IJSES
Volume 10, Issue 2, pp. 68-76, 2026.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

Application of Machine Learning in Forecasting International Tourist Arrivals to Vietnam Based on Google Trends Data and Macroeconomic Factors

PhD Thi-Huyen-Trang Vu[1], Thi-Thanh-Hai Trinh[2], Duc-Manh Le[3], Minh-Vu Hoang[3], Tien-Dat Dinh[3], Quynh-Lien Hoang[3]

[1]Faculty of Mathematical Economics, Thuongmai University, Hanoi, Vietnam; email: trang.vth@tmu.edu.vn
[2]K60V1, Faculty of Mathematical Economics, Thuongmai University, Hanoi, Vietnam;
[3]K60V2, Faculty of Mathematical Economics, Thuongmai University, Hanoi, Vietnam

Abstract— *This study develops and evaluates models for forecasting international tourist arrivals in Vietnam by integrating online search behavior data from Google Trends and macroeconomic factors such as GDP growth, exchange rates, consumer price index, interest rates, oil prices, and foreign direct investment. Monthly time series data from 2008–2025 are used to estimate and compare the forecasting effectiveness between traditional econometric models and machine learning models such as RF, XGBoost, and EXTrees. Empirical results show that machine learning models achieve higher forecasting accuracy, especially during periods of significant market volatility. The study contributes to clarifying the role of online behavior data as an early indicator in forecasting tourism demand and proposes an integrated approach to improve the effectiveness of tourism development strategy planning in Vietnam in the context of digital transformation and global economic fluctuations.*

Keywords— *Tourists, Vietnam, Macroeconomic Factors, Google Trends.*

I. INTRODUCTION

In the context of strong global economic and social development, the digitalization process and the ubiquity of the Internet have fundamentally changed the way people access information and make consumption decisions. The development of technology infrastructure, smartphones and online search engines has made the behavior of seeking information an inevitable part of the decision-making process, especially for services related to tourism. Consumers seek information about destinations, prices, visas, airline tickets or weather conditions before making a travel decision. As a result, online search query data, compiled through Google Trends, can reflect the level of interest and travel intentions of international visitors over time.

In parallel with the development of digital technology, tourism is considered a "smokeless industry" and plays an important role in the economic structure of many countries, including Vietnam. According to the Vietnam National Administration of Tourism, in 2019 Vietnam welcomes 18 million international visitors, nearly tripling from 2010, directly and indirectly contribute about 9–10% of national GDP and create jobs for millions of direct and indirect workers. However, the tourism industry has also shown high sensitivity to global economic and health shocks. The COVID-19 pandemic caused a serious decline in international visitors, showing the economy's significant dependence on international visitor flows and the vulnerability of the industry to external fluctuations.

In this context, accurate forecasting of international tourism has become particularly important for policy making, resource allocation and sustainable development strategy. Traditional economic models, although capable of explaining the role of macro factors such as income, exchange rates, or inflation, are limited in reflecting changes in tourist behavior in a timely manner. In contrast, GoogleTrends data provides early signals about travel interest trends, while Machine Learning allows for nonlinear relational exploitation and integration of various data sources in the same forecasting framework.

Stemming from the intersection between the development of digital technology and the strategic role of the tourism industry in Vietnam's economy, the application of Machine Learning in forecasting international tourists based on Google Trends data and macroeconomic factors is necessary both academically and practically. This study aims to build an integrated forecasting model, capable of improving accuracy and early response in the context of an increasingly volatile tourism market.

II. THEORETICAL BASIS AND LITERATURE REVIEW

2.1. Theoretical basis

Concepts of Tourism and Tourism Demand

Tourism is defined as the activities of persons traveling to and staying in places outside their usual environment for not more than one consecutive year for purposes other than the exercise of an activity remunerated from within the place visited (UNWTO, 2010). From an economic perspective, tourism is a comprehensive service industry characterized by intangibility, perishability, and high sensitivity to socio-economic fluctuations (Song and Li, 2009).

International tourism refers to cross-border travel flows, comprising both inbound and outbound arrivals. Inbound tourism is regarded as a form of "invisible export" as it generates foreign exchange earnings for the host country. Compared to domestic tourism, international tourism is more significantly influenced by the income levels of origin countries, relative prices, exchange rates, and institutional barriers, reflecting the cross-border cost characteristics of

international tourism demand (Witt and Witt, 1995; Song and Li, 2009).

Tourism demand forecasting is grounded in the economic theory of demand, where the number of arrivals is determined by factors such as income, prices, and macroeconomic conditions (Song and Li, 2009). Variables such as GDP per capita, exchange rates, and inflation are frequently employed to reflect tourists' purchasing power and relative costs. Recently, online search data like Google Trends has been considered a leading indicator of demand, as it reflects real-time consumer behavior and intentions. Integrating these search queries with traditional economic variables enhances the predictive accuracy of forecasting models.

Factors Affecting International Tourism Demand

Domestic macroeconomic factors play a crucial role in determining international tourist arrivals to Vietnam. According to tourism demand theory, international travel demand is generally influenced by income levels, relative prices, and macroeconomic conditions. Vietnam's GDP growth reflects economic development, infrastructure quality, and service supply capacity. Higher GDP growth enables both public and private investment in airports, transportation networks, accommodation facilities, and destination marketing, thereby enhancing tourism competitiveness and destination attractiveness.

The exchange rate directly affects travel costs. A depreciation of the Vietnamese Dong relative to major foreign currencies increases Vietnam's price competitiveness, potentially stimulating inbound tourism demand. In contrast, higher inflation, measured by the Consumer Price Index (CPI), raises domestic price levels and may reduce the country's relative price advantage. Interest rates indirectly influence tourism demand through capital costs and investment in tourism infrastructure, while oil prices affect international air transport costs, thereby altering travel expenses.

Foreign Direct Investment (FDI) contributes to infrastructure improvement, development of international-standard hotels and resorts, and expansion of business-related travel. In addition to traditional macroeconomic determinants, Google Trends search indices serve as a leading behavioral indicator, reflecting travel interest and intention prior to actual visitation, and thus provide supplementary predictive information for forecasting international tourist arrivals.

Impact of International Tourist Arrivals on the Economy

International tourist arrivals have significant impacts on the Vietnamese economy, particularly in its recovery following the COVID-19 pandemic. This impact is evidenced through the expenditure channels of foreign visitors. As international arrivals increase, total foreign currency expenditure on accommodation, transportation, dining, and entertainment services rises, thereby increasing aggregate demand and contributing directly to GDP. According to the General Statistics Office of Vietnam, the country welcomed approximately 18 million international arrivals in 2019, marking the period when tourism contributed most significantly to economic growth before the pandemic. Conversely, a sharp decline in international arrivals during 2020–2021 led to a deep recession in related service sectors.

Simultaneously, international tourist expenditure creates a substantial source of foreign exchange (approximately 18–19 billion USD in 2019, according to the Vietnam National Authority of Tourism), contributing to the improvement of the balance of payments and foreign reserves. Beyond direct impacts, the growth in international arrivals also stimulates demand for international transport, hotels, and high-end consumption. According to the World Travel & Tourism Council, international tourist spending has a significant spillover effect throughout the economy. Consequently, fluctuations in international tourist arrivals are positively correlated with economic growth, foreign exchange sources, and the vibrancy of the service sector in Vietnam.

2.2 Literature Review

Studies applying traditional models combined with macroeconomic data in tourism demand forecasting primarily employ conventional econometric approaches such as linear regression, ARIMA, and ARIMAX models incorporating macroeconomic variables including GDP, exchange rates, and relative prices. Song and Witt (2006) indicate that income in the origin country and travel costs significantly affect international tourist arrivals. These models offer advantages in terms of interpretability and in clarifying the relationship between macroeconomic determinants and tourism demand. However, due to their assumption of linear relationships and temporal stability, forecasting performance may deteriorate when economic shocks or abnormal fluctuations occur.

Studies applying machine learning models combined with macroeconomic data have emerged as an extension of traditional forecasting approaches, as the development of nonlinear methods has significantly expanded tourism demand research, particularly when data are influenced by macroeconomic factors. In the international tourism forecasting competition, Athanasopoulos et al. (2011) compared the performance of various forecasting methods, including time-series models, econometric models with exogenous variables, and combination approaches. The results indicate that nonlinear methods and model combinations can improve forecasting accuracy in contexts where tourism demand is affected by economic factors such as income and prices. However, the study primarily relied on historical data and traditional macroeconomic variables, without incorporating online behavioral indicators within the same modeling framework.

Studies applying machine learning models combined with search behavior data (Google Trends) have attracted increasing attention in tourism forecasting. Online search data, particularly Google Trends, are considered leading indicators reflecting travel intention prior to actual visitation. Choi and Varian (2012) showed that search query data can improve short-term economic forecasting performance. In the tourism context, Yang et al. (2015) found that integrating search indices into forecasting models enhances predictive accuracy compared to models using only historical data. However, many existing studies either use search data independently or fail to integrate them simultaneously with macroeconomic factors, especially in the context of Vietnam.

IJSES
Volume 10, Issue 2, pp. 68-76, 2026.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

III. Research Data and Proposed Methodology

3.1. Research Data

Macroeconomic and International Tourist Data

The study utilizes monthly time-series data from January 2008 to December 2025. Data were aggregated from reputable sources including the General Statistics Office (GSO), Vietstock, Investing.com, Trading Economics, and the Vietnam National Authority of Tourism. To ensure accuracy, the authors performed cross-checks between official reports and credible databases to minimize input errors. Variable details are presented in Table 1.

TABLE 1. Variables Used in the Model

No.	Variable	Symbol	Unit	Source
1	Gross Domestic Product	GDP	%	Vietstock
2	Exchange Rate	EX	VND/USD	Investing.com
3	Consumer Price Index	CPI	**Index, month-on-month, base = 100)**	General Statistics Office (GSO)
4	Interest Rate	IR	%	Tradingeconomics.com
5	Oil Price	OIL	USD	Investing.com
6	Foreign Direct Investment	FDI	Billion USD	Tradingeconomics.com and Ministry of Planning and Investment (MPI)
7	Tourist Arrivals	T	Persons	Vietnam National Authority of Tourism (VNAT) and General Statistics Office (GSO)

Source: Author group

Google Trends Data

The utilization of Google Trends requires careful consideration to ensure the data accurately reflects the actual demand of international tourists. Consequently, keywords were selected via Google Suggest and subsequently screened using Pearson correlation analysis to ensure the accuracy and reliability of the statistical results.

Details

- The Google Suggest tool was employed to identify potential keywords related to Vietnam tourism. Keywords from the research by Ngo Van Son et al. (2024) were utilized as initial search terms. Specifically, the phrase "Visit to Vietnam" was used to prompt Google Suggest to generate potential keyword recommendations.
- Upon collecting the list of potential keywords from Google Suggest, the Pearson correlation coefficient was calculated.

To evaluate the relevance of each keyword to international tourist arrivals to Vietnam, the research team applied the Pearson correlation coefficient to the Google Trends data series for each keyword and the monthly international tourist arrival data series. This method was utilized to determine the degree of linear relationship between user search interest on Google and fluctuations in international tourist arrivals.

The formula for calculating the Pearson correlation coefficient is defined as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Following the calculation of the Pearson correlation coefficients for each keyword, these coefficients were compared to identify the keywords with the highest correlation with the international tourist arrival data series. Keywords with correlation coefficients closer to 1 were considered to have a stronger relationship with tourist arrival fluctuations and were thus prioritized for analysis and forecasting purposes. The results indicate that the keywords "Visit to Vietnam," "Best time to visit Vietnam," "Place to visit Vietnam," "Vietnam best place to visit," and "hanoi" yielded correlation coefficients exceeding 0.75 relative to the international tourist arrival data series. This demonstrates that variations in the search volume of these keywords are closely associated with changes in the number of arrivals.

3.2. Models

Random Forest (RF)

Random Forest is a machine learning algorithm within the supervised learning category, commonly utilized for classification and regression tasks. This algorithm comprises multiple Decision Trees. “After a large number of trees is generated, they vote for the most popular class. We call these procedures random forests.” (Breiman, 2001, p. 6). For classification problems, Random Forest aggregates the predictions from each tree and selects the result with the highest frequency. In regression tasks, the final output is the average of the predictions from all constituent trees.

In Random Forest, training subsets are generated via bootstrap sampling from the original training dataset (Breiman, 2001). Each decision tree within the Random Forest performs classification or regression by partitioning the feature space into sub-regions. These partitions are determined based on splitting conditions at each node within the tree to create new nodes or leaf nodes, where the output value of that node is used as the final prediction.

Extremely Randomized Trees (EXTrees)

Extremely Randomized Trees (Extra Trees) proposed by Geurts et al. (2006) is an ensemble learning method based on the construction of multiple decision trees and the combination of their predictive results to produce a final forecast. This model is grounded in the principle that aggregating multiple individual models with high diversity enhances stability and generalization capabilities. The distinguishing feature of Extra Trees lies in the high degree of randomness during the tree-building process: at each node, the algorithm randomly selects a subset of features and split values rather than searching for the optimal split point across the entire dataset. Due to this mechanism, the trees within the model exhibit lower correlation, thereby reducing variance and enabling the model to operate effectively on non-linear or noisy data while maintaining computational efficiency.

Extra Trees and Random Forest share similarities in their use of multiple decision trees and aggregation mechanisms; however, Random Forest typically employs bootstrap sampling to create data subsets and searches for the optimal split point at each node. In contrast, Extra Trees increases the level of

randomness by "splits nodes by choosing cut-points fully at random and that it uses the whole learning sample (rather than a bootstrap replica) to grow the trees" (Geurts et al., 2006, p. 6). Consequently, Extra Trees often achieves faster training speeds and lower variance, though it may occasionally trade off some bias error. Incorporating Extra Trees into this study provides a machine learning method with robust generalization capabilities, leveraging the advantages of ensemble tree models while partially addressing limitations related to speed and overfitting.

Extreme Gradient Boosting (XGBoost)

Traditional machine learning models, such as Decision Trees and Random Forest, offer the advantages of interpretability and ease of implementation; however, they often face limitations in accuracy when processing complex data structures or non-linear relationships. In this context, XGBoost was developed as an advanced machine learning algorithm to enhance forecasting effectiveness while optimizing speed and computational performance.

According to Chen and Guestrin (2016), XGBoost is an optimized implementation of gradient tree boosting aimed at improving predictive accuracy and computational efficiency. The model constructs decision trees sequentially in an additive manner, where each newly added tree is trained to minimize a differentiable loss function. The theoretical foundation of gradient boosting was established by Friedman (2001), who formulated boosting as a gradient descent procedure in function space, in which each newly added tree approximates the negative gradient (i.e., the residual) of the loss function with respect to the current model. Building upon this theoretical framework, XGBoost incorporates a regularized objective function and employs second-order Taylor expansion to utilize both first- and second-order derivative information, thereby enhancing generalization performance and controlling model complexity.

3.3. Model Comparison Metrics

Mean Absolute Error (MAE)

MAE represents the average error magnitude in the original units of the data, facilitating ease of interpretation. When comparing models, a lower MAE value indicates a superior model, as its average forecasting error is lower.

$$MAE = \frac{\sum_{i=1}^{n}|e_t|}{n} = \frac{\sum_{i=1}^{n}|y_t - \hat{y}_t|}{n}$$

Where:

y_t the actual value of the dependent variable at the i-th observation

$\hat{y}_t$ the forecasted (or predicted) value of the dependent variable at the i-th observation

n the number of observations

Root Mean Square Error (RMSE)

In statistics, RMSE is the standard deviation of the residuals. RMSE is calculated as the square root of the average squared differences between the actual and predicted values.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\Sigma_{i=1}^{n} e_t^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(y_t - \hat{y}_t)^2}{n}}$$

Mean Absolute Percentage Error (MAPE)

MAPE is a simple metric that allows for the comparison of accuracy across different models with varying time periods and observation counts.

$$MAPE = \frac{1}{n}\Sigma_{t=1}^{n}\left|\frac{(y_t - \hat{y}_t)}{y_t}\right| \cdot 100$$

IV. RESEARCH RESULTS AND DISCUSSION

4.1. Data Statistics and Description

Descriptive statistics provide a comprehensive overview of the data, including the mean, standard deviation, minimum, and maximum values for each variable. These indicators reflect the distributional characteristics and the degree of dispersion within the dataset. Such analysis is crucial for assessing variable fluctuations over time and identifying potential outliers that could bias the regression results. The statistical results describing the variables shown in Table 2 show that the variables in the model are not abnormal.

Google Trends index for the keywords "Visit to Vietnam", "Best time to visit Vietnam", "Place to visit Vietnam", "Vietnam best place to visit", "hanoi".

TABLE 2. Descriptive statistics of the data

Index	CPI (Index)	Interest rate (%)	Exchange rate (VND/USD)	FDI (billion USD)	Oil Prices (USD)	GDP (%)	Tourist Arrivals (Persons)
Medium	100.46	6.95	21.951	1.40	78.02	5.89	761,700
Median	100.30	6.50	22.642	1.30	74.88	6.06	629,348
Standard deviation	0.72	2.83	2.342	0.64	23.81	1.95	513,627
Smallest	98.46	4.00	15.930	0.30	26.35	-6.02	7,100
Largest	103.91	15.00	26.427	4.00	139.83	13.71	2,070,466

Source: Authors' calculation using Microsoft Excel.

TABLE 3. Description of collected data

Keywords / Time	"Visit to Vietnam (GT1)	"Best time to visit Vietnam" (GT2)	"Place to visit Vietnam" (GT3)	"Vietnam best place to visit" (GT4)	"hanoi" (GT5)
2008-01	7	8	7	0	40
2008-02	4	0	0	0	40
2008-03	4	0	0	0	39
…	…	…	…	…	…
2025-10	73	76	55	46	89
2025-11	90	88	60	51	100
2025-12	100	89	94	85	95

Source: Author group

4.2. Forecasting and analysis of results

The dataset only contains macroeconomic data.

TABLE 4. Measurement of accuracy on the Training Dataset

Parameter	Random Forest	XGBoost	Extra trees
R2_score	0.9839	0.9990	0.9341
MAE	39553.62	11316.65	90709.26
RMSE	65341.52	16372.37	132339.51
MAPE	34.31	7.16	90.76

Source: Authors' calculation using Python.

Based on training results from macroeconomic data, XGBoost is the best-performing model with a near-perfect R2_score and extremely low MAPE error, demonstrating highly effective tracking of economic fluctuations. Ranked second, Random Forest with an R2_score of 0.9839; however, its MAE and RMSE values are noticeably higher compared to the leading model. Meanwhile, Extra Trees yields the most modest results with a MAPE of 90.76%, reflecting a forecasting efficiency that is not yet fully compatible with the specifics of this dataset. Overall, XGBoost provides the highest predictive reliability due to its superior error minimization capability.

On the monitoring set, XGBoost maintained its leading position with an R2_score of 0.9057 and the lowest MAPE among the three models. However, compared to the almost perfect results in the training set, the significant drop in performance across all three models indicates signs of overfitting. This is because the models learned too deeply into the noise of the macroeconomic data instead of grasping the general patterns, leading to reduced accuracy in actual forecasting. Nevertheless, with the best error control capabilities, XGBoost remains the most reliable choice for application.

TABLE 5. Measurement of accuracy on the monitoring Dataset.

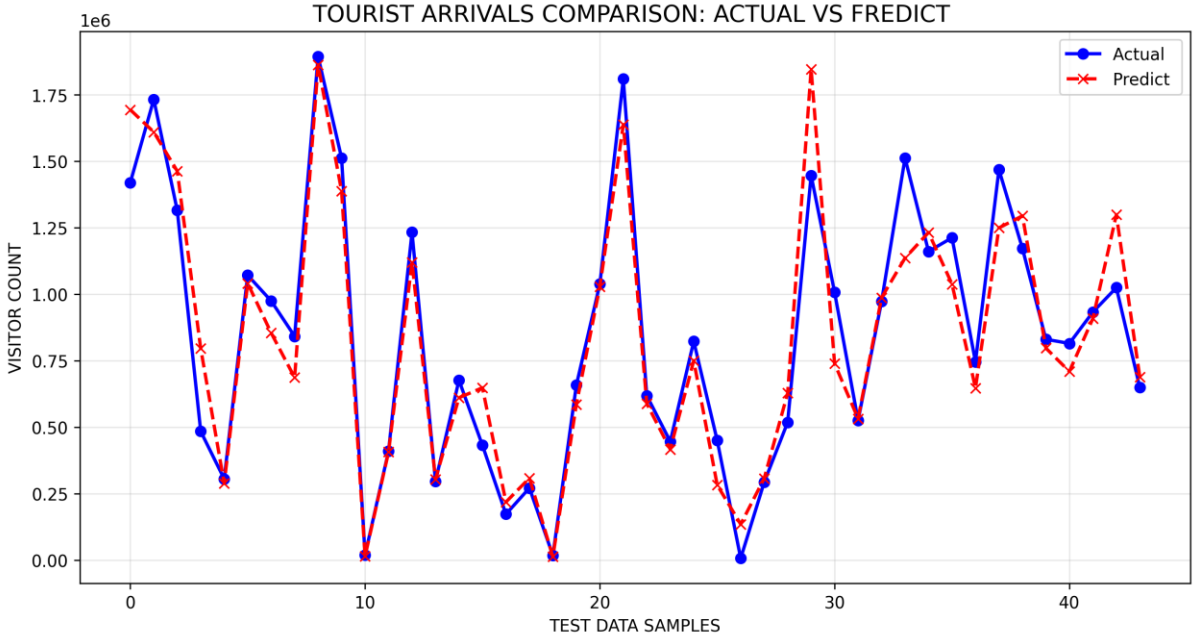
Parameter	Random Forest	XGBoost	Extra trees
R2_score	0.8844	0.9057	0.8193
MAE	112367.73	110618.78	142641.59
RMSE	167000.21	150892.42	208815.60
MAPE	79.55	54.94	186.56

Source: Authors' calculation using Python.

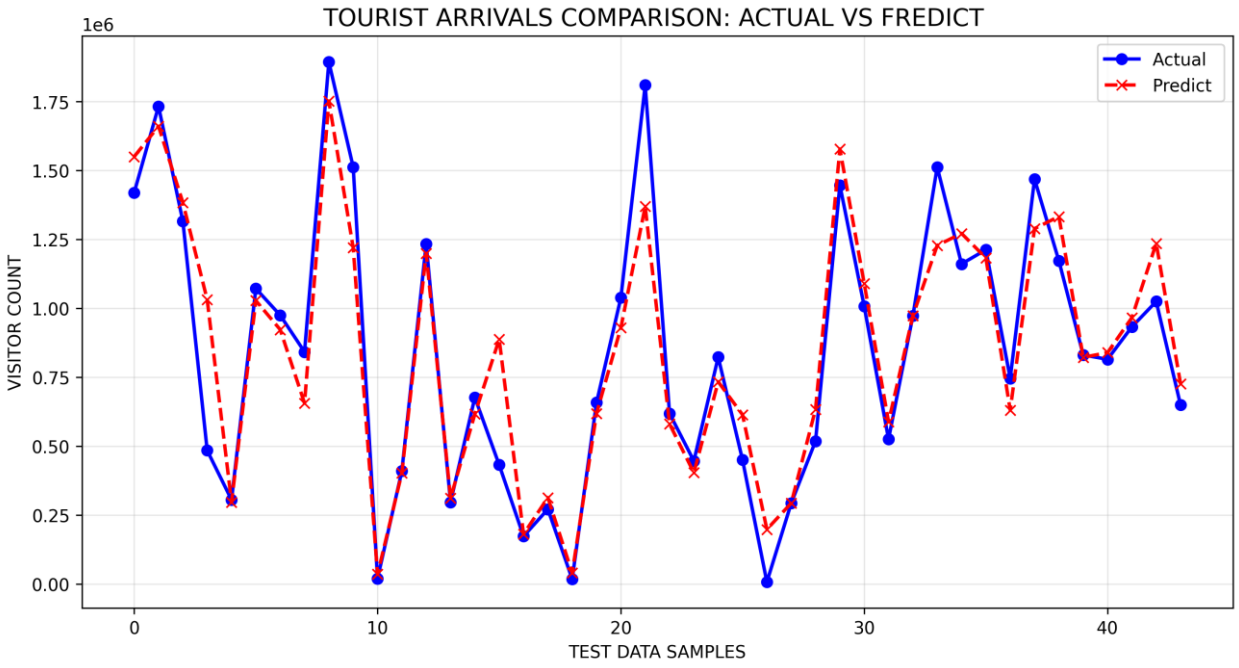
In summary, based on evaluations across both the training and monitoring sets, XGBoost maintains a clear performance advantage over Random Forest and Extra Trees. Although achieving near-perfect results on the training set, the performance decline and increased error on the monitoring set indicate that all three models exhibit signs of overfitting due to learning too deeply into the noise of the macro data. However, thanks to its superior error control capabilities despite this common limitation, XGBoost remains the optimal and most reliable choice.

All three models fairly accurately simulate tourist trends from macroeconomic data, with XGBoost performing best because of its higher sensitivity and its effectiveness in capturing turning points and peak fluctuations. Conversely, Random Forest and Extra Trees tend to be safer, underestimating peaks due to data smoothing mechanisms, making it difficult to capture sudden growth shocks. Overall, if the goal is the most accurate forecast for action planning, XGBoost is the optimal choice due to its ability to quickly reflect market changes. However, if a comprehensive view of long-term trends with less noise is needed, Random Forest will be a more reliable option.

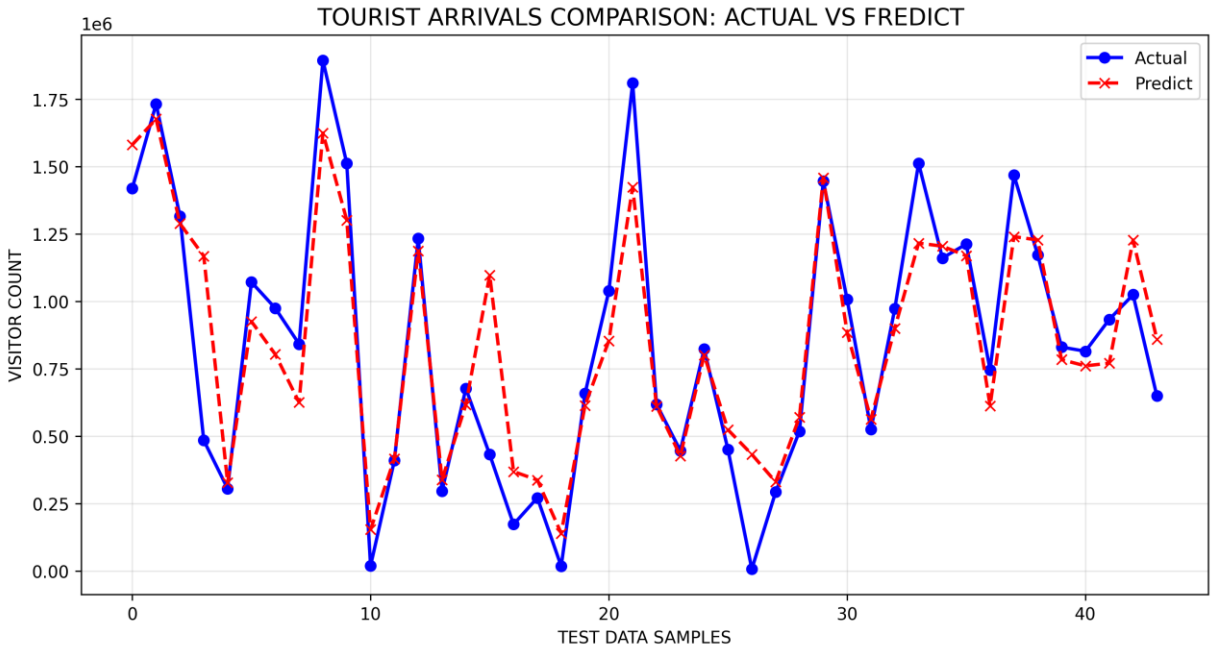
(a)



(b)



(c)



Note: (a) XGBoost model, (b) Random Forest model, (c) Extra Trees model
Figure 1: Graph comparing the predicted and actual tourist numbers of the model.
Source: Authors' calculation using Python.

Combining macroeconomic data and Google Trends

TABLE 6. Measurement of accuracy on the Training Dataset.

Parameter	Random Forest	XGBoost	Extra trees
R2_score	0.9889	0.9947	0.9674
MAE	35896.74	28024.98	67174.99
RMSE	54359.72	37617.99	93127.33
MAPE	7.47	12.16	41.28

Source: Authors' calculation using Python.

On the, all three models exhibited a strong fit, demonstrating good data capture capabilities. XGBoost achieved the best performance with the highest R2_score and the lowest error metrics (MAE and RMSE), suggesting minimizing bias. Random Forest was also a good choice, showing stability and particularly impressive with the lowest MAPE. Conversely, Extra Trees performed the weakest, recording high error rates compared to the other two algorithms. Overall, XGBoost and Random Forest delivered significantly

International Journal of Scientific Engineering and Science
Volume 10, Issue 2, pp. 68-76, 2026. ISSN (Online): 2456-7361

improved accuracy, whereas Extra Trees produced comparatively less satisfactory results.

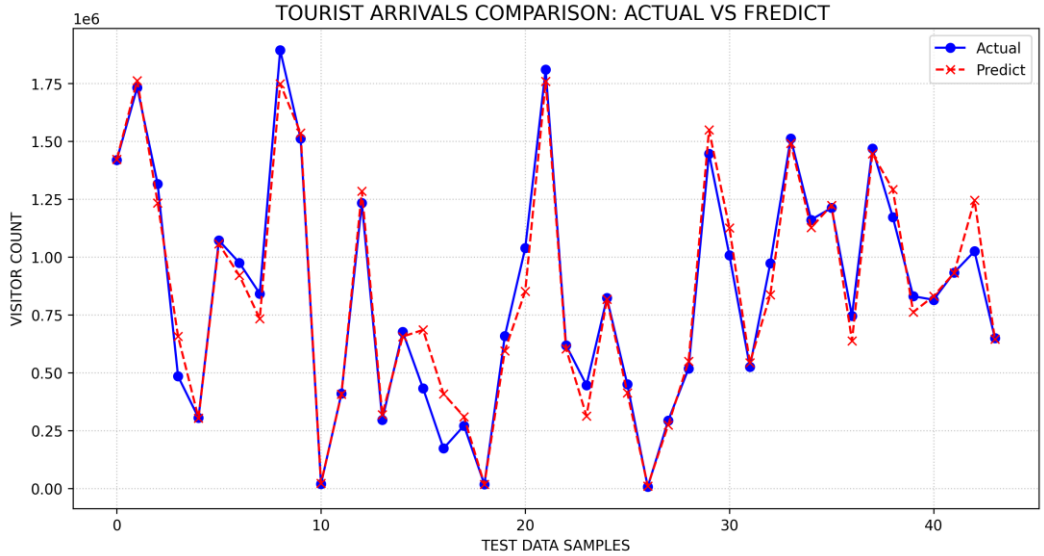
TABLE 7. Measurement of accuracy on the monitoring set

Parameter	Random Forest	XGBoost	Extra trees
R2_Score	0.9426	0.9637	0.9311
MAE	76546.54	63793.72	91425.84
RMSE	117715.79	93609.54	128948.69
MAPE	17.90	12.54	45.39

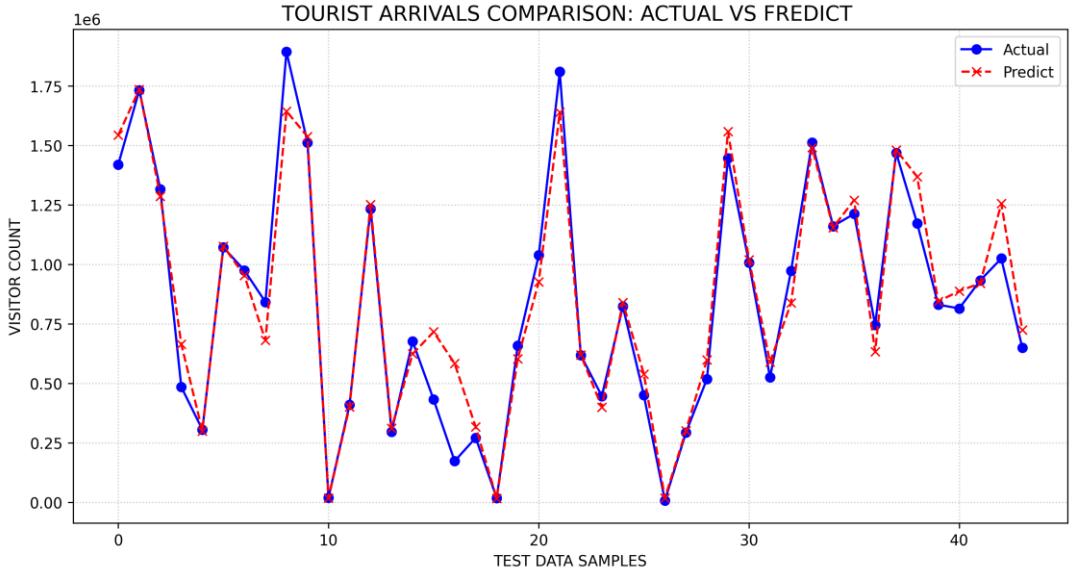
Source: Authors' calculation using Python.

On the monitoring set, the XGBoost model continued to assert its leading position with the best performance, achieving an R2_score of 0.9637 and the lowest MAPE. Following closely behind was Random Forest with a fit of 0.9426, while Extra Trees ranked last. Notably, although Random Forest closely followed and even had a better MAPE than XGBoost in the training set, the gap between the two models widened significantly in the monitoring set due to a decrease in accuracy. This divergence in results stems from overfitting, where the model learns too deeply into the noisy details of the training set, leading to poor generalizations on real-world data. While Random Forest and Extra Trees suffered from overfitting, XGBoost maintained stable performance thanks to its moderation mechanisms that controlled complexity and focused on the general patterns of the data. With only a slight and negligible increase in error and reliable forecasting results on new datasets, XGBoost is the optimal choice to ensure the accuracy of future forecasts.

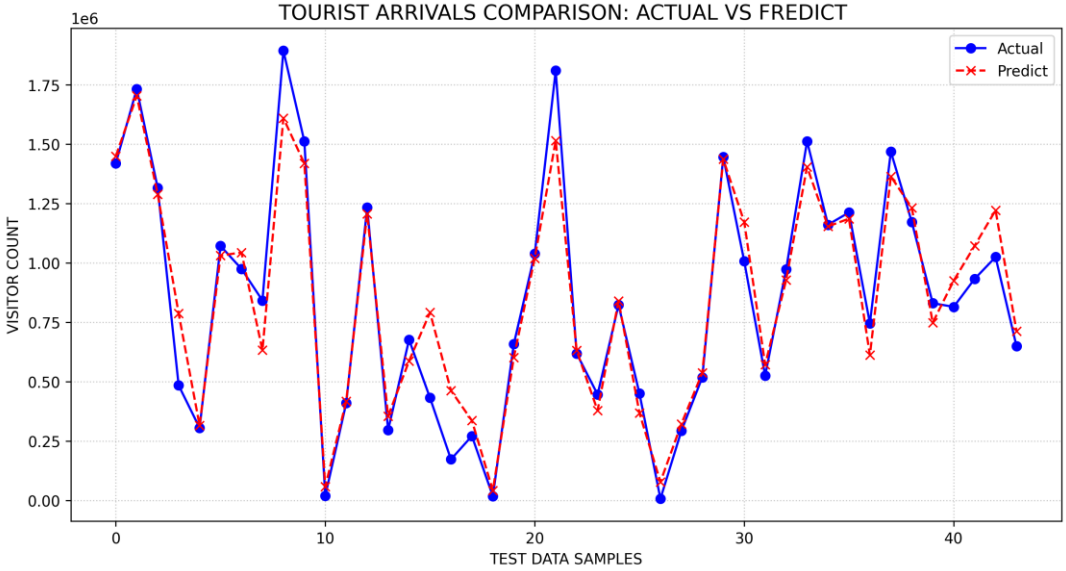
(a)



(b)



(c)



Note: (a) XGBoost model, (b) Random Forest model, (c) Extra Trees model
Figure 2: Graph comparing the predicted and actual tourist numbers of the model.
Source: Authors' calculation using Python.

Based on comparative charts, XGBoost demonstrates superior performance due to its ability to closely track extreme points and peaks of actual visitor volume fluctuations, exhibiting good generalization and being least affected by overfitting compared to other models. Meanwhile, Random Forest and Extra Trees reveal limitations, learning well but performing poorly in application, as the forecast line frequently deviates from the jump points on the monitored set, leading to large errors at times of high data volatility. Thus, with its stable R2_score and ability to fit the actual chart line most accurately, XGBoost represents the most appropriate model for improving forecasting accuracy in tourism demand analysis.

V. CONCLUSION

Experimental results show a significant difference in performance when changing the input data structure. Using only macroeconomic data makes models prone to overfitting, as evidenced by XGBoost's R2_score of 0.9990 on the training set but only 0.9057 on the monitoring set. This reflects the potential lack of responsive variables, causing Random Forest or Extra Trees models to often underestimate peak points and fail to catch sudden growth shocks.

A clear turning point emerged with the integration of Google Trends data, significantly reducing the MAPE of the best model to just 12.54%. In this scenario, XGBoost demonstrated superior performance with a stable R2_score of 0.9637 thanks to its ability to control complexity and more accurately capture turning-point dynamics relative to Random Forest. In summary, combining multidimensional data with the XGBoost algorithm not only helps to reduce overfitting but also creates a powerful forecasting tool that accurately reflects the rapid changes in the tourism market in practice.

REFERENCES

[1]. Athanasopoulos, G., Hyndman, R. J., Song, H., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822–844.
[2]. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
[3]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.
[4]. Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record, 88*(s1), 2–9.
[5]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.
[6]. General Statistics Office of Vietnam. (2020). *Vietnam welcomes record number of foreign visitors in 2019*. Statistical Publishing House. Accessed at 7:00p.m on 23/2/2026, from https://en.vietnamplus.vn/vietnam-welcomes-record-number-of-foreign-visitors-in-2019-post166291.vnp
[7]. General Statistics Office of Vietnam. (2025). *Socio-economic statistics of Vietnam*. Accessed at 3:00p.m on 2/2/2026, from https://www.gso.gov.vn/
[8]. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42.
[9]. Google. (2026). *Google Trends data for search terms "Visit to Vietnam," "Best time to visit Vietnam," "Place to visit Vietnam," "Vietnam best place to visit," and "Hanoi"* [Data set]. Accessed at 3:00p.m on 4/2/2026, from https://trends.google.com/trends/
[10]. Investing.com. (2026). *Exchange rate data for VND/USD and global crude oil prices*. Accessed at 4:00p.m on 1/2/2026, from https://vn.investing.com/
[11]. Ministry of Planning and Investment of Vietnam. (2026). *Foreign direct investment (FDI) data in Vietnam*. Accessed at 3:00p.m on 1/2/2026, from https://www.mpi.gov.vn/portal/Pages/Dau-tu-nuoc-ngoai.aspx
[12]. Ngo, V. S., Le, V. H., Thai, T. P., Hoang, T. L., & Vo, V. M. N. (2024). Enhancing tourism demand forecasting efficiency using Google Trends. *Hue University Journal of Science: Engineering and Technology, 133*(2A), 83–98.
[13]. Song, H., Li, G. (2009). *Tourism demand modelling and forecasting—A review of recent research*. Tourism Management, 29(2), 203–220.
[14]. Song, H., Witt, S. F. (2006). Forecasting international tourist flows. *Tourism Management, 27*(2), 214–224.

[15]. Trading Economics. (2026). *Interest rate and foreign direct investment indicators in Vietnam*. Accessed at 2:00p.m on 2/2/2026, from https://tradingeconomics.com/vietnam/
[16]. United Nations World Tourism Organization. (2010). *International recommendations for tourism statistics 2008*. United Nations. https://unstats.un.org/unsd/publication/Seriesm/SeriesM_83rev1e.pdf
[17]. Viet Nam National Authority of Tourism. (2026). *International tourism statistics*. Vietnam National Authority of Tourism. Accessed at 5:00p.m on 1/2/2026, from https://vietnamtourism.gov.vn/statistic/international
[18]. Vietnam National Authority of Tourism (2020). *Vietnam tourism annual report 2019.* Ministry of Culture, Sports and Tourism. Accessed at 8:00p.m on 23/2/2026, from https://vietnamtourism.gov.vn/
[19]. Vietnam National Authority of Tourism. (2020). *Vietnam tourism annual report 2019*. Accessed at 8:30p.m on 25/2/2026, from https://vietnamtourism.gov.vn
[20]. Vietstock. (2026). *Gross domestic product (GDP) growth data of Vietnam*. Accessed at 6:00p.m on 1/2/2026, from https://vietstock.vn/
[21]. Witt, S. F., & Witt, C. A. (1995). *Forecasting tourism demand: A review of empirical research*. International Journal of Forecasting, 11(3), 447–475.
[22]. Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). *Forecasting Chinese tourist volume with search engine data*. Tourism Management, 46, 386–397.