Application of Machine Learning Methods in Econometric–Identifying Critical Variables Through Random Forest and SHAP

Sirui Zhu

School of Business, Macau University of Science and Technology, Macau SAR, China

1230010820@student.must.edu.mo

Abstract. This study integrated explainable machine learning with econometrics to analyze economic drivers of China–U.S. mobile communication device exports. A SHAP-based Random Forest approach with both time-series and resampling robust validation was employed to assess variable quality. Furthermore, dependent relationships between variables were shown through kinds of graphs of SHAP values, and false negative variables from statistical test were identified. The hybrid framework effectively captured and interpreted complex non-linear relationships often overlooked by linear regression of classical econometrics, offering a practical tool for robust variable selection and mechanisms explanation. However, the method did not produce explicit coefficient estimates, limiting traditional hypothesis testing. Overall, it demonstrated how interpretable ML can enhance empirical economic analysis where data dynamics violate classical assumptions.

Keywords: Time-Series Analysis, Explainable Machine Learning, Econometrics

1 Introduction

The method of classical OLS linear regression was used in econometrics for many years, since it was the most explainable model that clearly declared the relationship between observations and control variables. Also, OLS linear regression had undergone a complete process of statistical test to prove the significance of model's relevance. It was made to be the most acceptable method for economics. However, the OLS linear regression had a weak assumption of linear relationships between observations and control variables [11], it made it too difficult to fit the non-linear relationship between data and led to weak performance of the model. Although higher-order terms and interaction terms could be added to regression formula to improve fitting level, other more flexible non-linear relationships are still not be welly captured in the model. Furthermore, the OLS linear regression assumed that coefficients of variables remain unchanged during the whole time periods, which meant some critical variables might be statistically insignificant since their dynamic relationship with dependent

K. Othman et al. (eds.), *Proceedings of the 2026 5th International Conference on Mathematical Statistics and Economic Analysis (MSEA 2026)*, Advances in Economics, Business and Management Research 402,
https://doi.org/10.2991/978-94-6239-774-3_11

variable in observation periods. Nowadays, with the development of computer science and artificial intelligence, methods of machine learning were considered to remedy the disadvantages of OLS linear regression in traditional econometrics.

The method of machine learning had developed various algorithms for regression tasks, and they could easily perform better than the OLS linear regression through optimal parameters. Simple models like Random Forest, Support Vector Machine (SVM) and Extreme Gradient Boosting Decision Tree (XGBoost) were easier to train but weaker to learn deep level relations between different variables, advanced models like Convolution Neural Network (CNN) and Recurrent Neural Network (RNN) had higher requirement on computational power but were able to learn implicit and deep relations. However, fitting level did not mean everything in econometric, to show explainable relations between variables with a reasonable economic logic was more crucial in this area. A consensus in machine learning is, more complicated the models are, less explainable the models are due to so called "Black box" effect [12]. The relations found by machine learning models would be difficult to understand and explain, which was biased from the core target of econometric. In this research, I focused on a traditional economic question, trying to find the impact factors of China's export of mobile communication devices to the U.S. An effort was made to combine the OLS linear regression and explainable machine learning methods, aiming to establish effective and interpretable econometric models. The random forest model would be applied and a validated method that using SHAP value of features in the model was proposed to find out highly relevant variables that contributed more to the description of observations' variance and were robust in resampling.

The essay would be divided into 5 sections. Section 1 and 2 was introduction and literature review, and some background information and literature support were introduced. Section 3 and 4 were datasets processing and proposed methods, in which variables selection, data collection and method explanation were completed. Section 5 and 6 were experiments and conclusions, where proposed method was applied and results would be concluded.

2 Literature Review

Classical econometrics faced various limitations derived from its linear regression model. Its weak assumption of linearity had extremely strict requirements to distribution of data samples. A linear model based on OLS usually was not able to cover comprehensive variance of observed data samples since the realistic confidence interval conducted by the model was significantly smaller than the nominal level of confidence interval [5]. The probability of type I error when estimating coefficients of variables would significantly increase when data samples were critically biased distributed [8]. Some research also revealed that the probability of type I error had a high level of fluctuation depending on specific distribution of data samples [18].

The integration of Explainable Machine Learning (XML) and SHAP into econometrics marked a significant shift from linear modeling to non-linear frameworks. Recent studies indicated that this hybrid approach effectively resolved the opacity of

complex models while retaining the interpretability necessary for economic policy analysis [17]. Methodologically, researchers found that ensemble and deep learning models consistently out-perform traditional econometric techniques. Teruel-Gutiérrez et al. [16] observed that Gradient Boosted Models surpassed penalized logistic regression in predicting CO2-GDP decoupling by capturing complex non-linearities. Similarly, Shahid et al. [13] reported that ML approaches identified a significantly broader range of risk factors for child malnutrition compared to conventional logistic regression.

The application of SHAP as the integrated way further enhanced these models by quantifying feature contributions. Proposed by Lundberg et al. [10], SHAP was a unified framework for interpreting prediction output of machine learning models, especially tree models. It explicitly showed contributions of each sample based on mean value of observations and thus allowed researchers to deeper study the desired relationships. Li et al. [9] utilized XGBoost- SHAP to determine that population size and energy intensity were primary drivers of urban carbon emissions. Shakeel et al. [14] introduced a hybrid VAR-GRU framework, using SHAP to assess the inequality of feature contributions in financial volatility forecasting. Additionally, Wang et al. [17] revealed non-linear relationships between income and electricity consumption that standard regression models failed to detect. Except for quantifying feature contributions, SHAP plots graphs provided effective way to visualize and study the non-linear relationships between dependent and independent variables. “U-shape” relationships of focused variables were identified by SHAP graphs analysis [4], which revealed the explainable complex relationships without staying unchanged over all data samples. And dual variables interaction was found by Antolini et al. [1] through SHAP interaction plots, so that cooperated contribution mechanism of two different variables was exactly learned and shown.

Although there were more applications of XML in econometrics, some research like key variables’ mechanisms explanation and false negative identification in hypotheses test still remained to explored. Therefore, a deeper learning of variables in econometric models was conducted based on method of XML through this research.

3 Variables and Datasets

In this research, economic variables were first selected by traditional way in econometrics, which depended on economic intuitive of researchers. Reviewing historical work, there were some broadly accepted control variables of classic international trade models that were considered. In particular, the exchange rate of RMB over U.S. dollars was expected to reflect the influence of currency inflation or deflation on export. Gross Domestic Production (GDP) of China was supposed to represent the overall productivity of the nation, which would reflect the capability to satisfy export need. Consumer Price Index (CPI) of The U.S. was able to define the consumption ability, which decided need for Chinese products. And a dummy variable of trade war was also considered to capture the variance brought by trade shock. Except for these necessary control variables, some specific variables for the case were introduced.

Foreign Direct Investment (FDI) reflected directly used production capitals from foreign states. The FDI of China was thought to be important since mobile communication devices export to The U.S. was occupied by foreign companies like Apple Inc. and Samsung Inc. over 80 percent. A seasonal dummy variable was introduced as well to extract the seasonality of mobile communication devices market. Considering most of mobile communications devices companies, September to November in each year were thought as intensive periods of new products launch [7]. Thus, the seasonal dummy variable was set depending on this rule.

The data was collected in a monthly frequency, which would be the highest frequency recorded officially. The scope of the research was 142 months, from 2016.01 to 2025.07. Data of the macro variables— GDP, CPI and FDI— were both collected as growth of months on month because of the requirement of seasonal adjustment. And unit variation of variables was processed effectively through quantities replacement with growth rate. To briefly introduce, detailed information about sources of data was shown in Table 1. The FDI variable was too macro for proposed model, which consisted of long-term strategic investment and short-run operational investment in several areas. So, in this case, the GFDI data was lagged in 7 unit months, since an average of seven months period was needed for the operational investment part to complete production, which was also supported by some research [19]. GDP data of China was declared at the highest frequency of seasons, so that the seasonal growth rate of GDP was evenly divided into three values, as the approximate estimations of included month data. Also, for exchange rate of RMB, public data had a high-level frequency of hours. Thus, an average of daily means transaction price in each month was calculated to represent the exchange rate of the month. After preprocessing, the time sequences data was recorded with time stamps in a table and saved as table format. In the end, we got 115 samples, with 1 observations and 6 variables data points in each sample. Some descriptive statistics were shown in Table 2.

Table 1. Detailed Information about Variables

Notes	Variables	Sources	ViF
Exp	export of MCDs	General Administration of Customs of PRC	–
GFDI	growth of FDI	National Bureau of Statistics of PRC	1.3868
TW	trade war	None	1.0375
PL	new products launch	International Data Company	1.3867
GCPI	growth of CPI	Investment.com	1.0936
ER	growth of exchange rate	People's Bank of China	1.0249
GGDP	growth of GDP	National Bureau of Statistics of PRC	1.0663

Table 2. Descriptive Statistics of Datasets

Variables	Count	Mean	Std	Min	Max	Skewness
Exp	115	9.3976	0.2093	8.7283	9.8345	-0.6286
GFDI	115	0.2404	0.2437	0.0210	1.3902	2.2729
TW	115	0.7739	0.4201	0.0000	1.0000	–

PL	115	0.2348	0.4257	0.0000	1.0000	–
GCPI	115	0.0026	0.0030	-0.0080	0.0130	0.4623
ER	115	0.0008	0.0109	-0.0261	0.0371	0.4818
GGDP	115	0.4658	0.9243	-3.5000	4.2667	-0.2593

4 Proposed Methods

In this part, basic units of XML would be introduced in detail. Decision tree and Random Forest were primary machine learning models that used in the research, where I would explain their theories for regression below. Further more, specific method of bootstrapping and testing framework for time series would be clarified which were particularly designed for this economic problems. In the end, a scoring framework for variables would be established and some graphs would be studied for explainable mechanisms.

4.1 Decision tree

Decision tree was a simple machine learning method that could be applied to regression and classification [15]. Each node of it was naturally a binary linear classifier that divided features space into two parts through an optimal divided point. To find the optimal divided points, each node of decision tree would traverse all features data of current features space, and the feature value point that minimize sum of mean square error of the two divided features space's data. After reaching target max depth, the decision tree would establish a regression model according to features space division of each node, where leaf nodes represent the prediction of the tree. It was easily discovered that the fitting accuracy of decision tree greatly depended on max depth of the tree, which represented times of division of the features space.

4.2 Random Forest with SHAP

Random Forest was a machine learning method that performed well in regression tasks. Introduced by Breiman [3], it represents an ensemble learning technique that constructs multiple decision trees during training and outputs the mean prediction of individual trees for regression tasks, as shown in Formula (1). Compared to a single decision tree, random forest

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} T_i(\mathbf{x}) \tag{1}$$

regression model performed more robustly on data which was outside of training set, since it randomly sampled features and data from original datasets with replacement for each decision tree. Generally, larger volume of trees in the random forest model led to higher robustness as a whole and lower sensitivity to uncommon data.

However, it was complicated to explain the process of model establishment in a traditional way just like in a single decision tree because random forest might include

hundreds of trees with different training sets. The key of applying random forest model to econometrics was to know the importance of each features and their direction of contribution to the model. Shapley additive explanations (SHAP) introduced by Lundberg et al. [10] based on cooperative game theory was a broadly accepted way to explain machine learning models. The core of SHAP was SHAP value, which was a quantitative measurement of a single data point in certain features and certain samples. The sign of SHAP value represented the direction of contribution, which meant increased or decreased predictions. Through calculating average of all data samples' SHAP value in the feature, feature importance could be detected, and critical variables would be found.

4.3 SBB and TSCV

The robustness of features importance was critical for statistical significance of variables. Robustness of a specific variable included sampling robustness, which meant the variable was significant in any subset of its datasets, and time robustness, which meant the variable was significant in the whole time series. To prominently validate the robustness of features, Sea- sonal Block Bootstrapping (SBB) and Time Series Cross-Validation (TSCV) were applied in random forest model. Bootstrap was a random sampling method with replacement, it aimed to test the standard deviation of features' mean SHAP value in random subset and identify critical features with sampling robustness. SBB had made some improvements for economic situation. Instead of randomly sampling various single data points, SBB randomly sampled numerous data blocks to integrate validated sets. It ensured that endogenous periodicity of variables would be kept through suitable block length [6]. TSCV was a method that created numerous folds of subsets following chronological order. Unlike standard K-fold cross-validation, which may break natural order of data, TSCV strictly respected chronological order, which was the soul of time series data. It revealed whether the same economic variables maintain their influence across different historical contexts.

4.4 Variable scoring framework

Traditional variable testing and selection methods often overlook the possibility of mechanism changing in macroeconomic, assuming that coefficients of variables remained the same across the entire sample. This oversight risks might cause misdeletion of significant variables that had changing relationships with dependent variable. This study proposed a variable scoring framework that integrated Time Series Cross-Validation (TSCV) with Seasonal Block Bootstrapping (SBB). The objective was to systematically identify critical variables which stably and strongly forced in the whole economic period, focusing on every data samples of variables without only knowing the coefficients. The core of this approach lied in acknowledging the dynamic evolution of economic relationships and testing overall significance of economic variables.

As mentioned above, the framework included two key mechanisms: Seasonal Block Boot- strap (SBB) driven by ACF and time robustness test driven by Time

Series Cross-Validation (TSCV). The SBB was employed to ensure that the inherent periodicity and time serial cor- relation of dependent variable data remained intact during resampling. The critical step here was the determination of the optimal block length, which we defined endogenously using the Auto-Correlation Function (ACF). For a time series consisting of T data samples, the value of ACF function in lag k was calculated as Formula (2). The process was begun by calculating the internal correlation index of the dependent

$$r_k = \frac{\sum_{t=1}^{T-k}(y_t-\bar{y})(y_{t+k}-\bar{y})}{\sum_{t=1}^{T}(y_t-\bar{y})^2} \quad (2)$$

variable series by ACF. The number of lags required for the ACF value to decay into statistical insignificant level of 0.95 reflects the" memory" duration of the variable. The block length is then set to cover this period of significant autocorrelation. If correlation period was shorter than 12 months, block length would be set as 12 months to capture at least one year completed economic cycle. An exception was that, when correlation duration lasted over 24 months, to ensure enough sample volume, block length would be set as multiples of cube root of data volume according to Bertail et al. [2]. This made sure that the endogenous dynamics were mostly preserved within each sampled block. Compared to the common Bootstrap, the synthetic samples generated by SBB statistically resembled the actual macroeconomic environment, thereby enhancing the credibility of the variable importance statistics derived from them. To respect chronological order of datasets, TSCV was introduced to generate numerous testing folds with different volume of data, utilizing "expanding window" strategy to chronologically increased testing samples. Then SBB with 50 times random sampling was employed in each fold of TSCV. Mean absolute SHAP values of each variables' data samples were calculated as variables' SHAP values in single sampling, and mean values and standard deviation of multi-times sampling were calculated as variables' SHAP values and their stability in each fold. After finishing calculations of each fold, statistics of every folds were integrated. Average of all folds mean SHAP values was seen to represent overall significance of variables. Further more, overall sampling volatility was calculated as standard deviation of all folds SHAP values standard deviation divided by average of all folds SHAP values standard deviation, and cross-periods volatility was calculated as standard deviation of all folds mean SHAP values divided by average of all folds mean SHAP values. A significant level of CVs was set to 0.3, which meant that variables with CVs lower than 0.3 would be considered as excellent robust variables. In the end, variables were scored through 3 dimensions, and final scores of each variable were weighted integration of average of all folds mean SHAP values, overall sampling volatility and cross-periods volatility, with weights of 0.5, 0.25 and 0.25. This setting of weights meant that what I focused first was mean absolute contribution of each variable, and coming behind were robustness in resampling and across time series. Comparing final scores of each variables, a variables order would be obtained to be an instruction for econometric model.

4.5 Econometric model and explain

After identifying critical variables through the scoring framework, the OLS linear econometric model could be established. Focusing on statistics of regression result, some cross validation could be conducted. Variables with high combing score would usually be statistically significant in econometric model, it revealed strong and stable relationships between those variables and independent variable. When there were conflicts between the two methods' results, numerous graphs of SHAP would be learned to further analyze the specific relation- ships. With the graphs, some insignificant variables could be expected to be explained without being incorrectly abandoned. Critically, the SHAP graphs serve only an exploratory and illustrative function in the framework, further conclusions still needed to be drawn from rigorous statistical tests.

5 Experiment

5.1 Model establishment and variables scoring

The first step was to establish a random forest regression model and make regression on the whole datasets. The relationship between dependent variable and independent variables was denoted as Formula (3):

$$E\hat{x}p_t = \mathcal{F}(GFDI_t, TW_t, PL_t, GCPI_t, ER_t, GGDP_t) \tag{3}$$

The mobile communication devices export from China to The U.S., which was dependent variable, was denoted as Exp_t in month t. Among the independent variables, the growth of FDI in China month on month in time t was denoted as $GFDI_t$, and the seasonal dummy variables of new products launch was denoted as PL_t. Other macro control variables were listed behind to capture necessary environment variance. In the month time t, growth of China's GDP month on month was denoted as $GGDP_t$, growth of American CPI month on month was denoted as $GCPI_t$. The growth of exchange rate at time t was denoted as ER_t, and the dummy variables of trade war was denoted as TW_t. The function F was a nonlinear-tree regressor to fit the unknown relationship. To prevent from overfitting, max depth of each tree was limited to 5 and 100 trees were integrated in the model. As the fitting result, the random forest model achieved an R^2 of 0.87. Then time series cross validation would be applied in 10 folds of datasets with different size, and Seasonal Block Bootstrap was conducted 50 times in each fold.

Statistics of all tests would finally be recorded and calculated, which were shown in the next subsection. Figure 1 showed a summary plot SAHP values based on the model over the whole datasets as form of beeswarm, which briefly extracted distribution of SHAP value of each variables according to their values, and provided a reference for resulted statistics in TS- CV. The resulted statistics would be accessible after conducting TS-CV, and were recorded in table format. As Table 3 shown, variables were sorted according to their combined scores. Each column recorded 4 critical measurements of different variables. The ranking of variables' scores was PL, GFDI, ER, GGDP, GCPI and TW. It was not difficult to discover that PL, GFDI and ER variables has excellent performance in the test, which meant they had higher mean absolute contribution while keeping acceptable level of two coefficient of variation

(CV) values. It revealed that contribution mode of the 3 variables remained comparably stable in different sub-samples and different chronological contexts. The macro control variables of GGDP and GCPI gained lower combined scores, since they had relatively lower mean absolute contribution and greater CV. With a great CV value, especially for CV-overall-sampling of GCPI, it represented that there was huge volatility in variables' contributions and thus their mean absolute contribution might be ineffective and unrealistic. The variable of TW performed worst in the test, both poor in mean and CV. The trend and distribution of each variable's SHAP values in time series were shown in Figure 2 and Figure 3, in which we could find that SHAP values of GFDI and GGDP represented some kinds of cycle and repetition, while other variables varied without obvious rules in chronological order. According to the ranking, econometric model of OLS linear regression could be established through step-wise backward selection.

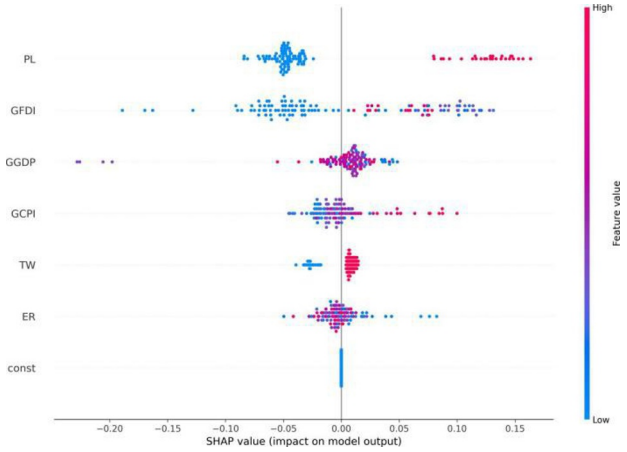

Fig. 1. Summary of SHAP Values

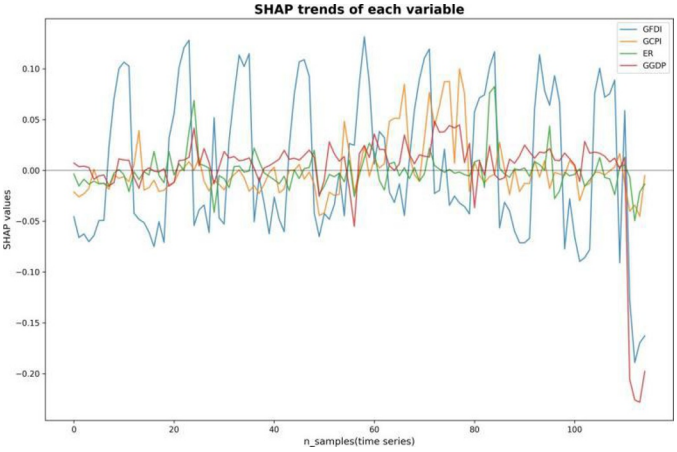

Fig. 2. Trend of SHAP values

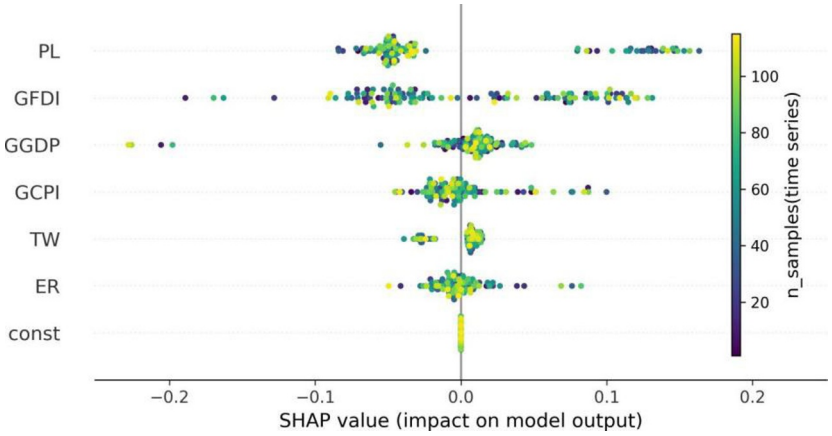

Fig. 3. Chronological Distribution Beeswarm

Table 3. Variables Testing Results

Variables	Mean-SHAP-means	CV-overall-sampling	CV-cross-folds	Combined scores
PL	0.057422	0.307621	0.144874	0.629272
GFDI	0.044180	0.231495	0.169172	0.550802
ER	0.020061	0.173678	0.166282	0.391376
GGDP	0.035683	0.326332	0.341104	0.310703
GCPI	0.013856	0.493286	0.305883	0.120654
TW	0.004322	0.739009	0.679314	0.037629
const	0.000000	–	–	0.000000

5.2 Econometric model and statistical test

As beginning all 6 variables were included in the econometric linear model, whose regression result was shown in Table 4 column (1). The variables of GFDI and ER were statistically insignificant, while TW variable was weakly significant. Since mobile communication devices were not included in U.S. tariffs policies against China, and the coefficient of TW was even positive which did not make sense, the variable's significance seemed to be incredible. Considering uncommon CV and the lowest score of the variable in the test, the first variable to be deleted was TW. A new econometric model was established without the variable of TW, as Table 4 column (2) shown. It was obvious that deletion of TW was effective and accurate. The significance of other variables approximately had not changed while avoiding unacceptable lost of R-square score and Durbin-Watson index, which could be concluded as irrelevance of TW variable. After removing variable of TW, I started to consider the next target which was capable for deletion. The GCPI variable was not robust enough in the machine learning test but was strongly significant in econometric statistical test. As for trying, the variable was deleted from model, whose result was shown in Table 4 column (3). The absence of GCPI variable had made a terrible outcome, some other

variables became greatly insignificant while the Durbin-Watson index and R square score dropped seriously. The result meant that critical variance of independent variable was captured by GCPI variable, which further implied the deletion should be rejected.

While the variables of GFDI and ER performed excellent in mean absolute SHAP values and CVs of both two dimensions, it was necessary to seek out the reason why the two variables were statistically insignificant in the econometric linear model. A dependence plot which aimed to precisely depict dependent relationships between SHAP values of variables and original values of variables was shown in Figure 4. The colors of samples represented values change of an interacted variable with largest probability to have inter-relations, which was automatically identified by SHAP mechanism. It was shown that SHAP values of GCPI variable had relatively strong linear relationship with increasing the variable's values, which represented a simply positive relevance between dependent variable and GCPI. Associating with Figure 1, it was not difficult to find that distribution of GCPI variable's SHAP values was sparse in larger values range, which could led to not outstanding mean absolute importance and CVs performance in the result of variables test. When focusing on the plot of GFDI variable, it could be found that a stronger positive relevance comparing to GCPI variable was existed between SHAP values and the variable's values, when values of GFDI lower than approximately 0.5. This kind of relevance disappeared when values of GFDI greater than 0.5 and instead transited to distribution with invisible rule. The plot was expected to explain why the variable of GFDI earned high score about mean absolute importance and CVs but was statistically insignificant in the econometric linear model. GFDI seemed to be a critical variable with excellent performance in both different samples and chronological context, but its positive relationship with dependent variable was circumstantial to the range of the variable's values, which was possibly the reason for its insignificance. When came to the plot of ER variable, things had been different. The SHAP values' points were approximately distributed beside zero horizon randomly and evenly without visible dependent relationship. Along with Figure 1, the SHAP values of ER variable kept fluctuation in a relatively stable range, which consisted of result in Table 3 that the variable had eared high combined score. After removing variable of ER, the regression result was shown in Table 4 column (4). It was proved that the variable also played unimportant role in the model. However, it was not rigorous to conclude that the variables of GFDI and ER had no relationships with the dependent variable when there were no visible relationships in the plots. Further tests of non-linear relationships needed to be conducted to draw conclusions, which would not be included in this research.

Table 4. Regression Results with Different Variables

	(1)	(2)	(3)	(4)
GFDI	0.209	0.160	0.141	0.197
TW	0.080*	–	–	–
PL	0.000***	0.000***	0.000***	0.000***
GCPI	0.000***	0.000***	–	0.000***

ER	0.149	0.180	0.455	–
GGDP	0.002***	0.006***	0.544	0.003***
Const	0.000***	0.000***	0.000***	0.000***
Count	115	115	115	115
R^2	0.417	0.401	0.354	0.399
Adj. R^2	0.385	0.374	0.330	0.377
D-W	0.945	0.945	0.775	0.943

Notes: Robust standard errors are in parentheses. *(p <0.10), **(p < 0.05), ***(p < 0.01).

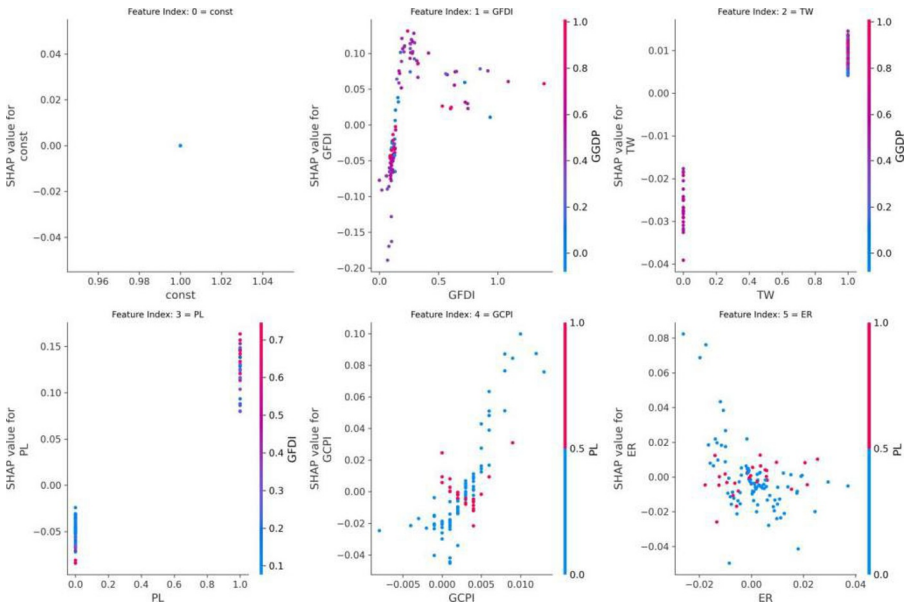

Fig. 4. Dependence of SHAP Values

5.3 Results analysis

After stepwise deletion process mentioned above, the final econometric model was found and the most critical economic variables about export of mobile communication devices from China to The U.S. were identified. In the end only four economic variables from original selected variables were kept and their relevant mechanism was further explored.

The variable of TW was deleted from model firstly, mainly because of its poor performance of stability. When reviewing The U.S. policies of tariffs including 301, 232, 201, 338, 122 sections, there was not enough evident to prove that these policies would cause effect on export of mobile communication devices from China to The U.S. Both results of machine learning model and econometric model had showed a positive relationship between TW variable and dependent variable, meaning that outbreak of trade war facilitated export from China to The U.S., which were unreasonable and ridiculous. Considering extremely high volatility of TW's mean absolute con-

tribution, the dependent relationship shown in results was not a stable outcome, which meant its influenced mechanism had changed hugely in different subsets of the same time period and different time periods. The weak significance shown by TW variable in the econometric model might be an illusion, which caused by occasionally absorbing dependent relationship of unknown variables.

Percentage change of exchange rate was also abandoned. Although the ER variable per- formed excellent on contribution stability test, earned great scores on the two CVs of overall sampling and cross time folds, its dependent relationship was insignificant in both machine learning model and econometric model. It could be concluded that inflation or deflation of RMB against U.S. dollars had no effect on export of mobile communication devices in the research periods of ten years. A possible reason might be that fluctuation of exchange rate did not critically influenced demand of mobile communication devices, whose price for retail was comparably low and not so sensitive to be perceived by consumers even with tiny higher cost of exchange. It was said that the American might have greater inelastic demand for mobile communication devices ever than we expected.

Growth of FDI and CPI were both significant driving factors for the economic problem. Among them, the variable of GCPI had a significantly positive relevance with export of mobile communication devices in the whole research periods. This relevance was easy to explain because growth of CPI in certain months directly reflected consuming power of American in those months, which was thought to be a more critical factors rather than exchange rate that influenced demand. Another variable of GFDI was found to be positive relevant with export of mobile communication devices only when values of the variable were smaller than 0.5, which implied that growth rate greater than 50% of FDI would loss its regular driving mechanism and turned to be irrelevant. This phenomenon was kind of overthinking, whose reason might be that growth over 50% included focused investment on other industries which did not impact on the research industry of mobile communication devices.

Remaining variables GGDP and PL were proved to be necessary macro factors in the economic problem. Periods of new products launch effectively captured seasonality of the focused export, showed significant force mechanism when in periods with frequent products launch. Growth of GDP in China as a variable had shown unstable contribution in differ- ent subsets and context, but was validated to be critical for econometric model. I inferred the reason might be that impact of this macro index was not so great and explicit, and it depended on interaction with other variables. However, it was proved that growth of GDP which represented production capability of China did have positive relationship with export.

6 Conclusion

In this research, an effective integrated methodology based on both XML and econometric was developed. In particular, it performed well in mechanism explanation of variables' contribution. To expand, it allowed implicit relationships between variables with biased distributions to be effectively captured, and mechanism transitions ap-

peared as changing coefficients would also be identified without being tested under weak linear assumption of econometric. In summary, XML indeed provided a new and acceptable method to solve traditional problems of classical econometrics under strict assumption of data distribution and linearity, and produced potential for researchers to detailly study explainable-local-dependent relationships. However, there were some limitations of current XML methods. The latest model based on SHAP could not generate explicit functions and precise coefficients of desired relationships, which was a constraint for statistical hypotheses test that broadly used in econometric. Furthermore, variances introduced by random resampling were still underexplored, and were looking forward to further research.

References

1. Antolini, F., Cesarini, S., & Terraglia, I. (2025). Explaining tourism expenditure patterns in italy: A comparative study of domestic and incoming tourists using xgboost and shap. *Social Indicators Research*, *180*(1), 369–382.
2. Bertail, P., & Dudek, A. E. (2024). Optimal choice of bootstrap block length for periodically correlated time series. *Bernoulli*, *30*(3), 2521–2545.
3. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.
4. Chen, J.-e., & Jing, A. (2025). Recent advances in causal machine learning and dynamic policy learning. *Wiley Interdisciplinary Reviews: Computational Statistics*, *17* (4), e70050.
5. Cuesta-Albertos, J. A., García-Portugués, E., Febrero-Bande, M., & González-Manteiga, W. (2019). Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics*, *47* (1), 439–467.
6. Dudek, A. E., Leśkow, J., Paparoditis, E., & Politis, D. N. (2014). A generalized block bootstrap for seasonal time series. *Journal of Time Series Analysis*, *35*(2), 89–114.
7. Giachetti, C., & Marchi, G. (2010). Evolution of firms' product strategy over the life cycle of technology-based industries: A case study of the global mobile phone industry, 1980–2009. *Business History*, *52*(7), 1123–1150. Retrieved from https://doi.org/10.1080/00076791.2010.523464 doi: 10.1080/00076791.2010.523464
8. Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior research methods*, *53*(6), 2576–2590.
9. Li, Y., He, Y., Yang, F., An, H., Li, J., & Xie, Y. (2025). An xgboost-shap analysis of the driving factors of carbon emissions in china's first-tier cities. *Scientific Reports*.
10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
11. Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*(52), 145–158. Retrieved 2026-02- 09, from http://www.jstor.org/stable/621706
12. Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200-42216. doi: 10.1109/ACCESS.2020.2976199
13. Shahid, M., Yahya, M. A., Song, J., Naveed, H. M., Yuksel, S., Dincer, H., & Ali, M. (2025). Machine learning vs. traditional logistic regression: predictive performance and risk factor identification for child nutritional outcome in pakistan. *BMC public health*, *1*(1).

14. Shakeel, M. R., Siddiqui, T. A., Alam, S., & Ahmad, M. (2025). Leveraging explainable ai, gru and var based framework to uncover relationships between endogenous and exogenous volatility indices for effective trading strategies. *Computational Economics*, 1–31.
15. Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*.
16. Teruel-Gutiérrez, R., Fernandes da Anunciação, P., & Teruel-Sánchez, R. (2026). Modeling absolute co2–gdp decoupling in the context of the global energy transition: Evidence from econometrics and explainable machine learning. *Sustainability*, *18*(2), 758.
17. Wang, Y., Hu, L., Hou, L., Wang, L., Chen, J., He, Y., & Su, X. (2024). A shap machine learning-based study of factors influencing urban residents' electricity consumption-evidence from chinese provincial data. *Environment, Development and Sustainability*, *26*(12), 30445–30476.
18. Warton, D. I., Lyons, M., Stoklosa, J., & Ives, A. R. (2016). Three points to consider when choosing a lm or glm test for count data. *Methods in Ecology and Evolution*, *7* (8), 882–890.
19. Zhang, Y., Cheng, Z., & He, Q. (2019, 11). Time lag analysis of fdi spillover effect: Evidence from the belt and road developing countries introducing china's direct invest- ment. *International Journal of Emerging Markets*, *15*(4), 629-650. Retrieved from https://doi.org/10.1108/IJOEM-03-2019-0225 doi: 10.1108/IJOEM-03-2019-0225