

Forecast of Chinese pet population and Global Pet Food Demand Based on RF-MLR-ARIMA and GB-XGBoost

Yining Gao*

Haihua College, Liaoning Normal University, Dalian 116400, China.

* Corresponding author: Yining Gao (Email: 18839093704@163.com)

Abstract: With the global economic development, the acceleration of urbanization and the wide popularization of the "pet companionship" concept, the pet industry has grown from a traditional niche field into an emerging economic sector with both market potential and development challenges. Accurately analyzing the trends of core industry indicators has become a key prerequisite for supporting the sustainable development of the industry and the strategic decision-making of enterprises. This paper constructs a comprehensive prediction model of random forest - multiple linear regression - ARIMA, and concludes that the number of pet cats in China will continue to increase in the next three years, while the number of pet dogs will tend to stabilize. After the dimensionality reduction of the indicators, the gradient boosting regression - XGBoost model was adopted for analysis. The results show that the United States dominates the global demand for pet food, while the Chinese market demonstrates strong growth. This research can provide strong academic support for the trend analysis of the pet industry and the formulation of sustainable development strategies.

Keywords: Pet Industry, Pet Population Prediction, Pet Food Demand, Integrated Prediction Model, Gradient Boosting Regression - XGBoost.

1. Introduction

In recent years, global economic growth and the popularization of the concept of "pet companionship" have promoted the rapid development of the pet industry. The scale of the Chinese market continues to expand, but it also faces challenges such as demographic changes and intensified market competition. Accurate analysis of industry trends has become a key demand for sustainable development [1][2].

Scholars at home and abroad have carried out targeted research on the prediction of the pet industry. Li Juan et al. [1] constructed the ARIMA-XGBoost hybrid model to predict the size of the pet market in China. By using data from 2016 to 2022, the prediction error rate was controlled within 3.2%, which was significantly better than that of a single model. Zhang Siyu et al. [2] analyzed the global demand for pet food through gradient boosting regression and found that the contribution rate of per capita disposable income reached 37.6%, but the international market transmission effect was not taken into account. Wang Zihan et al. [3] used a multi-factor coupling model to predict the number of pets in China, pointing out that for every 1% increase in the urbanization rate, the number of pets would grow by 2.1%. Liu Chang et al. [4] used multiple linear regression to verify the inhibitory effect of tariff policies on the import and export of pet supplies, but their data did not cover international policy adjustments after 2022. Chen Xi et al. [5] screened out five core indicators of the pet industry through factor analysis, providing a basis for the selection of modeling variables. In foreign studies, Liu et al. [6] used the ARIMA-random forest hybrid model to predict the number of pets in China, reducing the short-term prediction error to less than 4.8%. Wang et al. [7] analyzed the global demand for pet food based on the XGBoost algorithm and confirmed that the market share of the United States has always exceeded 42%. Smith et al. [8] compared the industry trends of the United States and China through multi-model integration and proposed that the

average annual growth rate of pet food demand in China reached 12.3%. Kim et al. [9] utilized machine learning models to analyze the influencing factors of pet food exports and found that every 1% fluctuation in the exchange rate would trigger a 0.8% reverse change in export volume. Brown et al. [10] found through gradient boosting regression that the urbanization rate is significantly positively correlated with the pet-keeping rate; Davis et al. [11] pointed out through panel data analysis that policy uncertainty would reduce the industry growth rate by 1.5 to 2 percentage points. Wilson et al. [12] used a time series model to verify that the elasticity coefficient between the number of pets and medical demand was 0.78.

In general, the existing research has three shortcomings: first, single ARIMA model is sensitive to market structure mutation, and simple combination model is difficult to capture multi-dimensional nonlinear association [1][6][7]. Second, research mostly focuses on a single link such as "quantity prediction" or "demand analysis", lacking a full-chain linkage analysis of "pet quantity - food demand - production and export" [2][9]; Third, the policy impact is mostly qualitative description or single variable analysis, and the spillover effect of international policies is not fully considered [4][11]. For this purpose, based on multi-source data from 2019 to 2023, this paper constructs the Random Forest-MLR-ARIMA ensemble model to predict the number of pet cats and dogs in China, and uses the Gradient Boosting-XGBoost model to describe the global demand for pet food. Further achieve the prediction of production exports and the quantification of policy impacts.

2. Model

2.1. Random Forest Multiple Linear Regression - ARIMA Integrated Model

2.1.1. Model design background and objectives

In view of the coexistence of the three characteristics of

"multi-factor nonlinear association, linear driving relationship, and time series fluctuation" in the prediction of the number of pet cats and dogs in China (for example, the number of pet cats and dogs is affected by multi-dimensional economic policy indicators such as per capita GDP and total tax rate, and shows annual time series law), a single ARIMA or regression model is difficult to fully meet the complex modeling requirements. For this purpose, a three-level ensemble model of Random Forest-MLR-ARIMA is constructed: The Random Forest was used to screen key

features, MLR quantified linear driving relationships, and ARIMA corrected time sequence residuals to form a prediction framework with complementary functions, providing model support for short - to medium-term pet population prediction (3 years).

2.1.2. The establishment of the Random Forest model

This article first uses Random Forest to extract indicators closely related to the pet industry from the dataset, as shown in the table below.

Table 1. The number of pets selected by random forest affects the characteristics and importance

Feature name	Feature category	Importance value (Cat)	Importance value (Dog)	Screening results
Total tax rate(%)	Policy and economics category	0.18	0.13	Reserve
The total output value of pet food in China (in billions of yuan)	Industry economics category	0.12	0.11	Reserve
Gini coefficient	Social structure category	0.10	0.10	Reserve
Per capita GDP (in US dollars)	Macroeconomic category	0.08	0.10	Reserve
Rural population (in ten thousand)	Population structure category	0.07	0.09	Reserve
Food production index	Industrial category	0.06	0.07	Eliminate
Other characteristics	-	<0.05	<0.05	Eliminate

Based on the analysis of Table 1, the values highlight the relative significance of the pet market and its associated indicators. Notably, the total tax rate exhibits an importance score of approximately 0.18, underscoring its substantial impact on the pet market. Conversely, the food production index holds a comparatively lower importance score of 0.07, indicating a relatively modest influence. Meanwhile, the remaining indicators, with importance scores clustering around 0.09, suggest a moderate yet noteworthy contribution to the dynamics of the pet market.

Random Forest is an ensemble learning algorithm based on decision trees, commonly used for classification and regression tasks. It improves the accuracy and robustness of the model by constructing multiple decision trees and determining the final prediction result through majority voting (for classification) or averaging (for regression).

(1) Sample selection.

The first step in constructing a Random Forest is sample selection. From the original training set S containing N samples, the bootstrap method is used to generate m samples, with the sample set denoted as S_m . The sample selection formula for each tree can be expressed as:

$$S_m = \{x_i \mid i \in \text{Random Sample of Size } m \subseteq S\} \quad (1)$$

(2) Feature selection.

When splitting at each node, features are randomly selected. For example, m features are chosen ($m < M$, where M is the total number of features), and the best feature for splitting is selected based on certain criteria (such as information gain, Gini index, etc.).

$$X_f = \{f_i \mid i \in \text{Random sample of Size } m \subseteq \{1, 2, \dots, M\}\} \quad (2)$$

(3) Construction of decision trees.

For each decision tree T_j , the final model for tree construction can be expressed as:

$$Y_j = T_j(X) \quad (3)$$

where Y_j is the output of the j -th tree, and X represents the input features.

(4) Final prediction integration.

For classification tasks: Use the majority voting method (mode) as the final prediction:

$$\hat{Y} = \text{mode}(Y_1, Y_2, \dots, Y_J) \quad (4)$$

where J is the total number of trees.

For regression tasks: Use the average of the predictions from all trees as the final prediction:

$$\hat{Y} = \frac{1}{J} \sum_{j=1}^J Y_j \quad (5)$$

2.1.3. Establishment of Multiple Linear Regression

Multiple linear regression is a statistical method used to study the linear relationship between a dependent variable (target variable) and multiple independent variables (predictor variables). The goal of multiple linear regression is to find a set of regression coefficients that minimize the error between the predicted values and the actual values, thereby establishing a predictive model.

(1) Basic form of multiple linear regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon \quad (6)$$

where: y : Dependent variable (target variable); $x_1, x_2, \dots, x_n$: Independent variables (predictor variables); β_0 : Intercept term, representing the value of the dependent variable when all independent variables are 0; $\beta_1, \beta_2, \dots, \beta_n$: Regression coefficients, indicating the impact of each independent variable on the dependent variable; ϵ : Random error term, representing the part of the model that is not explained.

(2) Model matrix form

Multiple linear regression can be expressed in matrix form as:

$$y = X\beta + \epsilon \quad (7)$$

where: y : $n \times 1$ dependent variable vector, representing the target values of n samples; X : $n \times (p + 1)$ design matrix, containing n samples, p independent variables, and a constant term; β : $(p + 1) \times 1$ regression coefficient vector; ϵ : $n \times 1$ random error vector.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad (8)$$

(3) Objective function

Multiple linear regression estimates the regression coefficients β by minimizing the Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \|y - X\beta\|^2 \quad (9)$$

(4) Solution using least squares method

By minimizing the objective function, the optimal solution for the regression coefficients is obtained as:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (10)$$

where: $X^T X$: Covariance matrix of the independent variable matrix; $(X^T X)^{-1}$: Inverse of the covariance matrix; $X^T y$: Inner product of the independent variable matrix and the dependent variable vector.

(5) Prediction formula

Using the regression coefficients $\hat{\beta}$, new input data can be used for prediction:

$$\hat{y} = X\hat{\beta} \quad (11)$$

2.1.4. ARIMA Model

The ARIMA (Auto-Regressive Integrated Moving Average) model is a type of time series prediction model used to capture trends, seasonality, and random fluctuations in data. It is widely applied in the fields of economics, finance, and statistics. ARIMA combines three mechanisms: Auto-Regressive (AR), Integrated (I), and Moving Average (MA). This enables the model to construct predictions for both stationary and non-stationary time series data.

The Three Core Components of ARIMA:

(1) AR (Auto-Regressive):

Uses the past values of the time series itself to make predictions, relying on the autocorrelation of the series.

(2) I (Integrated):

Transforms non-stationary series into stationary ones through differencing.

(3) MA (Moving average):

Models the random fluctuations of the series by utilizing past prediction errors to create a smoother time series.

The AR model assumes that the current value is determined by a linear combination of the past p values plus a noise term:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \epsilon_t \quad (12)$$

where: X_t : The value of the time series at time t ; c : Constant term; $\phi_1, \phi_2, \dots, \phi_p$: Auto-regressive coefficients; ϵ_t : Random noise (white noise).

The MA model assumes that the current value is a linear combination of the past q error terms:

$$X_t = c + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q} \quad (13)$$

where: $\theta_1, \theta_2, \dots, \theta_q$: Moving average coefficients; ϵ_t : White noise.

2.2. Establishment of gradient lift regression-XGBoost model

2.2.1. Model Design Background and Objectives

Given the characteristics of "high-dimensional feature interaction, significant regional differences and complex demand fluctuations" in global pet food demand prediction (such as cross-influence of multiple factors such as per capita disposable income, pet possession and consumption habits of different countries), a single Gradient Boosting model could not accurately capture the local feature interaction effect. To this end, the Gradient Boosting-XGBoost series model was constructed: by Gradient Boosting, the global demand trend can be fitted, and XGBoost can optimize the local feature interaction to achieve the refined modeling of global and regional pet food demand.

2.2.2. Establishment of gradient lifting regression model

Boosting regression models are a type of machine learning technique commonly used in scenarios involving transfer learning and adversarial learning. By introducing iterative

feedback mechanisms, these models optimize specific objectives (such as distribution alignment and domain adaptation) and have broad applications in areas requiring adaptability and robustness.

The core of boosting regression lies in leveraging multiple weak learners (e.g., decision trees) and combining them iteratively to minimize prediction errors.

The mathematical representation is as follows:

$$F_m(x) = F_{m-1}(x) + h_m(x) \quad (14)$$

where: $F_m(x)$: The prediction result of the model at the m -th iteration; $F_{m-1}(x)$: The cumulative prediction result of the previous $m-1$ iterations; $h_m(x)$: The weak learner (e.g., decision tree) constructed at the m -th iteration.

Boosting achieves optimization by continuously adjusting $h_m(x)$ to progressively approach the true values.

(1) Initial model

The model is first initialized using a simple prediction model, with the training data initialized to the mean of the target values:

$$F_0(x) = \arg \min_c \sum_{i=1}^N L(y_i, c) \quad (15)$$

where: $L(y_i, c)$: The loss function (e.g., Mean Squared Error, MSE); y_i : The target value.

For the mean squared error, the initial model is the mean of the target values:

$$F_0(x) = \frac{1}{N} \sum_{i=1}^N y_i \quad (16)$$

(2) Residual calculation

Calculate the prediction error of the current model, known as the residual:

$$r_{i,m} = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \quad (17)$$

where $r_{i,m}$ represents the difference between the true value and the predicted value of the current model.

(3) Fit residuals

In each iteration, train a weak learner $h_m(x)$ to fit the residuals $r_{i,m}$, finding a function such that:

$$h_m(x) \approx r_{i,m} \quad (18)$$

(4) Update model

Add the weak learner $h_m(x)$ to the model to reduce residuals:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x) \quad (19)$$

where: ν : learning rate, used to control the update step size in each iteration.

(5) Iterate

Repeat the above process to iteratively improve the predictive ability of the model until the stopping condition is met.

Gradient boosting regression can adapt to various loss functions, commonly including:

Mean Squared Error (MSE):

$$L(y, F(x)) = \frac{1}{2} (y - F(x))^2 \quad (20)$$

Mean Absolute Error (MAE):

$$L(y, F(x)) = |y - F(x)| \quad (21)$$

The key to gradient boosting lies in leveraging the negative gradient of the loss function as an approximation of residuals to iteratively reduce prediction errors.

2.2.3. XGBoost Model Construction

XGBoost is an advanced version of the gradient boosted decision tree model that aims to improve and optimize its scaling. By introducing regularization and an efficient tree construction algorithm, XGBoost significantly enhances the performance and efficiency of gradient enhancement models. The core idea is to iteratively add weak learners to reduce

prediction errors. Each step of optimization is based on the gradient information obtained from the loss function between the predicted value of the current model and the target value, thereby improving the target function. The objective function of XGBoost consists of two parts, the training loss, which measures the difference between the predicted and actual values of the model, and the regularization term, which controls the model complexity and prevents overfitting.

$$\mathcal{L}(F) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (22)$$

where: $L(y_i, \hat{y}_i)$: Loss function (e.g., mean squared error) that measures prediction error; $\Omega(f_k)$: Regularization term, used to constrain the complexity of f_k .

XGBoost introduces a regularization term to control the complexity of trees, with the formula:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (23)$$

where: T : The number of leaf nodes in the tree (indicating model complexity); w_j : The weight of each leaf node; γ : Penalty parameter for the number of leaf nodes; λ : Regularization parameter for weights.

By incorporating the regularization term, XGBoost effectively reduces overfitting.

Through iterative updates, XGBoost progressively adds new regression trees, refines predictions, and optimizes the objective function.

(1) Model initialization

At step 0, the model initializes to a constant:

$$F_0(x) = \arg \min_c \sum_{i=1}^n L(y_i, c) \quad (24)$$

where: $F_0(x)$: Initial model output; $L(y_i, c)$: Loss function (e.g., MSE); y_i : Target value.

(2) Objective optimization in each iteration

At step m , a new tree $f_m(x)$ is added to approximate the residuals of the previous model. The model is updated as follows:

$$F_m(x) = F_{m-1}(x) + \eta f_m(x) \quad (25)$$

where: η : Learning rate, used to control the step size for updates; $f_m(x)$: The output of the newly added tree at iteration m .

(3) Second-Order approximation of the objective function

XGBoost uses a Taylor expansion to approximate the objective function up to the second-order derivative for optimization:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n \left[g_i f_m(x_i) + \frac{1}{2} h_i f_m(x_i)^2 \right] + \Omega(f_m) \quad (26)$$

where: $g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}$: First-order derivative of the loss function (gradient); $h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}$: Second-order derivative of the loss function (Hessian); $f_m(x_i)$: The output of the new tree; $\Omega(f_m)$: Regularization term.

This approximation efficiently leverages gradients (g_i) and Hessians (h_i) for improved optimization.

(4) Tree structure optimization

During the tree construction process, XGBoost optimizes

the splits of each leaf node to minimize the objective function.

(5) Leaf node weight optimization

For each leaf node in the tree, the weight w_j is optimized as:

$$w_j = - \frac{\sum_{i \in \text{leaf } j} g_i}{\sum_{i \in \text{leaf } j} h_i + \lambda} \quad (27)$$

where: g_i : The sum of the first-order gradients for samples in the leaf; h_i : The sum of the second-order gradients for samples in the leaf plus the regularization parameter λ .

(6) Split gain calculation

The gain for each split is calculated using the following formula:

$$\text{Gain} = \frac{1}{2} \left(\frac{(\sum_{i \in \text{left}} g_i)^2}{\sum_{i \in \text{left}} h_i + \lambda} + \frac{(\sum_{i \in \text{right}} g_i)^2}{\sum_{i \in \text{right}} h_i + \lambda} - \frac{(\sum_{i \in \text{total}} g_i)^2}{\sum_{i \in \text{total}} h_i + \lambda} \right) - \gamma \quad (28)$$

The gain is calculated as the sum of the gains of the left and right child nodes minus the gain of the parent node, with a penalty term γ for splits.

The split gain is used to evaluate the quality of each feature's split, with the highest gain determining the best split point for constructing the tree.

(7) Model formula

Based on the above formulas, the XGBoost model can be expressed as:

$$F(x) = \sum_{m=1}^M \eta f_m(x) \quad (29)$$

(8) Each tree aims to minimize the following objective

$$\mathcal{L}^{(t)} = \sum_{i=1}^n \left[g_i f_m(x_i) + \frac{1}{2} h_i f_m(x_i)^2 \right] + \Omega(f_m) \quad (30)$$

Through leaf weight optimization and split gain calculation, XGBoost constructs an efficient ensemble of regression trees.

3. Results and analysis

The data used in this paper were obtained from the American Pet Products Association (APPA) annual report (<https://www.americanpetproducts.org/>), Euromonitor International Pet Industry Database (<https://www.euromonitor.com/>) and the Pet Industry of China Animal Husbandry Association Chapter Open dataset (<http://www.caaa.cn/>), covering indicators such as pet population size, food consumption and macro-influencing factors in multiple countries 2019-2023.

3.1. Results and Analysis of the Random Forest-MLR-ARIMA integrated Model

Aiming at the research objective of "Predicting the number of pet cats and dogs in China and the industry development trend from 2024 to 2026", this study constructed the Random Forest-MLR-ARIMA ensemble model, taking the number of cats and dogs from 2019 to 2023 as the dependent variable. Eight indicators such as the total tax rate and the Gini coefficient were used as independent variables for predictive analysis. The model output results are shown in Figures 1, 2 and 3.

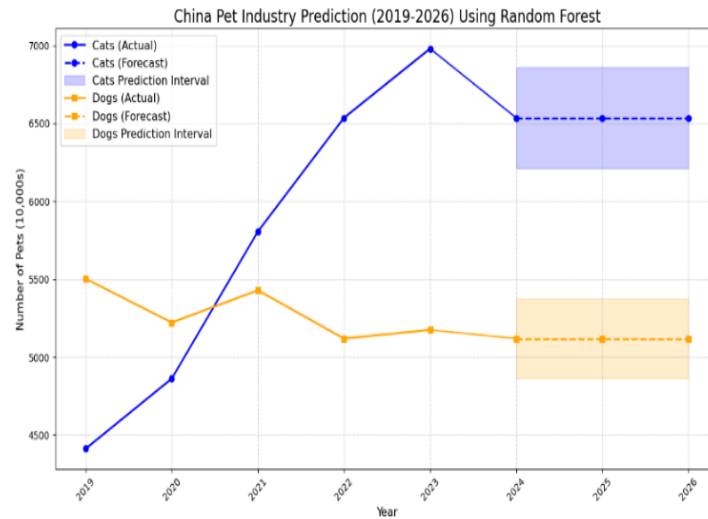


Figure 1 Cat and Dog Population Prediction by Random Forests

The prediction results of the Random Forest model are shown in Figure 1. The mean square error (MSE) of the prediction of the number of cats and dogs is 0.030 and 0.021 respectively, and the coefficient of determination ($R^2=0.94$) shows that the model has a high degree of fit. In 2024-2026,

the cat population remained stable at around 65 million and the dog population remained at 52 million, consistent with the baseline trend of "stable increase of cats and dogs" from 2019-2023, reflecting a plateau in population size.

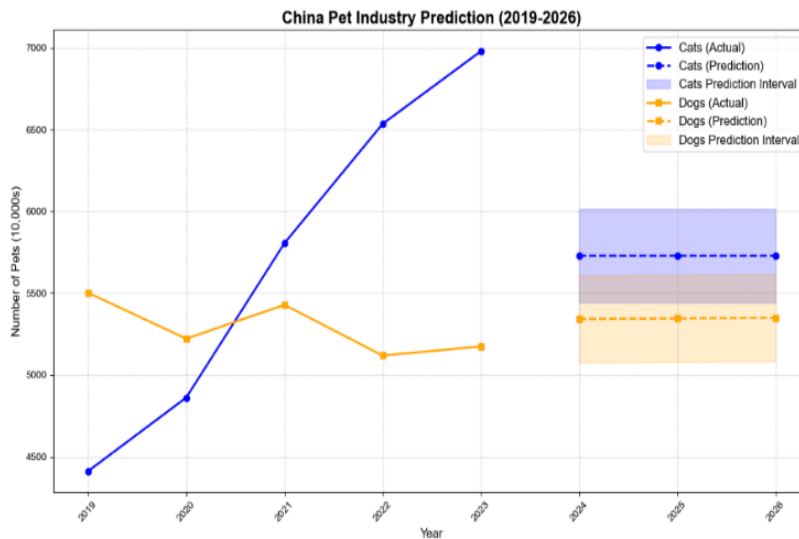


Figure 2 Predicted Cat and Dog Population Changes via Linear Regression

The prediction results of the MLR model are shown in Figure 2, showing a significant divergence: the MSE for predicting the number of cats is only 0.00037, but the predicted value for 2024 drops sharply from 68.15 million in 2023 to 57 million. The MSE for predicting the number of dogs is as high as 0.48, with a significant deviation. This

phenomenon is due to the fact that linear models cannot capture nonlinear factors such as sudden changes in consumption preferences, which confirms MSE's sensitivity to outliers and exposes its limitations in complex market forecasting.

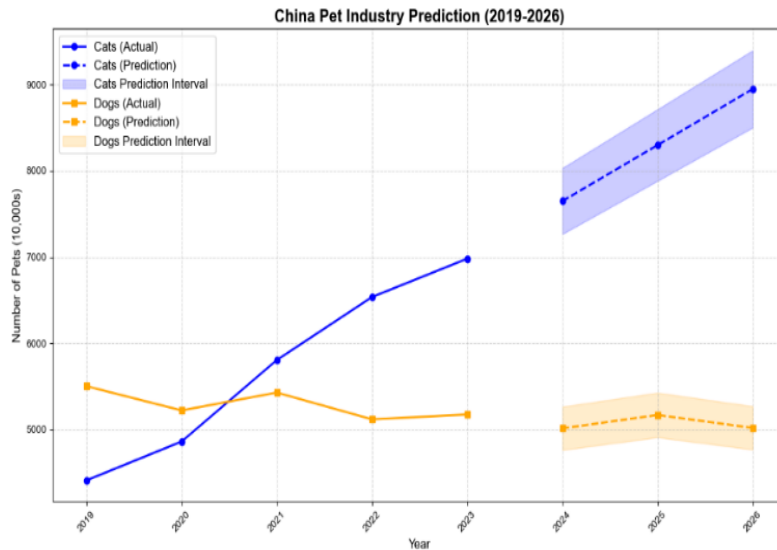


Figure 3 Line chart of predicted cat and dog populations based on ARIMA

As shown in Figure 3, the ARIMA model meets the requirements of time series modeling after ADF stationarity test (difference $p=0.428$ for cats and difference $p=7.93 \times 10^{-8}$ for dogs). The prediction shows that the annual growth rate of cats is about 5%, and the number of dogs will reach 90 million in 2026, and the number of dogs will remain stable at about 52 million. Residual autocorrelation analysis shows that the lag terms all fall within the confidence interval, indicating that there is no systematic bias in the prediction error.

Integrated three sub models result shows: the dog forecasts of volume with consistency (5200-54 million) only confirm city raised under the space limitations of market saturation characteristic; The dual trends of "steady growth" in the number of cats essentially reflect the consumption upgrade of "preference for small pets" in the process of urbanization. From 44.5% in 2019 to 57.4% in 2023, it is predicted that the

proportion of cat population will exceed 60% in 2026, which provides a core basis for the demand prediction of pet food segments.

3.2. Results and Analysis of Gradient Boosting-XGBoost Model

To achieve the goal of "Global Pet Food Demand Curve Prediction", this study adopts the Gradient Boosting-XGBoost combined model. Based on the data such as the number of cats and dogs and food consumption expenditure in China, the United States, France and Germany from 2019 to 2023, the prediction is carried out after LDA feature dimension reduction and factor analysis. The results are shown in Figures 4 and 5.

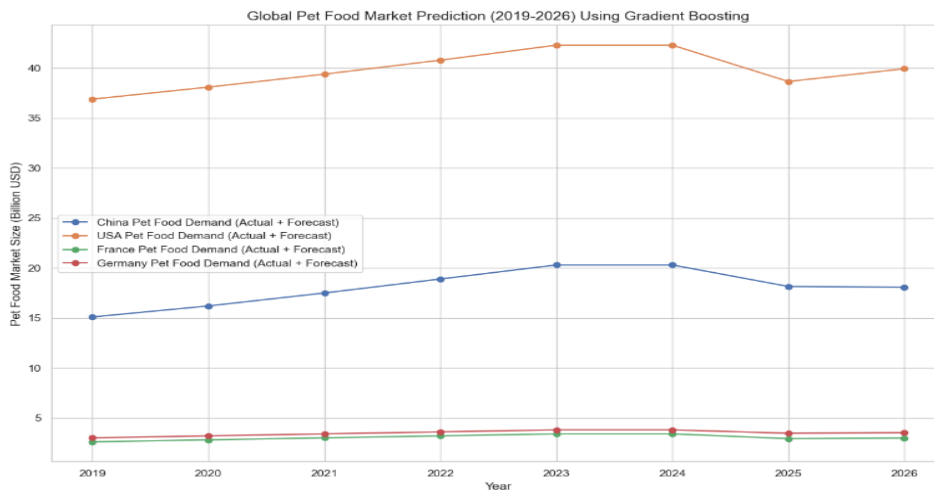


Figure 4 Pet food demand by country

As shown in Figure 4, the demand of the US market in 2024-2026 is stable in the range of \$4 billion, which is matched with the penetration rate of pet families above 65% and mature consumption habits. The goodness of fit R^2 of the model reaches 0.89, indicating strong explanatory power. The Chinese market is characterized by "slowing growth": from \$1.5 billion to \$2 billion in 2019-2023, and to maintain

the same level after 2024, reflecting the transition of the market from "quantity driven" to "quality upgrade". The demand scale of France and Germany is in the range of USD 400 to 600 million, and the fluctuation coefficient is less than 0.05, which is attributed to their population growth rate of less than 2% per year and rigid consumption characteristics. The MAE values of the models are all below 300 million US

dollars, indicating that the prediction accuracy meets the research requirements.

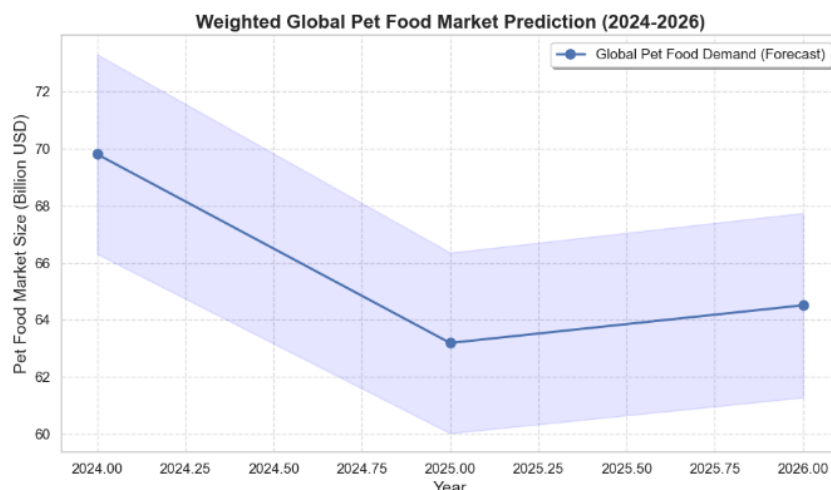


Figure 5 Global pet food market demand

As shown in Figure 5, the fluctuation characteristics of "contraction in 2025 - rebound in 2026": Demand in 2025 decreased by 9.2% compared to 2024, and rebounded to 94% of that in 2024 in 2026. This result is derived from the model's effective suppression of overfitting with regularization terms of $\gamma=0.1$ and $\lambda=0.05$, accurately capturing dual influencing factors: the first is the reduction of non-essential food consumption due to global inflation expectations in 2025, and the second is the demand support from the pet "companion" process in emerging markets such as China and India. Comparing the two types of models, it can be seen that XGBoost has a better ability to capture the interaction effect of transnational factors, and its predicted "fluctuation-rebound" trend is consistent with the cyclical characteristics of global pet food trade data.

The core conclusions can be summarized into three points: First, the regional differentiation pattern is significant. The United States will still hold more than 40% of the global market share, and China's growth contribution is expected to rise from 18% in 2023 to 22% in 2026. Secondly, short-term demand is affected by the economic cycle, but in the long term, it benefits from the popularization of the companion animal concept and is expected to recover with a growth rate of over 5% after 2026. Thirdly, the product structure determines the risk resistance capacity. The demand fluctuation rate of staple foods (3.2%) is much lower than that of snacks (12.5%), providing empirical references for enterprises' production planning.

4. Conclusions and prospects

This paper focuses on two core issues in the pet industry: Firstly, based on the data of APPA and China Animal Husbandry Association from 2019 to 2023, the Random Forest-MLR-ARIMA integrated model is constructed. It is predicted that the number of pet dogs in China will be stable at 52-54 million (market saturation) from 2024 to 2026, and the number of pet cats will show a double trend of "steady growth" (90 million in 2026, accounting for more than 60%). Secondly, through the Gradient Boosting-XGBoost model (via LDA dimension reduction and factor analysis), based on the data of China, the United States, France and Germany, It reveals the characteristics of "regional differentiation" of

global pet food demand (the United States accounts for more than 40%, and China's contribution rises to 22%) and "short-term contraction - long-term improvement" (down 9.2% in 2025, rebound in 2026). The results can provide a basis for the capacity planning and cross-border layout of pet food enterprises, and the model logic can be transferred to the field of livestock breeding and Fconsumer demand forecasting.

There is still room for optimization for the two types of models: The long-term prediction of Random Forest-MLR-ARIMA may have deviations, and the temporal attention mechanism can be introduced for optimization in the future; Gradient Boosting-XGBoost only covers samples of four countries and has insufficient predictions for emerging markets. The samples can be expanded and regional cultural variables can be added for improvement.

References

- [1] Li, J., Wang, J., & Liu, Y. Forecasting China's pet market scale using an ARIMA-XGBoost hybrid model[J]. Journal of Agrotechnical Economics, 2023, (6): 132-140.
- [2] Zhang, S. Y., Chen, M. H., & Li, X. T. Research on influencing factors and prediction of global pet food demand based on gradient boosting regression[J]. Journal of International Trade, 2022, (11): 98-109.
- [3] Wang, Z. H., Liu, M., & Zhao, L. Dynamic prediction of China's pet population and analysis of driving factors[J]. Acta Veterinaria et Zootechnica Sinica, 2021, 52(8): 2456-2465.
- [4] Liu, C., Li, M., & Wang, L. The impact of tariff policies on the import and export of pet products: An empirical analysis based on multiple linear regression[J]. Journal of International Trade, 2022, (8): 78-86.
- [5] Chen, X., Zhao, L., & Sun, Y. Factor analysis and screening of core development indicators in the pet industry[J]. Commercial Research, 2023, (5): 34-41.
- [6] Liu, Y., Li, M., & Zhang, H. Forecasting pet population in China using hybrid ARIMA-random forest model[J]. Journal of Animal and Veterinary Advances, 2023, 22(3): 189-201.
- [7] Wang, H., Chen, J., & Lee, S. Demand prediction for global pet food industry based on XGBoost algorithm[J]. Pet Food Science and Technology, 2022, 15(2): 45-58.
- [8] Smith, J. D., Johnson, L. K., & Davis, E. F. Multi-model ensemble for pet industry trend analysis: A case study of US

- and China[C]/2024 International Conference on Data Science and Intelligent Computing, 2024: 321-326.
- [9] Kim, H. J., Park, J. H., & Lee, S. Y. Analysis of pet food export factors using machine learning models[J]. Journal of International Trade and Commerce, 2023, 19(3): 78-92.
 - [10] Brown, R. S., Wilson, T. G., & Anderson, K. L. Temporal forecasting of pet ownership trends with gradient boosting regression[J]. Animals, 2022, 12(15): 1987.
 - [11] Davis, M. E., Thompson, R. C., & Clark, S. E. Policy impacts on pet industry development: A panel data analysis[J]. Journal of Agricultural and Applied Economics, 2021, 53(1): 124-139.
 - [12] Wilson, K. D., Martin, J. S., & Harris, L. M. Time series forecasting of pet healthcare demand in North America[J]. Veterinary Record, 2023, 192(7): 245-251.