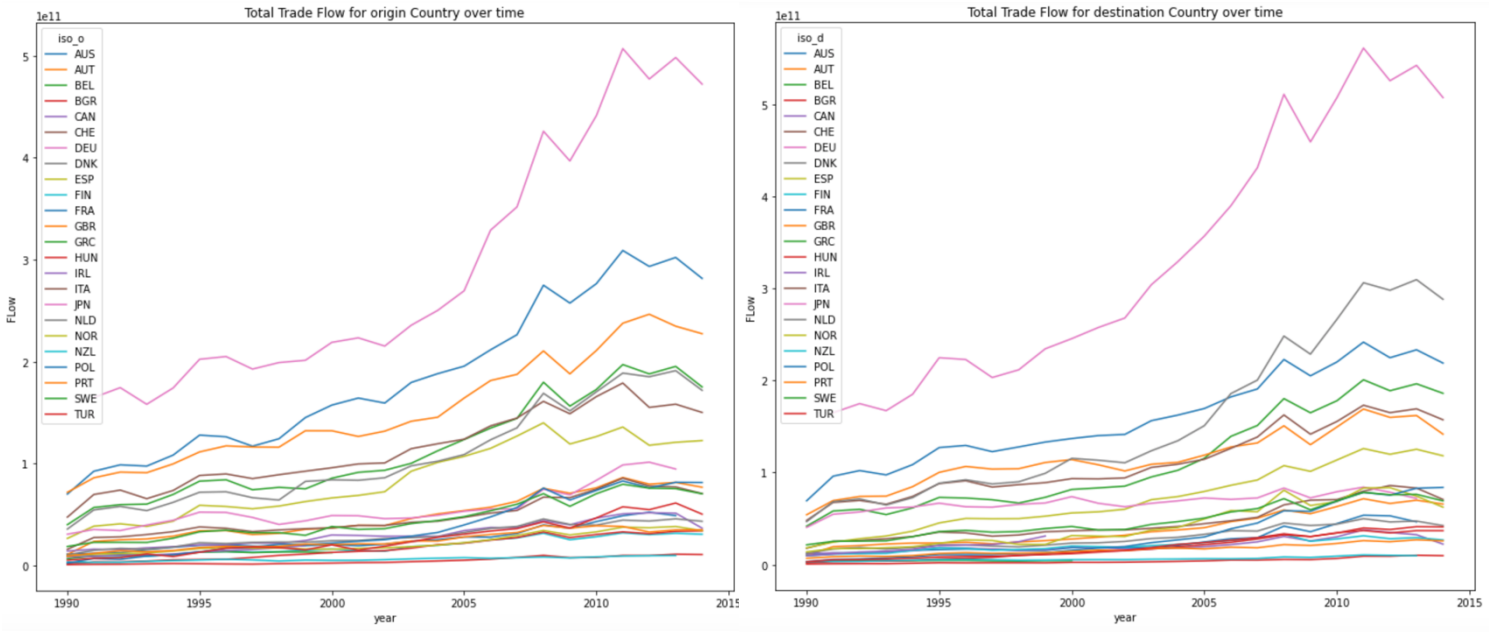
4. Results

Before we look at the out-of-sample prediction from each model, let's have a look at the trade pattern for each country. Figure 2 shows the changes in trade volume for both origin and destination countries from 1990 to 2014. While Japan has the highest trade flow both as origin and destination, France has a larger trade flow as an origin but Netherland started to have higher trade volume since 2006 as a destination. Overall, trade flow presents a persistent increasing trend and approximately linear over time. Therefore, we should have an expectation that linear regression will also achieve a good result in predicting bilateral trade.

Linear regression has achieved a r-squared of 0.9763, which is better than 0.9385 in the LASSO case. One possible reason for this could be that LASSO only selects 9 out of 32 features, which potentially reduces the model's predictive power by eliminating less relevant variables when the trend in outcome is easier to track. CART and RF achieve even better out-of-sample prediction results, 0.993 and 0.999 respectively if we use the optimal tree depth (10 in this case) and optimal number of trees grown in a forest (15 in this case). What's more, in RF we have a mean squared error of 0.002, which is smaller than 0.015 in CART case. This could because of the reason that RF algorithm is a bagging of CART trees. By averaging the prediction results over multiple trees, we preserve the stable relationship among bootstrap samples and average out the noise.

One interesting thing to notice is that none of the three ML models give higher weights to GDP in both origin and destination when weighted value of GHG emissions are included in the models. Particularly in LASSO case, the coefficients of GDP in a country pair are penalized to zero, which is counterintuitive to what gravity model includes. Besides, CART and RF consider the interaction term of API between origin and destination to be important in the models, indicating that instead of the impact of inflation from a single country, country-pair inflation is more important in determining bilateral trade flow by interacting with export or import values.

Figure 2: Total Trade Flow for both Origin and Destination



5. Discussion and Conclusion

We should also include country-pair fixed effects in our benchmark model, but we do not figure out how we should code them in python, so we leave them out this time. Besides, we should also apply Poisson Pseudo Maximum Likelihood (PPML) to tackle with zero trade flows and potential heteroskedasticity presented in error terms, and use it as one of our benchmark models. We have tried Gravity Modeling Environment (gme) package, but also did not manage to finish it.

From the penalty term of LASSO and its nature, we can find out two potential problems that could impose caveats to the results. First, the penalty term does not depend on weight ω_i and treats every feature always to the same degree, which could be a problem in different time periods. As from figure 2 we can see that, similar trade pattern displays in almost all countries from 2007 to 2010, meaning that the financial crisis in US influence almost all of the countries in the data but to different degrees, the most severe ones are in Japan, France, Great Britain, and Netherland. Therefore, we should expect that coefficients of variables that are related to this crisis should be penalized differently in different countries over time, but LASSO fails to do so. One possible solution is using Adaptive Least Absolute Shrinkage and Selection Operator (adaLASSO) proposed by Zou (2006) to include weighting parameter ω_i in the penalty terms and to avoid heteroskedasticity (Medeiros and Mendes 2016) if there is any. Second, as LASSO relies on irrepresentable condition, which leads to the problem of inconsistent model selection, we can try Ridge Regression (RR) which shrinks the coefficient of less relevant variables to nearly zero but not exactly zero for any size of λ, thus relaxing the irrepresentable condition. But in any cases, LASSO will be a good starting point in regularization.

What's more, from our results we can see that linear regression has already achieved a good out-of-sample prediction and ML methods only slightly outperform our linear regression in predicting bilateral trade flow. Two reasons could be considered. First, trade flow is persistent and increases over time, meaning that we can use history data to predict future and that observations are not purely independent on each other. When using non-parametric methods, we need to assume independence between observations. If this assumption fails (i.e., in time-series analysis), parametric models would even outperform. Second, the

outperformance of CART and RF comes from the nonlinearity which allows interaction among variables (i.e., country experiences higher GDP growth or have larger GHG emission will also engage more in international trade). Therefore, ML methods will be more superior than conventional econometric methods when the output variable is much more dynamic and unpredictable over time. And when predicting the outcome, ML methods, especially RF, will achieve a better prediction by capturing the interaction among covariates to better fit the nonlinearity trend presented in the outcome.

References

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32.

[2] Cristea A. D., Hummels D. L., Puzzello L., and Avetisyan M. (2011). Trade and the Greenhouse Gas Emissions from International Freight Transport. NBER Working Paper No. w17117.

[3] Hastie, T., Tibshirami, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction, New York: Springer.

[4] Medeiros, M., and Mendes, E. (2016). $\ell 1$-Regularization of High-Dimensional Time-Series Models with Non-Gaussian and Heteroskedastic Errors. Journal of Econometrics, 191, 255–271.

[5] Meinshausen N., and Yu B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist, 37 (1), 246 – 270.

[6] Stockman, A. C. (1985). Effects of Inflation on the Pattern of International Trade. The Canadian Journal of Economics, 18(3), 587-601.

[7] Zhao, P., and Yu B. (2006). On model selection consistency of Lasso. Journal of Machine Learning Research, 7, 2541–2563.

[8] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. Journal of the American Statistical Association, 101, 1418–1429.