IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 9, 2022, accepted April 6, 2022, date of publication April 11, 2022, date of current version April 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3166157

Predicting Airline Additional Services Consumption Willingness Based on High-Dimensional Incomplete Data

JIANING CHEN, MINGGUANG DIAO, AND CHUYAN ZHANG
School of Information Engineering, China University of Geosciences (Beijing), Beijing 100083, China

Corresponding authors: Mingguang Diao (dmg@cugb.edu.cn) and Chuyan Zhang (zcy@cugb.edu.cn)

This work was supported in part by the National College Students Innovation and Entrepreneurship Training Program of China under Grant 2022A231, and in part by the Fundamental Research Funds for the Central Universities of China under Grant 2652019026.

ABSTRACT Prediction of the purchase willingness of passengers has great benefits for airlines to promote auxiliary services, however, the datasets stored in passenger travel information systems are often high-dimensional and incomplete. This study develops a prediction method of airline additional service consumption willingness based on high-dimensional and incomplete datasets with a triple-layer hybrid PSO-XGBoost model, which consists of an incomplete data processing layer, a high-dimensional data processing layer, and a predicting layer. The raw dataset is converted into a complete and low-dimensional dataset through the first two layers and inputted into the predicting layer to train and optimize the XGBoost model together with the PSO algorithm and 10-fold cross-validation. The experimental results show that the proposed method outperforms other traditional machine learning models, presenting the highest prediction score with 0.9879 in terms of AUC. The findings help predict airline additional services consumption intentions of passengers and are beneficial to efficient and low-cost precise marketing for airlines.

INDEX TERMS Airline addition services, consumption willingness, XGBoost, machine learning.

I. INTRODUCTION

Conducting airline additional services brings new profit growth for airline companies. However, the attention and related studies are very limited. The existing literature about the consumption willingness of airline additional services mostly uses the questionnaire method [1], [2] or scenario hypothesis method [3], [4] to collect data and adopt the theory of consumer psychology to build traditional statistical models. The literature [5], [6] mines customer opinion survey data to analyze the factors affecting customer satisfaction by methods such as exploratory factor analysis and classifiers such as Parsimonious Bayes. The real passenger travel data which is subjective and limited, however, has not been used in any form of study. The literature [7] uses the Potluck Problem method to predict cargo demand for a given airline on a given route, but the method is not applied to the field of exploring customer consumption intentions. With the development of modern information technology such as the Internet plus and big data, relying on the real travel records of a large number of passengers provides objective conditions for accurately mining the consumption willingness of additional services of passengers.

The problems of different passenger travel data storage structures and access methods among passenger service information subsystems have caused a large amount of missing and redundant data [8]. In addition, due to the small consumer group of additional services, the insufficient data collection leads to an incomplete and high-dimensional air passenger travel dataset, which further affects the accuracy of prediction. However, there are insufficient attention and sparse solutions to the above-mentioned problems.

Given the above situation, seat selection, being regarded as the most valuable additional service for long-distance passengers, is selected as the research object, and the specific dataset related to seat selection is used. The eXtreme Gradient Boosting (XGBoost) model is introduced as a machine learning model into the field of consumer behavior prediction of airline additional service, and a triple-layer Particle Swarm Optimization (PSO) algorithm optimized XGBoost model is proposed for prediction. An incomplete data processing layer and a high-dimensional data processing layer are respectively set up to address the defects of severe missing, unbalancedness, and high-dimensionality of datasets. Then the processed datasets are inputted into the predicting layer.

The associate editor coordinating the review of this manuscript and approving it for publication was Tallha Akram.

The three-layer PSO-optimized XGBoost prediction model proposed in this paper achieves accurate prediction of passengers' purchase of paid seat selection ancillary services on incomplete and high-dimensional airline passenger travel dataset, which has the advantages of fast speed and high accuracy. This study can provide effective decisions for airlines' accurate marketing of ancillary services. Additionally, this paper mines real airline passenger travel and paid seat selection service consumption data to build a paid seat selection passenger characteristics portrait, which provides a scientific basis for airlines' auxiliary service marketing campaign settings.

The main content of this paper is organized as follows. Section 2 describes in detail the incompleteness and high dimensionality problems of airline passenger travel datasets. Section 3 specifies the design details of the proposed three-layer PSO-optimized XGBoost prediction model, including the theoretical basis of the model, the data pre-processing process, and the working process of the proposed three-layer prediction model. Section 4 evaluates the effectiveness of the proposed three-layer prediction model with data incomplete processing layer, data high-dimensional processing layer, and PSO optimization settings. The superiority of the proposed method by comparing it with the single-layer traditional machine learning model is demonstrated. Conclusions are presented in Section 5.

II. AIR PASSENGER TRAVEL DATA

The study focuses on exploring the classification of passengers by their consumption willingness for seat selection with data obtained from 23,432 passenger travel samples collected by an airline from 2016 to 2020.

The dataset contains 652 features as inputs, such as passenger gender, flight number, count of seat selection, preferred flights, accumulated mileage, etc., and a binary output for whether a passenger has purchased seat selection services. The dataset includes five types of features, as illustrated in Table 1. Table 1 shows that the input space is high-dimensional, containing more redundant features as well as various aspects of passengers' flight-related information. It not only brings difficulties for experts to find the exact critical effectors of consumption intentions for additional services but also reduces the prediction accuracy and efficiency.

The missing data rate and variance distribution of the collected dataset are shown in Figure 1. It needs to be noted that the missing values in the raw dataset have already been replaced with 0 by the data storage system such that the 0 value in numerical features can not indicate whether it is missing or real. Therefore, the information loss in the dataset is reflected by both the missing rate of categorical features and the variance distribution of normalized numerical features. Figure 1(a) shows that only about 5% of the categorical features are relatively complete, and the rest 95% have over 60% missing data. Figure 1(b) illustrates that about 25% of numerical features have 0 variances (less than 10^{-10}), and most of them have small variance. Owing to the traits mentioned above, the dataset has serious information loss.

TABLE 1. Feature types of the dataset.

No.	Feature type name	Explanations	Count
1	Demographics of passengers	Basic demographic information of passengers.	11
2	Current flight information	Flight information of this sample.	7
3	Historical flight preferences of passengers	Statistical preference information of passengers about flight.	250
4	Historical travel information of passengers	Aviation-related historical statistical travel information of passengers.	119
5	Historical consumption information of passengers	Aviation-related historical statistical consumption information of passengers.	265

There are 1,475 positive samples of purchasing seat selection additional services in the dataset, accounting for only 6.29%. This data amount is far less than the negative sample amount, presenting an imbalanced distribution among classes. The model's judgment of the target will be affected by the imbalance since the model is biased to classify the test sample into the category with high-cardinality to guarantee accuracy. However, this is not what airlines expect.

In summary, there are several prominent problems of high dimensionality and incompleteness including missingness and unbalance in the dataset. The accuracy and credibility of the prediction results will be greatly reduced if the above problems are not addressed.

III. METHODS

Owing to the incomplete and high dimensional dataset, a method based on a triple-layer Particle Swarm Optimization modifying the eXtreme Gradient Boosting (PSO-XGBoost) model is proposed to predict passengers' consumption willingness about specific seat selection additional services. The architecture is shown in Figure 2.

First, the raw dataset is roughly cleaned and encoded to make the model focus more on prominent problems and more universal. After that, the cleaned dataset is inputted into the triple-layer model where PSO-XGBoost is used as the base model. Second, in the triple-layer model, the incomplete data processing layer (IDP-layer) is set to handle the problems of data missing by imputation and imbalance by resampling. Then the dimension of the complete dataset is reduced in the high-dimensional data processing layer (HDP-layer) according to the feature importance obtained by XGBoost. Finally, in the predicting layer (P-layer), the processed dataset is used to train the XGBoost model. A state-of-the-art heuristic, the Particle Swarm Optimization algorithm, is applied to tune the hyperparameters of XGBoost to enhance the accuracy obtained by 10-fold cross-validation. The optimal XGBoost classifier attained is adopted to predict the willingness of new passengers to purchase seat selection.

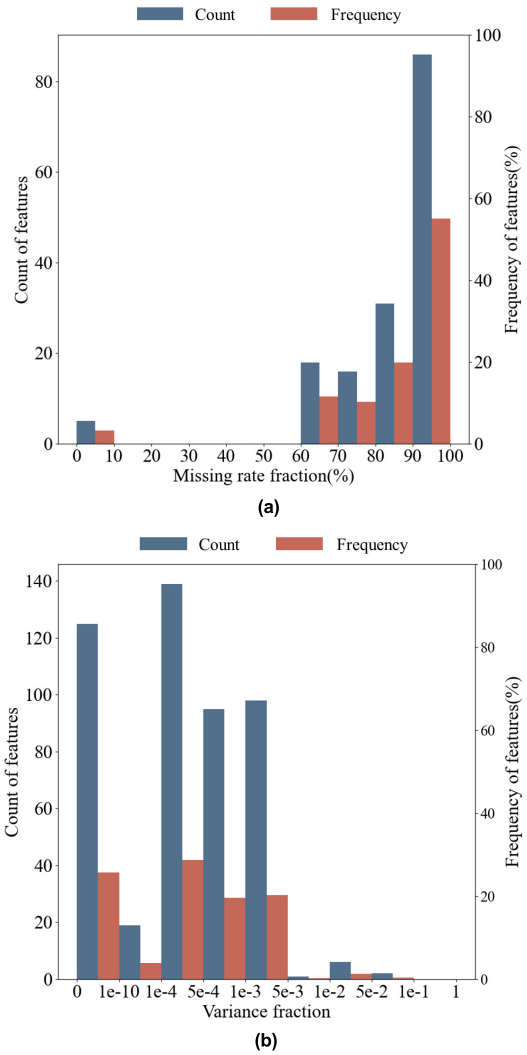

FIGURE 1. Data missing rate and variance distribution of the dataset: (a) Missing rate of classification features. (b) Variance distribution of normalized numerical features.

A. DESIGN OF THE BASIC MODEL

The base model PSO-XGBoost of airline additional service consumption willingness prediction method proposed is a hybrid model [9]–[11], combining Particle Swarm Optimization algorithm (PSO) and eXtreme Gradient Boosting model (XGBoost).

XGBoost [12] modifies the traditional gradient boosting tree, which is widely used in the field of consumer behavior prediction with its superior generalization performance, prediction accuracy, and outstanding parallel computing rate [13]. The XGBoost is an additive algorithm consisting of t CART model, the objective function of which comprises a loss function and a regularization term, defined as

$$O_t = \sum_{i=1}^{n} L(y_i, f_{t-1}(x_i) + h_t(\boldsymbol{x_i})) + \Omega\left(h_t\left(\boldsymbol{x}\right)\right) \quad (1)$$

$$\Omega\left(h_t\left(\boldsymbol{x}\right)\right) = \gamma J_t + \frac{\lambda}{2}\sum_{j=1}^{J_t}\omega_{tj}^2 \quad (2)$$

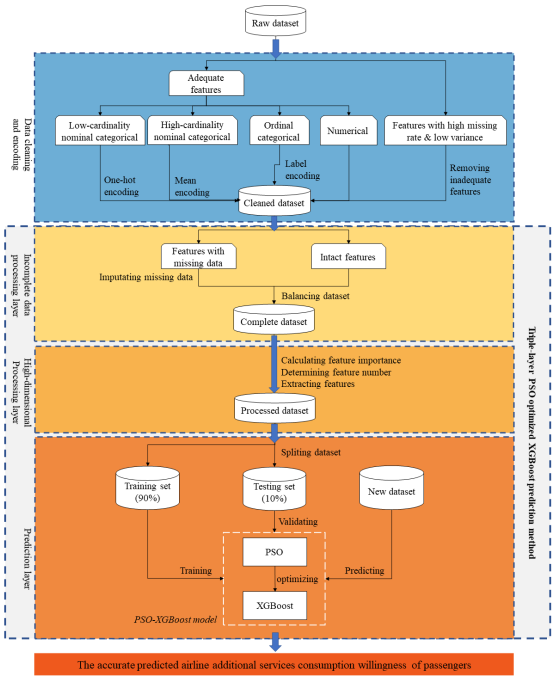

FIGURE 2. The architecture of prediction method of airline additional service purchase willingness based on triple-layer PSO-XGBoost model.

where, $L(\cdot)$ represents the loss function, $\Omega(\cdot)$ is a regularization term based on tree complexity, which is beneficial to reducing the risk of over-fitting.

Based on the loss function, the second-order Taylor expansion is introduced. Then, the optimal weight of each leaf node is solved iteratively to obtain the final objective function, as follows,

$$O_t = -\frac{1}{2}\sum_{j=1}^{J}\frac{\left(\sum_{x_i \epsilon R_{tj}} g_i\right)^2}{\sum_{x_i \epsilon R_{tj}} h_i + \lambda} + \gamma J \quad (3)$$

The PSO algorithm [14] is a modern heuristic algorithm adopted to solve combinatorial optimization problems. The basic idea is to use the position $\boldsymbol{x}_i$ of i-th particle to represent a candidate solution to the problem, and use a fitness function value to evaluate the superiority of positions. All particles update their velocity through (4) and position through (5) by sharing the individual and global position with others, and will finally gather in the extremum area after multiple iterations.

$$\boldsymbol{v}_{i+1} = v_i + c_1 \times r \times \left(pbest_i - \boldsymbol{x}_i\right) + c_2 \times r \times \left(gbest - \boldsymbol{x}_i\right)\boldsymbol{v}_{i+1} \quad (4)$$

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \boldsymbol{v}_{i+1} \quad (5)$$

Machine learning models contain many hyperparameters to be set externally due to their complex construction. The appropriate hyperparameters are crucial to the results since they can improve performance while reducing the risk of the overfitting of models. Manual or traversal search is used as a conventional parameter tuning method,

TABLE 2. The tuned parameters and tuning ranges.

No.	Parameter	Description	Tuning range
1	N	The number of iterations.	[10,500]
2	η	Learning rate.	[0,1]
3	max_$depth$	The maximum depth of a tree.	$[0,+\infty]$
4	γ	The minimum loss reduction required to make a split.	$[0,+\infty]$
5	min_child_weight	The minimum sum of instance weight.	$[0,+\infty]$

which requires, however, a time-consuming and arduous process. Many studies have verified that adopting the PSO algorithm to search near-optimal hyperparameters makes models more automated and robust [13]. Therefore, in this study, XGBoost optimized by PSO is used as the base model. Five hyperparameters of XGBoost are selected for optimization, as shown in TABLE 2.

B. DATA CLEANING AND ENCODING

Before being inputted into the triple-layer model, the raw dataset is firstly cleaned and encoded. The following steps were carried out:

(1) Rough cleaning. Filling the features with an extremely high missing rate will introduce more noise, and retaining the features with low variance will get sparse information but lose more efficiency. Therefore, features with the above defects are directly deleted to preliminary reduce the dimension of the dataset.

(2) Date features handling. The dataset includes more date-related features, such as travel time and preferred travel month, which contain rich information but do not have a direct relationship with consumption willingness. Considering that passenger flow and physical fatigue level will largely affect the passenger's requirement for seat comfort, departure year and month variables are mapped to the traveler flow with the samples share representing. Additionally, the departure day variable is divided into a fatigue-prone period (from 20:00 to 8:00 the following day) and a non-fatigued period (from 9:00 to 19:00) for binarization. Other year and month variables are replaced by the average of departure year or month values.

(3) Categorical feature encoding. The dataset after step (1) contains 34 categorical features, such as passenger gender, flight cabin, etc. Since the machine learning model has a better performance on numerical features, the categorical feature should be transformed into their numerical counterparts by encoding [15], [16]. Among them, the flight cabin is subjected to label encoding [17] due to the ordinal values. Passengers' gender and flight destination are subjected to one-hot encoding [17] which maps the 1-dimensional N-value feature to the N-dimensional binary feature because of the low-cardinality and nominal values. The rest features are processed by mean encoding [18] due to their high cardinality and disorder.

After the above processing, the number of features in the dataset is reduced from 652 to 137, with no extremely serious missingness, low variance, or categorical feature values. The cleaned dataset then will be inputted into the triple-layer model.

C. THE INCOMPLETE DATA PROCESSING LAYER

The IDP-layer solves the problems of data missing and imbalance of the air passenger travel dataset.

In the first stage, the missing values are imputed. Uniform constant values are often used to impute missing data in past studies [13], which does not perform well in severely missing datasets. Machine learning models are widely used in imputation recently, considering other complete features of samples to make the imputed values closer to real values [19]. Therefore, PSO-XGBoost is adopted to impute missing data in this stage. The following five steps are carried out:

(1) Sort missing features in ascending order according to their missing rate.

(2) Take the least missing feature X as the prediction target Y, use the sample set with observations on Y as the training set to train the model, and then use the sample set with missing values on Y as the test set to predict the corresponding Y values by the model.

(3) During training and prediction, missing values of the remaining missing features are temporarily filled by 0.

(4) Update the dataset by imputing feature X with the predicted values.

(5) Repeat steps (2) to (4) until the dataset contains no missing values.

The optimal coefficient of determination R^2 and the optimal mean-square error percentage for filling each missing feature is shown in Figure 3. The 30 features containing missing values from the preprocessed dataset are filled with missing values using the PSO-XGBoost model. As can be seen from the figure, the R^2 of most of the features filled are close to 1, and the mean square error percentages are less than 1%. With that indicated the method can fill the missing values of the features better and can improve the prediction accuracy of the model without affecting the overall performance of the missing data.

In the second stage, the dataset is resampled to balance. In the related literature, random undersampling, random oversampling, or synthetic minority oversampling techniques (SMOTE) [15] are often used for resampling. Considering that the former two approaches may lead to the risk of data loss or model overfitting, SMOTE [20] based on the K-nearest neighbor idea is applied to balance the dataset. Figure 4 shows the sample distribution of the dataset before and after SMOTE processing on the two feature dimensions with the lowest missing rate, which shows that the number of positive samples for purchasing seat selection services increases, and the sample distribution transforms into a balance after resampling by SMOTE.

As a result, the cleaned dataset is changed from missing and unbalanced to complete and balanced.

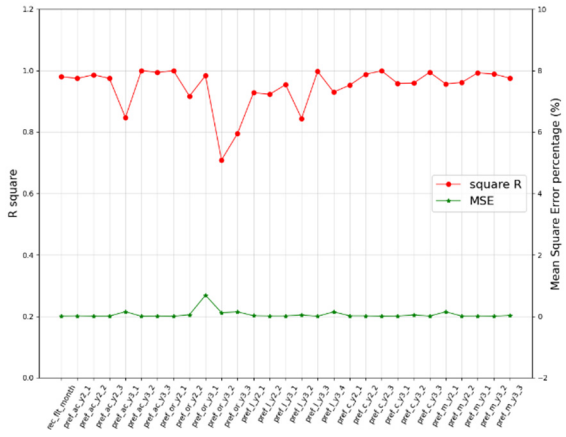

FIGURE 3. Effect of missing data imputation with PSO-XGBoost.

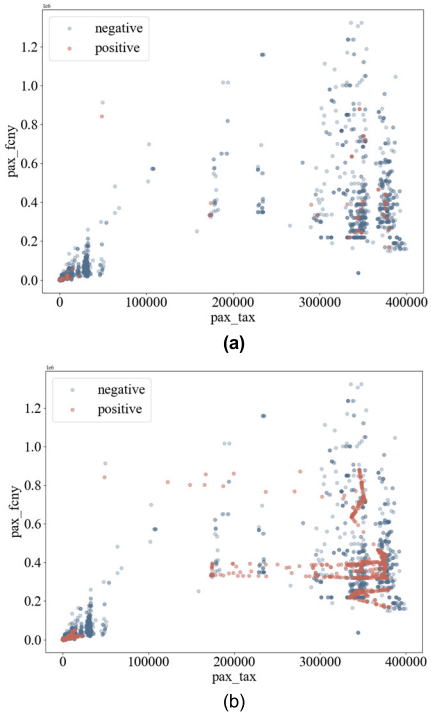

FIGURE 4. Distribution of samples before and after SMOTE resampling: (a) Before SMOTE resampling. (b) After SMOTE resampling.

D. THE HIGH-DIMENSIONAL DATA PROCESSING LAYER

HDP-layer solves the problem of the high dimension of air passenger travel dataset and can be approached by the following three steps.

In the first step, the feature importance is calculated using XGBoost and ranked. Feature importance is a product in the XGBoost training process, which is calculated from the node splitting gain every time. The gain is given by

$$gain = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma. \quad (6)$$

In the second step, the loss of XGBoost on the dataset containing the number of features of all combinations from 1 to 137 is calculated from the ordered features. Then

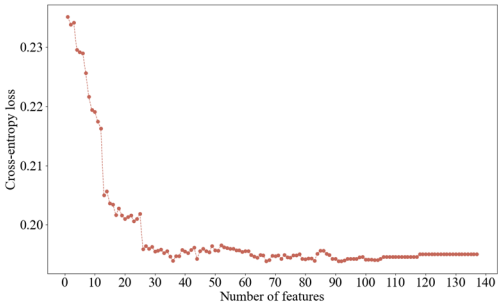

FIGURE 5. Relationship between feature number and loss of XGBoost model.

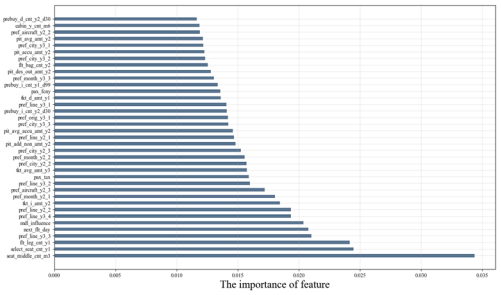

FIGURE 6. Top 37 most important features ranking and their importance.

the number n of features with the lowest error is selected as the optimal solution.

In the third step, the top n important features to form the processed dataset are selected as the final input to the P-layer.

Figure 5 illustrates the relationship between the number of features and the cross-entropy loss of XGBoost, and it can be seen that the loss of the classifier is the lowest when the number of features is 37 or 90. Finally, the top 37 important features are extracted considering the operation efficiency. The top 37 most important features ranking and their importance are shown in Figure 6.

As can be seen from Figure 6, the top 37 important features consist of 19 passenger history airline preference features, 8 passenger history consumption information features, 7 passenger history travel information features, 2 current flight information, and 1 basic passenger information, of which 51.13% are passenger history airline preference features, indicating that passengers' travel habits and preferences largely determine their willingness to purchase paid seat selection ancillary services. The top two features are the times to sit in the middle seat in the past three months (seat_middle_cnt_m3) and times to select a seat in the past year (select_seat_cnt_y1), indicating that the passenger's seat preference has the most significant impact on the willingness to pay for seat selection. Therefore, airlines should focus on tapping passengers' airline preferences, in other words, to focus on passengers' seat preferences, and put targeted marketing advertisements for passengers with different flight preferences as well as seat preferences.

Meanwhile, statistics show that, in terms of quantity, the characteristics related to preferred travel route (pref_line) and preferred arrival city (pref_city) account for the largest

TABLE 3. Optimized parameter values by PSO algorithm.

hyperparameter	T	η	mtd	γ	mw
value	50	0.39	16	0.33	1

proportion, accounting for 29.72%; in terms of the time dimension, most of the characteristics are passengers' long-term travel preferences or information, accounting for 45.9% of data within 2 years and 21.62% within 3 years. It can be seen that airlines should focus on studying the long-term travel characteristics of passengers with different preferred routes and cities to set up accurate marketing of paid seat selection auxiliary services.

As a result, a complete and reduced-dimensional processed dataset with 37 input features and a binary label is formed and set as final input into the P-layer for training and prediction.

E. PREDICTION LAYER

Based on the complete and low-dimensional processed dataset, prediction of consumption willingness about seat selection additional services is performed through XGBoost in the P-layer. First, the hyperparameters of XGBoost are specified by the PSO algorithm, and the average model accuracy through 10-fold cross-validation is used as the fitness function value to guide the evolutionary direction of the particle swarm. The final optimal model obtained is applied to predict new airline passengers' willingness to purchase seat selection additional services.

IV. RESULTS AND DISCUSSION

The performance measures are essential instruments to evaluate the reliability and validity of models. Five commonly used evaluation metrics are selected to represent the effect of setting IDP-layer, HDP-layer, and using PSO to specify hyperparameters in the evaluation phase. Finally, the triple-layer PSO-XGBoost model is compared with other existing single-layer machine learning models.

A. EVALUATION METRICS AND METHODS

Five commonly used evaluation indexes for classification problems are selected for various evaluations, including accuracy (Acc), precision (Pre), recall (Rec), F1 score (F1), and area under the ROC curve (AUC).

However, Acc is impractical if the sample distribution is unbalanced as well as Pre, Rec, and F1 lack comprehensive reflection of model performance. Apart from them, AUC is recognized as a reliable tool for evaluating several machine learning models in several situations. Therefore, among all these metrics, this study primarily focuses on AUC value.

To avoid the impact of dividing the dataset on model performance evaluation, the training time overhead and model performance of the model under different division ratios were considered [21]. A 10-fold cross-validation method was adopted to evaluate the performance of the model, dividing the dataset into 90% training set and 10% verification set. The model was trained and verified 10 times, and the average was used to represent the score of the model on each metric, respectively.

B. EFFECT EVALUATION OF INCOMPLETE DATA PROCESSING LAYER

To verify the validity of the IDP-layer, datasets are processed by different methods and used to train the XGBoost model, and the experimental results are shown in TABLE 3. According to the results, the Rec value of XGBoost is significantly improved on SMOTE-balanced datasets, which indicates that balanced datasets highly enhance the model. Meanwhile, it can be seen that the missing value imputation method based on PSO-XGBoost is superior to the traditional imputation method, especially with AUC score reaching 0.9587 on the balanced dataset, which shows that the quality of the dataset has been improved after being processed by the IDP-layer.

C. EFFECT EVALUATION OF HIGH-DIMENSIONAL DATA PROCESSING LAYER

To evaluate the effect of dimensionality reduction on the dataset, the XGBoost model was trained using the dataset before and after the HDP-layer, and the model performance is shown in TABLE 4. The dataset after dimensionality reduction by the HDP-layer can significantly reduce the training time overhead while keeping the model performance stable.

D. EVALUATION OF THE EFFECT OF PSO OPTIMIZATION

To evaluate the effectiveness of the PSO algorithm in optimizing the XGBoost model, the performance of models using different optimization methods on the processed dataset is shown in TABLE 5. The graph shows that the performance of the model with default hyperparameter is poor, while all metrics of the optimized model are improved. Among them, the performance of the hyper-parameters obtained by the PSO algorithm is equivalent to that of the fine traversal search method, but the search time is greatly reduced, which improves the efficiency of model optimization.

The optimal hyperparameter values searched by the PSO algorithm are shown in TABLE 6.

E. EFFECT COMPARISON OF TRIPLE-LAYER PSO-XGBoost MODEL

The triple-layer PSO-XGBoost model is compared with other widely used machine learning models, including Logistic Regression [22], Random Forest [23], BP Neural Network [24], Naive Bayes [25], Decision Tree [26], Support Vector Machine, K Nearest Neighbor, Long Short Term Memory and single-layer XGBoost model to verify the effectiveness of the proposed method, and the basic dataset used is only cleaned, encoded and mean imputed. The performance of the model is shown in TABLE 7.

As can be seen from TABLE 7, the performance of the latter three ensemble tree-based models is excellent, with

TABLE 4. Comparison of performance on datasets with different processing methods.

Imputation method	Balance	Acc	F1	Pre	Rec	Auc
Mean value	Imbalance	0.9385	0.1134	0.6293	0.0626	0.8098
Mean value	Balance	0.8438	0.8485	0.8241	0.8747	0.9356
PSO-XGBoost	Imbalance	0.9404	0.2641	0.7067	0.0942	0.8386
PSO-XGBoost	Balance	0.8827	0.8835	0.8778	0.8895	0.9587

TABLE 5. Comparison of performance on datasets with different number of features.

Number of features	Acc	F1	Pre	Rec	Auc	Time(s)
137	0.8827	0.8835	0.8778	0.8895	0.9587	26.7449
37	0.8881	0.8839	0.9189	0.8514	0.9636	9.8101

Note: Time(s) is the total time of 10-fold cross-validation.

TABLE 6. Performance comparison of XGBoost with different optimization method.

Optimization method	Acc	F1	Pre	Rec	Auc	Time(s)
None	0.8881	0.8839	0.9189	0.8514	0.9636	0
Traversal search	0.9571	0.9569	0.9624	0.9515	0.9895	> 3,600
PSO algorithm	0.9520	0.9517	0.9584	0.9451	0.9879	388.86

Note: Time(s) is the time for searching optimal hyperparameters.

TABLE 7. Comparison of performance between triple-layer PSO-XGBoost model and other machine learning models.

Model	Acc	F1	Pre	Rec	Auc
Logistic Regressor	0.9361	0.0026	0.0583	0.0013	0.4564
BP Neutral Network	0.9369	0.0720	0.3996	0.0405	0.7561
Gaussian Naïve Bayes	0.1419	0.1187	0.0634	0.9185	0.5355
Decision Tree	0.9148	0.3444	0.3343	0.3564	0.6792
Random Forest	0.9416	0.3483	0.5831	0.2489	0.8242
Support Vector Machine	0.9372	0.0067	0.4000	0.0034	0.7041
K Nearest Neighbor	0.9305	0.1094	0.2881	0.0680	0.6983
Long Short Term Memory	0.9370	0.0040	0.3000	0.0020	0.7789
Single XGBoost	0.9373	0.0517	0.5505	0.0272	0.8041
Triple layer PSO-XGBoost	**0.9520**	**0.9517**	**0.9584**	**0.9451**	**0.9879**

Note: Bold is the maximum value of each column.

AUCs above 0.8. The proposed triple-layer PSO-XGBoost model shows the most outstanding performance with a better metric of 0.9879 in terms of AUC, which fully proves that the proposed method can accomplish the prediction task well on a high-dimensional and incomplete dataset. The AUCs of both the linear model-based Logistic Regression and the Naive Bayesian-based on the assumption of inter-feature independence are lower, indicating that there is an obvious non-linear relationship between passengers' consumption willingness towards seat selection additional services, the 37 features mentioned, and the correlation between the features. In the comparison models, there are better Acc and worse Rec for all models except the Naive Bayesian model, indicating that the model classifies the vast majority of samples as negative for not purchasing additional services, while the Naive Bayesian classifies the majority of samples as positive. The comparison of the models illustrates that the untargeted dataset causes great fluctuations in model performance, which further confirms the stability of the proposed triple-layer model.

V. CONCLUSION

In this paper, a prediction method of airline additional services consumption willingness based on a triple-layer XGBoost model optimized by PSO is proposed, and the following conclusions are drawn.

(1) In the IDP-layer, compared with the unbalanced dataset imputed by traditional methods, the dataset imputed by the PSO-XGBoost model and SMOTE balanced has significantly improved prediction performance.
(2) In the HDP-layer, the XGBoost model is used for feature extraction and dimension reduction, which greatly reduces the running time and improves efficiency while ensuring the stability of the model.
(3) The modern heuristic PSO algorithm is used to optimize the XGBoost model. Compared with the traditional traversal search method, the time cost is greatly reduced, and the performance is significantly optimized based on the XGBoost model.

(4) Comparative experimental results show that the proposed triple-layer PSO-XGBoost model is better than single-layer XGBoost and other widely used machine learning models such as BP neural network. The proposed model not only greatly improves the prediction accuracy, but also reduces the training time expenditure, which can meet the demand of passenger additional service willingness prediction based on civil aviation passenger information big data system.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper significantly.

ACKNOWLEDGMENT

The authors thank Neusoft for providing an air passenger travel dataset. They would also like to thank the reviewers for their many insightful comments, which helped to improve this paper significantly.

DATA AVAILABILITY STATEMENTS

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] C. Morosan, "Understanding the benefit of purchasing ancillary air travel services via mobile phones," *J. Travel Tourism Marketing*, vol. 32, no. 3, pp. 227–240, Apr. 2015, doi: 10.1080/10548408.2014.896763.

[2] Y. Zhou, T. Zhang, Y. Mo, and G. Huang, "Willingness to pay for economy class seat selection: From a Chinese air consumer perspective," *Res. Transp. Bus. Manage.*, vol. 37, Dec. 2020, Art. no. 100486, doi: 10.1016/j.rtbm.2020.100486.

[3] C. Rice, N. K. Ragbir, S. Rice, and G. Barcia, "Willingness to pay for sustainable aviation depends on ticket price, greenhouse gas reductions and gender," *Technol. Soc.*, vol. 60, Feb. 2020, Art. no. 101224, doi: 10.1016/j.techsoc.2019.101224.

[4] P. Chiambaretto, "Air passengers' willingness to pay for ancillary services on long-haul flights," *Transp. Res. E, Logistics Transp. Rev.*, vol. 147, Mar. 2021, Art. no. 102234, doi: 10.1016/j.tre.2021.102234.

[5] H. Giao "Customer satisfaction of Vietnam airline domestic services," OSF, Chennai, India, Tech. Rep., 2017, doi: 10.31219/osf.io/dyze3.

[6] S. Roy, D. Kaul, R. Barna, S. Mehta, and A. Misra, "Prediction of customer satisfaction using naive Bayes, multiclass classifier, K-star and IBK," *Soft Comput. Appl.*, vol. 634, pp. 153–161, Oct. 2016, doi: 10.1007/978-3-319-62524-9_12.

[7] R. Totamane, A. Dasgupta, and S. Rao, "Air cargo demand modeling and prediction," *IEEE Syst. J.*, vol. 8, no. 1, pp. 52–62, Oct. 2014, doi: 10.1109/JSYST.2012.2218511.

[8] G. Li, W. Yuan, and H. C. Wang, "Civil aviation passenger loss prediction model for incomplete data," *Comput. Eng. Des.*, vol. 40, no. 10, pp. 2884–2891, 2020, doi: 10.16208/j.issn1000-7024.2020.10.031.

[9] D. A. Dias Jánior, L. B. da Cruz, J. O. Bandeira Diniz, G. L. França da Silva, G. B. Junior, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass, "Automatic method for classifying COVID-19 patients based on chest X-ray images, using deep features and PSO-optimized XGBoost," *Expert Syst. Appl.*, vol. 183, Nov. 2021, Art. no. 115452, doi: 10.1016/j.eswa.2021.115452.

[10] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network intrusion detection based on PSO-XGBoost model," *IEEE Access*, vol. 8, pp. 58392–58401, 2020, doi: 10.1109/ACCESS.2020.2982418.

[11] X. Zhang, H. Nguyen, X.-N. Bui, Q.-H. Tran, D.-A. Nguyen, D. T. Bui, and H. Moayedi, "Novel soft computing model for predicting blast-induced ground vibration in open-pit mines based on particle swarm optimization and XGBoost," *Natural Resour. Res.*, vol. 29, no. 2, pp. 711–721, Apr. 2020, doi: 10.1007/s11053-019-09492-7.

[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[13] W. XingFen, Y. Xiangbin, and M. Yangchun, "Research on user consumption behavior prediction based on improved XGBoost algorithm," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2018, pp. 4169–4175, doi: 10.1109/BigData.2018.8622235.

[14] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE ICNN*, vol. 4. Nov./Dec. 1995, pp. 1942–1948, doi: 10.1109/ICNN.1995.488968.

[15] K. Koc, Ö. Ekmekcioèlu, and A. P. Gurgun, "Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers," *Autom. Construct.*, vol. 131, Nov. 2021, Art. no. 103896, doi: 10.1016/j.autcon.2021.103896.

[16] S. S. Roy, P. Samui, R. Deo, and S. Ntalampiras, *Big Data in Engineering Applications*, vol. 44. Berlin, Germany: Springer, 2018.

[17] S. P. RM, P. K. R. Maddikunta, M. Parimala, S. Koppu, T. R. Gadekallu, C. L. Chowdhary, and M. Alazab, "An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture," *Comput. Commun.*, vol. 160, pp. 139–149, Jul. 2020.

[18] D. Micci-Barreca, "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems," *ACM SIGKDD Explor. Newslett.*, vol. 3, no. 1, pp. 27–32, Jul. 2001, doi: 10.1145/507533.507538.

[19] F. Tang and H. Ishwaran, "Random forest missing data algorithms," *Stat. Anal. Data Mining: ASA Data Sci. J.*, vol. 10, no. 6, pp. 363–377, Dec. 2017, doi: 10.1002/sam.11348.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002, doi: 10.1613/jair.953.

[21] Y. Zhang and Y. Yang, "Cross-validation for selecting a model selection procedure," *J. Econometrics*, vol. 187, no. 1, pp. 95–112, 2015, doi: 10.1016/j.jeconom.2015.02.006.

[22] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, no. 1, pp. 1–22, 2010, doi: 10.18637/jss.v033.i01.

[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[24] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Netw.*, vol. 21, nos. 2–3, pp. 427–436, 2008, doi: 10.1016/j.neunet.2007.12.031.

[25] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naíve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy," *Appl. Environ. Microbiol.*, vol. 73, no. 16, pp. 5261–5267, Aug. 2007, doi: 10.1128/AEM.00062-07.

[26] G. De'Ath and K. E. Fabricius, "Classification and regression trees: A powerful yet simple technique for ecological data analysis," *Ecology*, vol. 81, no. 11, pp. 3178–3192, 2000, doi: 10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.