International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 14 Issue 7 ǁ July 2026 ǁ PP. 104-111

Decomposing the Driving Factors of Tourism Carbon Emissions: An Empirical Study Based on SHAP Machine Learning Analysis and Its Policy Implications

Wang Zhongchen
(Jinan University, Shenzhen Campus, Guangdong 518053, China)

Abstract: *Guided by the "3060" dual carbon goals, all industries have embarked on green transformation initiatives. Tourism, a pillar sector of the national economy, also needs to accelerate its transition to a sustainable, low-carbon development pathway. Traditional carbon accounting methods for the tourism industry are inadequate when applied to the tourism system—a complex nonlinear system driven by multiple intertwined factors. To address this challenge, this study develops a robust machine learning framework with high prediction accuracy and strong interpretability through empirical analysis, aiming to achieve precise carbon emission forecasting and dynamic attribution of carbon emission drivers for the tourism sector.Based on an open tourism regional carbon emission dataset from Kaggle, this paper establishes a multi-level machine learning analytical framework, and conducts systematic comparative experiments across six mainstream models including linear regression, random forest, and LightGBM. The hybrid model constructed via the Stacking ensemble method achieves the optimal performance, with an RMSE of only 62.08, demonstrating a substantial advantage over the conventional linear regression model (RMSE > 300). To interpret the black-box machine learning model, the SHAP method is adopted to quantify the marginal contribution of each driving factor to prediction results.The empirical results indicate that regional carbon emission intensity and regional GDP are the dominant drivers of regional tourism carbon emissions, whereas annual tourist arrivals are not a primary contributing factor, as they account for only a minor share of total carbon emissions.Accordingly, this study concludes that advancing the low-carbon transition of tourism hinges on qualitative improvements—such as reducing emission intensity and optimizing industrial structure—rather than simply restricting tourist volumes (a metric that is also over-emphasized in many basic tourism data models). The findings provide solid evidence for decoupling tourism economic growth from environmental carbon burdens. Furthermore, this research explores pathways for national economic development under the dual carbon goals within the service sector, confirming that the tourism industry can make significant contributions to national sustainable development and serve as a vital foundation for achieving carbon peaking and carbon neutrality targets, whose role deserves full recognition.*

Keywords：Tourism Carbon Emissions, Machine Learning, SHAP, Driving Factor Analysis, Carbon Neutrality

---

Date of Submission: 13-07-2026 Date of Acceptance: 24-07-2026

I. Introduction

1.1Research Background and Significance

Global climate governance has now entered a practical carbon emission reduction phase framed by the Paris Agreement. Against this backdrop, China has put forward the dual carbon goals of achieving carbon peaking before 2030 and carbon neutrality before 2060, setting a critical agenda for green transformation in China's economic and social development[1].As a widely connected and large-scale industry, tourism serves as a key driver of high-quality economic development, while also bearing a substantial carbon reduction obligation[2].Studies by the World Tourism Organization indicate that the global tourism industry accounts for approximately 8% of total anthropogenic carbon emissions. Given its energy consumption and pollution impacts, tourism is an essential sector for delivering carbon neutrality[3].To align with national strategic objectives and meet the sustainable development needs of the tourism industry itself, it is vital to accurately predict the low-carbon transition trend of tourism and decompose the core carbon emission drivers. This is a prerequisite for the high-quality development of tourism, as well as a key measure to support the delivery of China's 2030 carbon peaking and 2060 carbon neutrality targets.

1.2Literature Review

Academic research on carbon emission driving factors has formed a relatively mature framework. Two mainstream methodological approaches dominate traditional studies. The first is index decomposition analysis represented by the LMDI method[4],which decomposes total carbon emission changes into effects including energy structure, energy intensity and economic activity, and has been widely applied in attribution research[5].The second is regression analysis based on the STIRPAT model[6], which constructs econometric models to examine the elastic impacts of population, affluence, technology and other factors on environmental pressure.Nevertheless, these classical methods have inherent limitations. LMDI is fundamentally a descriptive accounting framework lacking statistical inference capacity and cannot support reliable forecasting[7]; The STIRPAT model and its extended forms generally assume a log-linear relationship among variables, making it difficult to capture the prevalent complex nonlinear relationships and multivariate interactions in real-world systems[8].

To overcome these limitations, machine learning methods have developed rapidly in recent years and achieved remarkable progress. Models such as random forest, support vector regression (SVR) and gradient boosting machine (GBM) achieve far higher prediction accuracy than conventional econometric models, primarily because they require no strict parametric assumptions and can effectively handle high-dimensional nonlinear problems.However, most existing relevant studies suffer from a critical flaw: they focus heavily on algorithm comparison while neglecting in-depth model interpretation. This restricts evidence-based analysis and hinders the translation of research findings into actionable policy recommendations[9].Fortunately, explainable artificial intelligence (XAI) techniques represented by SHAP (SHapley Additive exPlanations) provide a theoretically sound solution to unpack black-box models and quantify the marginal contribution of each driver to individual predictions[10], enabling the integration of high-precision forecasting and in-depth factor attribution[11].

1.3Research Innovations

Against this context, this study constructs an interpretable machine learning framework with strong predictive performance to investigate the overall driving mechanisms of tourism carbon emissions. The research workflow includes: building and comparing multiple machine learning models to select the optimal base models[12]; constructing an ensemble prediction model via the Stacking strategy; and conducting quantitative model interpretation using the SHAP method.

Two core innovations are outlined as follows:

① Methodological innovation: The Stacking ensemble model is systematically applied to tourism carbon emission forecasting for the first time. Grid search and cross-validation are adopted to optimize hyperparameters before model integration.

② Interpretability innovation: Leveraging the explanatory strength of the SHAP method, this study moves beyond traditional linear attribution analysis. It further examines nonlinear relationships and interactions between superficial variables (such as tourist arrivals) and structural factors including emission intensity and GDP, and identifies the key leverage points for tourism decarbonization through analyzing interaction patterns among influencing factors[13].This research aims to clarify the actual impact of variables such as tourist volume, facilitate the green development of the tourism economy, and offer a new perspective for decoupling tourism economic growth from carbon emissions.

II. Research Design and Methodology

2.1Data Sources and Variable Description

The empirical data used in this study is derived from the publicly available Tourism-Economy-Carbon Dynamics Dataset on Kaggle. High-quality tourism carbon emission samples in the dataset reflect the interactions among tourism activities, carbon emissions and regional economic growth. The dataset contains 2,500 observations of tourism regions covering the period 2016–2019. Besides basic tourism indicators such as tourist arrivals, it includes multi-dimensional environmental data including carbon dioxide emissions from transportation, energy consumption in accommodation, and emissions from tourism services. Regional GDP, emission intensity and economic growth rate are also incorporated into the dataset, laying a foundation for constructing an accurate carbon emission prediction model and subsequent analysis of driving factors.

2.2Research Framework and Model Construction

To achieve accurate prediction and in-depth attribution of tourism carbon emissions, an integrated "prediction-explanation" machine learning framework is developed following standardized technical procedures. The implementation steps are as follows. First, in the data preprocessing stage, total emissions from transportation, accommodation and tourism services are summed as the target variable (Y) Total Emissions, while other economic and tourism indicators are taken as input features (X). The dataset is randomly split into a training set and a test set at an 8:2 ratio for model training and generalization validation. Second, in the model training and evaluation stage, three types of models are constructed and compared sequentially: a basic reference prediction model using traditional linear regression, core ensemble models with strong nonlinear fitting capability, and a Stacking ensemble model integrating all candidate base models. Finally, in the interpretable analysis stage, the optimal Stacking model is decomposed using the SHAP method to analyze the internal interaction mechanisms of various factors.

The model system constructed in this study includes multiple levels from simple to complex. The benchmark model is classic multiple linear regression, which assumes a linear relationship between dependent and independent variables, with the model specification shown below:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$$

This model provides a baseline reference for evaluating performance improvements of more complex models. The core ensemble models adopt two mainstream tree-based algorithms. The first is Random Forest (RF), which constructs a large set of weakly correlated decision trees via Bagging (bootstrap aggregating) and feature randomization, and averages predictions across all trees for regression tasks to reduce model variance and effectively prevent overfitting. The second is LightGBM, an efficient implementation of gradient boosting decision trees that iteratively fits the negative gradient of the loss function. Its core iterative formula is as follows:

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h_m(\mathbf{x})$$

LightGBM significantly accelerates training through innovative gradient-based one-side sampling and exclusive feature bundling techniques. Finally, a Stacking ensemble model is built to further improve prediction accuracy. It adopts a two-stage structure. In the first stage, heterogeneous models including Random Forest and LightGBM are used as base learners, and cross-validation is applied on the training set to generate secondary prediction features. In the second stage, these secondary features are used as new inputs to train a meta-learner for final prediction. Lasso regression is selected as the meta-learner, whose optimization objective is to minimize the loss function with an L1 regularization term:

$$\min_{\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

The L1 regularization term automatically shrinks the weights of base learners with low contribution to zero, enabling automatic base model selection and enhancing the stability and interpretability of the ensemble model.

2.3Model Evaluation and Interpretable Analysis

A comprehensive evaluation framework combining performance assessment and interpretable analysis is established to identify the optimal prediction model and explore its internal decision-making mechanism. For regression tasks, the Root Mean Squared Error (RMSE) is adopted as the core evaluation metric. RMSE directly measures the deviation between predicted values and actual values, sharing the same unit as the target variable for intuitive interpretation. A smaller RMSE value indicates higher model prediction accuracy. The formula is defined as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

However, pursuing pure high accuracy often results in an uninterpretable black-box model[10].To resolve this issue, the advanced explainable artificial intelligence method SHAP (SHapley Additive exPlanations) is applied. It interprets the predicted value for each individual sample based on the baseline prediction (the mean value across the whole dataset), attributes prediction deviations to each input variable, and quantifies the marginal contribution of each variable in shifting the prediction from the global mean to the specific sample value. This study uses SHAP summary plots for result interpretation, which convey three layers of information simultaneously: (1) Global feature importance ranking, ordered from left to right according to the descending mean absolute SHAP values; (2) Direction and distribution of feature effects on predictions: a point located on the left side of the horizontal axis indicates a negative effect on the corresponding sample, while a point on the right side indicates a positive effect; (3) Magnitude of feature values, represented by color gradient (darker colors denote higher feature values, lighter colors denote lower feature values). Within this three-layer interpretation framework, higher feature values generally correspond to greater impacts. The causal relationship can be clearly identified from color distribution and SHAP values; as shown in the plot, feature values gradually decrease regardless of color shade.

III. Results and Analysis

3.1Comparison of Model Prediction Performance

Models including KNN, support vector machine, linear regression, ElasticNet, Lasso regression, decision tree, LightGBM and random forest are constructed using the basic dataset, and different models are applied to make predictions on the test set. Based on the prediction results and root mean squared error values, the optimal model is selected. Table 1 presents the prediction results of tourism carbon emissions across different models, revealing substantial performance differences among them.

First, traditional linear models (including linear regression, ElasticNet and Lasso regression) exhibit large prediction errors, with RMSE values generally exceeding 300. This indicates that linear models cannot adequately capture the complex nonlinear relationships between tourism carbon emissions and their driving factors. In contrast, nonlinear models such as XGBoost and LightGBM achieve much higher prediction accuracy. After hyperparameter tuning via five-fold cross-validation, a single decision tree model achieves an RMSE of only 96.63, representing a roughly 70% reduction in error compared with linear models. This confirms that nonlinear models are necessary for accurate fitting.

Subsequently, based on ensemble learning principles, further error reduction is conducted for the random forest (RMSE ≈ 65.98) and LightGBM (RMSE ≈ 65.26) models (see Table 1), with both models achieving RMSE values between 65 and 67. This demonstrates the advantages of ensemble algorithms in improving model stability and prediction accuracy. Among all tested models, the optimally parameterized Stacking model achieves the best overall performance, with the lowest RMSE of 62.08 on the test set.

Model Name	Parameter Processing Method	Maximum Error	RMSE
KNN	Default parameters	1998.1	479.517
Support Vector Machine	Default parameters	1708.3	433.3665
Linear Regression	Default parameters	1373.5	319.976
Support Vector Machine	Optimal parameters via 5-fold cross-validation	1266.2	302.1075
ElasticNet	Default parameters	1293.7	301.6285
Lasso Regression	Default parameters	1266.1	301.401
Decision Tree	Default parameters	471.9	101.0371
Decision Tree	Optimal parameters via 5-fold cross-validation	448.5	96.631016
LightGBM	Optimal parameters via 5-fold cross-validation	272.7	66.7783
Random Forest	Optimal parameters via 5-fold cross-validation	276.5	65.9843
Random Forest	Default parameters	258.6	65.8305
LightGBM	Default parameters	284.1	65.2559
Simple Stacking Model	**Optimal model parameters**	**155.0**	**62.9288**
Weighted Stacking Model	**Optimal model parameters**	**187.3**	**62.633269**
Optimized Stacking Model	**Optimal model parameters**	**164.7**	**62.083856**

Table 1. Performance Comparison of Different Prediction Models

Figure 2 shows the predicted-versus-actual scatter plot and residual plot for the Stacking model on the test set. The left scatter plot indicates that most data points (blue dots) cluster closely around the red dashed perfect-prediction line, showing strong consistency between predicted and actual values. The model achieves a coefficient of determination of 0.9362, meaning it explains more than 93% of the variance in total tourism carbon emissions in the test set, reflecting excellent goodness of fit.

The right residual plot is used to diagnose model error distribution. Most residuals (green dots) are randomly and evenly distributed around the zero line, indicating no systematic overestimation or underestimation bias. However, residual dispersion slightly increases as predicted values rise, suggesting mild heteroskedasticity. This means the model has greater predictive uncertainty for high-emission samples and occasionally produces large errors, with residuals exceeding 400 for a small number of individual observations..

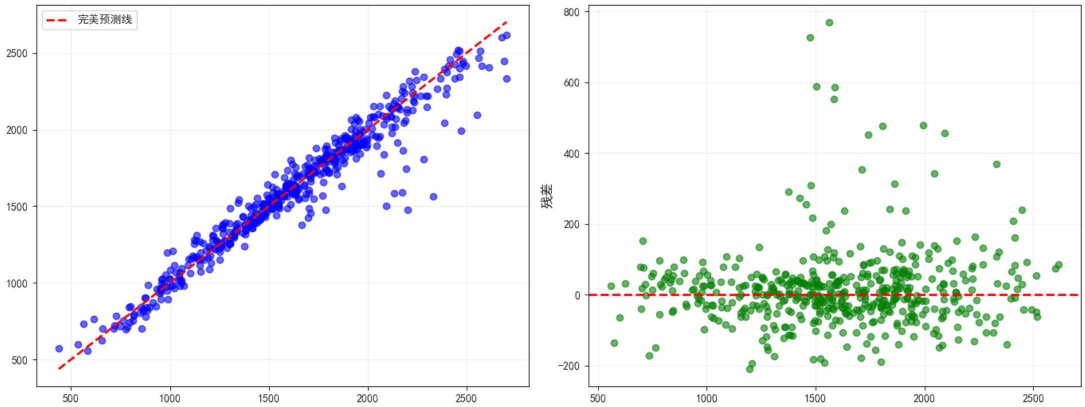

Figure 2. Predictive Performance and Residual Diagnostics of the Optimal Stacking Model

In summary, the Stacking framework intelligently integrates predictions from multiple strong base learners via a meta-learner, combining their strengths to achieve superior accuracy beyond any single model. Accordingly, this study adopts the optimally parameterized Stacking model as the baseline model for subsequent interpretable analysis of key driving factors, to ensure the reliability and depth of research conclusions.

3.2 SHAP Analysis of Key Carbon Emission Driving Factors

Having verified the outstanding predictive performance of the Stacking ensemble model, this section applies the SHAP method to interpret the model and identify the key driving factors of tourism carbon emissions and their underlying mechanisms.

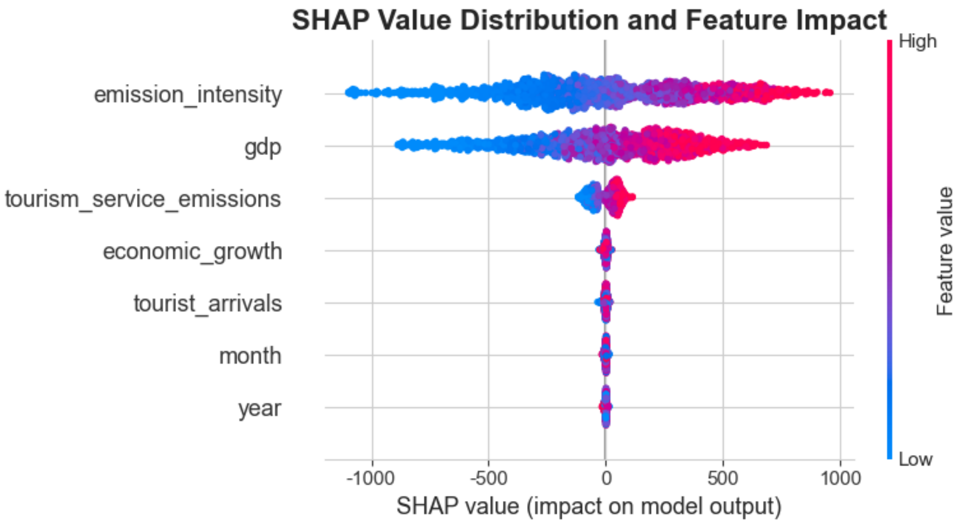

Figure 3. SHAP Summary Plot of Feature Importance and Impact Distribution

3.2.1Global Feature Importance Ranking

The SHAP summary plot in Figure 4 illustrates the overall impact direction and magnitude of all driving factors on total tourism carbon emissions. The results clearly show that emission intensity and regional GDP are the two dominant factors, with far higher mean absolute SHAP values than other variables, followed by tourism service emissions.

Notably, tourist arrivals, which are conventionally considered a major influential factor, exhibit relatively low global feature importance in the model and have a negligible overall effect. In addition, the accuracy of some baseline models is limited, which can inform future model development for improved prediction performance..

3.2.2Analysis of Core Driving Factors

Further analysis is conducted to examine the influence patterns of the core driving factors. For emission intensity, the most important variable, the pattern in Figure 4 is distinct: high emission intensity values (red dots) mostly correspond to positive SHAP values, while low emission intensity values (blue dots) correspond to negative SHAP values..

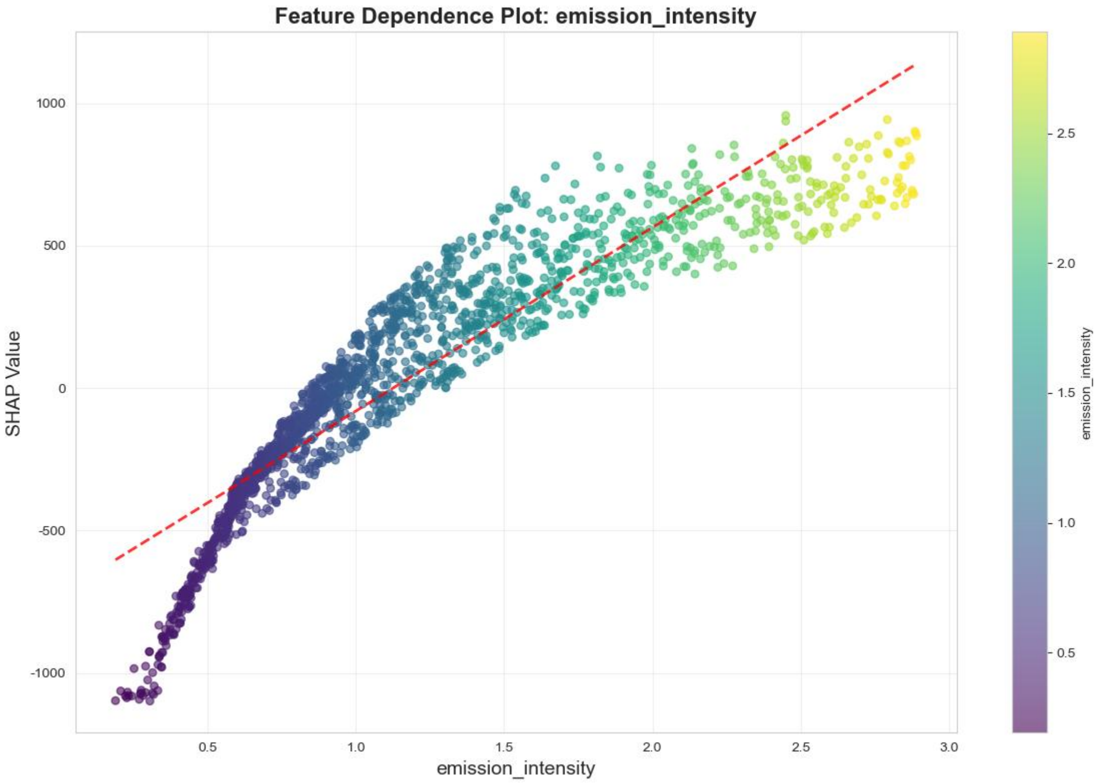

Figure 4. SHAP Dependence Plot for the Core Driving Factor 'Emission Intensity'

The SHAP dependence plot in Figure 5 reveals a strong approximate positive linear relationship between emission intensity and its SHAP values, indicating that each unit increase in the carbon intensity of the regional economy leads to a predictable and significant rise in tourism carbon emissions. For the second dominant factor, regional GDP, Figure 4 also shows a clear positive correlation: higher regional GDP is associated with a stronger upward driving effect on carbon emissions.

Figure 6 is a SHAP waterfall plot illustrating the multivariate drivers of tourism carbon emissions, using Sample 425 as an example. The model baseline value is 1579.73. Since the GDP value (798.45) of Sample 425 is lower than the average, it produces a strong negative effect of -541.15 on the final prediction. However, the sample has a relatively high emission intensity (2.011) and tourism service emissions (633.308), which generate strong positive effects of +489.3 and +67.09 respectively. Additional minor effects from other variables contribute a total effect of +0.63, shifting the baseline value upward to the final predicted value of 1604.82. This demonstrates how the model generates individualized predictions for each sample by aggregating the marginal effects of all features.

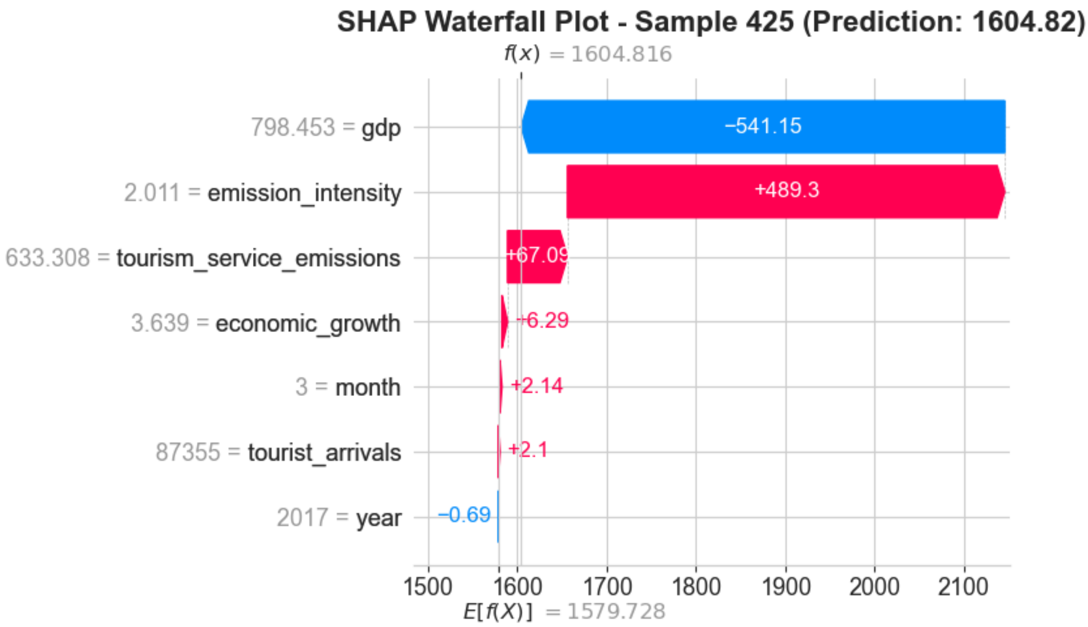

Figure 5. SHAP Waterfall Plot for a Single Prediction (Example: Sample 425)

IV. Discussion and Policy Implications

4.1Discussion of Research Findings

The results show that structural indicators such as regional economic emission intensity and total economic output have stronger predictive power for tourism carbon emissions than tourist arrivals themselves. This indicates that tourists are not the direct producers of tourism carbon emissions, but rather triggers of regional economic activities. The environmental footprint left by each tourist depends on whether they use carbon-intensive transport modes, the price of tourism services consumed, and the embodied carbon content of these services. Apart from travel modes, which can be directly controlled by tourists, most carbon emissions are determined by the local industrial structure and energy consumption patterns.

Accordingly, the model results reveal that the overall carbon emission level of regional tourism is not determined simply by the number of tourists, but by the quality of the economic system supporting tourism operations. This finding creates an insightful dialogue with existing studies that focus primarily on quantifying the environmental impact of tourist volumes. This study does not deny the correlation between tourism activities and carbon emissions, but argues that using tourist numbers as a single indicator oversimplifies the complex transmission mechanism of tourism-related carbon emissions.

The study confirms that local development patterns and quality constitute key mediating factors linking tourist behavior to environmental impacts. Recognizing this core mechanism establishes a clear logical foundation for designing targeted and effective tourism decarbonization strategies, which also helps improve the theoretical framework of tourism carbon emission research.

4.2Policy Implications

SHAP analysis identifies carbon emission intensity as the most influential driving factor, which means reducing carbon emissions per unit of GDP should be the core policy priority. Specific measures are proposed as follows: enforce strict green and energy-saving building codes and carry out energy retrofitting for existing buildings on a large scale; accelerate the development of green transport, using fiscal subsidies and road priority policies to expand public transport such as electric buses, rail transit and BRT that connect airports, railway stations, docks, urban areas and major tourist attractions. Develop smart tourism management based on big data technologies: construct regional smart tourism platforms powered by big data, 5G and the Internet of Things, digitally collect and analyze massive passenger and vehicle flow data, predict passenger flow patterns at tourism transport hubs, pre-emptively allocate resources and manage information flows to reduce energy waste caused by traffic congestion and information asymmetry.

GDP is the second major driving factor, and current economic growth is significantly carbon-locked. Breaking this carbon lock-in requires transforming the pattern of economic growth. First, vigorously develop low-carbon tourism formats. Use fiscal incentives, financing support and preferential land policies to prioritize eco-tourism, cultural heritage tourism, rural wellness tourism and other low-impact tourism models, and increase their overall share. Conduct strict carbon impact assessments for new high-carbon tourism projects such as large theme parks, indoor ski resorts and golf courses, and appropriately guide capital allocation.

In summary, under the dual carbon strategy, China's tourism decarbonization policy needs to shift from scale control to data-driven quality regulation. The model results confirm that economic emission intensity exerts a greater direct effect on carbon emissions than tourist arrivals. Restricting tourist numbers alone cannot achieve meaningful emission reductions and may even harm economic vitality. To achieve both tourism economic growth and green development, policymakers must improve the overall quality of economic operations and promote high-quality socioeconomic development through industrial structure optimization.

V. Conclusions and Future Research

Against the backdrop of China's dual carbon strategic goals requiring the green and sustainable development of tourism, this study constructs a machine learning framework for tourism carbon emission prediction and attribution analysis to address the urgent need for accurate carbon forecasting and driver identification. The optimal Stacking ensemble model is selected to precisely capture tourism carbon emission trends, and the SHAP interpretable method is further applied to unpack the black-box model and identify the dominant carbon emission drivers in terms of feature importance and influence mechanisms.

The core conclusions are as follows: (1) Regional tourism carbon emissions are mainly determined by economic emission intensity and regional GDP, rather than tourist arrivals. (2) Tourist arrivals are not a primary leading indicator for carbon forecasting. This challenges the conventional approach of directly limiting tourist volumes as the main decarbonization measure, and highlights the importance of quality improvement and efficiency enhancement for the low-carbon transformation of tourism.

This study has several limitations. First, the analysis is based on aggregate macro panel data. Future research may adopt micro-level data from cities, enterprises, hotels, airlines and individual tourists to achieve more precise carbon source accounting. Second, the current model does not explicitly incorporate specific environmental policies. Quantifying the actual emission reduction effects of targeted climate policies represents a promising future research direction. Furthermore, the combined Stacking ensemble and SHAP interpretable analysis framework can be generalized and applied across different regions and countries, to reveal geographical heterogeneity in tourism carbon emission drivers and provide refined decision support for sustainable tourism development worldwide.

References

[1] 胡鞍钢. 中国实现 2030 年前碳达峰目标及主要途径 [J]. 北京工业大学学报(社会科学版), 2021, 21(03): 1-15.
[2] 唐承财, 查建平, 章杰宽, et al. 高质量发展下中国旅游业"双碳"目标：评估预测、主要挑战与实现路径 [J]. 中国生态旅游, 2021, 11(04): 471-97.
[3] HALL C M. Constructing sustainable tourism development: The 2030 agenda and the managerial ecology of sustainable tourism [J]. Journal of Sustainable Tourism, 2019, 27(7): 1044-60.
[4] LIU L C, FAN Y, WU G, et al. Using LMDI method to analyzed the change of China's industrial CO2 emissions from final fuel use:: An empirical analysis [J]. Energy Policy, 2007, 35(11): 5892-900.
[5] MA X J, WANG C X, DONG B Y, et al. Carbon emissions from energy consumption in China: Its measurement and driving factors [J]. Sci Total Environ, 2019, 648: 1411-20.
[6] YORK R, ROSA E A, DIETZ T. STIRPAT, IPAT and ImPACT: analytic tools for unpacking the driving forces of environmental impacts [J]. Ecological Economics, 2003, 46(3): 351-65.
[7] XU S C, HE Z X, LONG R Y. Factors that influence carbon emissions due to energy consumption in China: Decomposition analysis using LMDI [J]. Appl Energy, 2014, 127: 182-93.
[8] FAN Y, LIU L C, WU G, et al. Analyzing impact factors of CO2 emissions using the STIRPAT model [J]. Environmental Impact Assessment Review, 2006, 26(4): 377-95.
[9] LINARDATOS P, PAPASTEFANOPOULOS V, KOTSIANTIS S. Explainable AI: A Review of Machine Learning Interpretability Methods [J]. Entropy, 2021, 23(1).
[10] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead [J]. Nature Machine Intelligence, 2019, 1(5): 206-15.
[11] GILPIN L H, BAU D, YUAN B Z, et al. Explaining Explanations: An Overview of Interpretability of Machine Learning; proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA), Turin, ITALY, F Oct 01-04 2018, 2018 [C]. 2018.
[12] GANAIE M A, HU M H, MALIK A K, et al. Ensemble deep learning: A review [J]. Engineering Applications of Artificial Intelligence, 2022, 115.
[13] MURDOCH W J, SINGH C, KUMBIER K, et al. Definitions, methods, and applications in interpretable machine learning [J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(44): 22071-80.