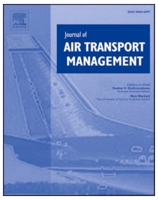
Contents lists available at ScienceDirect

Journal of Air Transport Management

journal homepage: www.elsevier.com/locate/jairtraman



The adverse impact of flight delays on passenger satisfaction: An innovative prediction model utilizing wide & deep learning

Cen Song [a,*], Xiaoqian Ma [a], Catherine Ardizzone [b], Jun Zhuang [b]

[a] *School of Economics and Management, China University of Petroleum, Beijing, 102249, China*
[b] *Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, NY, 14260, USA*

ARTICLE INFO

Keywords:
Flight delay
Feature interaction
Passenger satisfaction prediction
Wide & deep model
Interpretability analysis

ABSTRACT

This article addresses the substantial negative influence of flight delays on passenger satisfaction and aims to bridge the research gap in understanding passenger satisfaction during delayed flights. We present a passenger satisfaction prediction model leveraging a real dataset from Kaggle. Through an examination of the interplay between individual in-flight services and passenger characteristics using the Pearson correlation coefficient and PCA-K-means clustering methods, we introduce a novel satisfaction prediction model built upon the deep learning Wide & Deep algorithm. Additionally, we employ the DeepLIFT algorithm to interpret the deep learning model and elucidate the salient features impacting passenger satisfaction, as revealed through feature importance analysis. Our findings demonstrate that the prediction model outperforms benchmark models such as MLP, SVM, and Random Forest, achieving higher accuracy. This study contributes to an enhanced comprehension of the multifaceted factors influencing passenger satisfaction following flight delays, and it offers valuable insights and recommendations for the enhancement of service quality among airline companies.

1. Introduction

With the rapid growth of the global civil aviation industry, flight delays have seriously impacted passengers, airlines, and aviation authorities. According to statistics, the annual economic loss caused by flight delays is in the billions of dollars worldwide, including increased operating costs for airlines, time costs for passengers, and surcharges (Liu et al., 2019). Studies in China have shown that flight delays significantly negatively impact passenger satisfaction and loyalty (Du and Zhang, 2020). According to the Civil Aviation Administration of China's consumer complaint notification for the 12 months of 2022, flight delays are one of the most worrying issues for passengers, negatively affecting both the airline's reputation and customer satisfaction (CAAC, 2023).

To improve service quality, airlines need to consider each link that the customer experiences throughout the service chain, since these activities have a direct impact on service quality and the final satisfaction outcome. Passengers are more inclined to judge airlines based on their satisfaction with the in-flight service provided by the flight attendants and aircraft facilities, which are most directly consumer-facing (Park et al., 2004). It is clear that inflight service is a key part of the airline service chain and improving the quality of inflight service are one of the drivers of airline success.

In a delayed flight scenario, passengers make many travel decisions, such as waiting for the original flight, changing tickets, obtaining refunds of tickets, and changing transportation modes. For passengers who choose to continue to wait for the original flight and take the delayed flight, their satisfaction evaluation of the delayed onboard service is valuable for the airline's service strategy. This paper presents several innovations. Firstly, in investigating the impact of flight delays on passenger satisfaction, this paper emphasizes the significance of individual passenger satisfaction research, highlighting that existing research primarily focuses on satisfaction related to airports and airlines. Secondly, this paper constructs a predictive model based on the Wide & Deep algorithm, which effectively captures interactions among different features in the data for more accurate prediction of passenger satisfaction. Third, this paper examines the interpretability of prediction results, investigating feature importance and their influence on predictions. The application of such a model not only offers more accurate passenger satisfaction prediction but also provides a basis for airlines to formulate corresponding improvement measures.

The rest of this paper is organized as follows: Section 2 provides a

* Corresponding author. No.18 Fuxue Road, Changping District, Beijing, 102249, China.
E-mail addresses: songcen22@126.com (C. Song), 15230258417@163.com (X. Ma), crardizz@buffalo.edu (C. Ardizzone), jzhuang@buffalo.edu (J. Zhuang).

https://doi.org/10.1016/j.jairtraman.2023.102511
Received 27 May 2023; Received in revised form 8 October 2023; Accepted 18 October 2023
Available online 26 October 2023

literature review. Section 3 shows the data sources, the pre-processing process, and the classification of passengers into categories through data feature analysis. Section 4 presents the feature interaction-based Wide & Deep satisfaction prediction model and performs model optimization and performance comparison experiments. Section 5 presents a model interpretability analysis of the prediction results based on the DeepLIFT algorithm. Section 6 concludes and provides some future research directions.

2. Literature review

2.1. Satisfaction with in-flight service

There are numerous theoretical models related to in-flight service satisfaction research, with scholars primarily exploring dimensions such as service quality, expectation levels, and perceived value. Among these, the service quality model posits that factors like reliability, responsiveness, assurance, empathy, and tangible factors significantly impact passenger in-flight service satisfaction (Ryu et al., 2019). The expectation level model suggests a close correlation between passengers' expected service levels and their satisfaction. The perceived value model emphasizes the pivotal role of passengers' perceived value o in-flight services in evaluating satisfaction. Similarly, the SERVQUAL model (Giao & Vuong, 2021), a classic service quality assessment model, applies to airline in-flight service satisfaction research. This model evaluates the influence of service quality on customer satisfaction by measuring the gaps between customers' expectations and actual experiences across different dimensions of service quality.

In the study of in-flight service satisfaction, traditional statistical methods are commonly used for data collection and assessing passengers' satisfaction with various services. For researchers focusing on in-flight service satisfaction, they must also take into account, not only factors like service quality, expectation levels, and perceived value but also the influence of cultural, gender, age, and occupational factors on in-flight service satisfaction. Given the subjective nature of passenger evaluations of in-flight services, objectively and scientifically evaluating and measuring such satisfaction is a challenging aspect of research. In recent years, researchers have begun to use methods such as machine learning and natural language processing to predict and analyze in-flight service satisfaction (Walia et al., 2021). These methods not only allow for more efficient processing of large-scale data but also uncover potential correlations in the data and improve prediction accuracy.

2.2. Feature interaction based on the wide & deep model

Feature interaction as a key feature engineering method has been widely used in various machine learning tasks (Almuqren & Cristea, 2023). In the field of deep learning, feature interaction is usually implemented by introducing an interaction layer, which allows the model to automatically learn complex relationships between features, improving performance and expressiveness. The Wide & Deep model is a machine learning model that combines the Generalized Linear Model (GLM) and Deep Neural Network (DNN), proposed by Google in 2016 and applied to the field of recommendation systems (Cheng et al., 2016). Through feature interaction, the Wide & Deep model can explore the non-linear relationship between features and improve the accuracy of the model (Tompson et al., 2014). In the Wide & Deep model, feature interaction is mainly implemented through the embedding layer and the fully connected layer, where the combination and intersection of features can be achieved through different activation functions and weight matrices (Wang et al., 2015). For example, Pei et al. (2017) introduced a recommendation model based on the Interacting Attention-gated Recurrent Network (IARN), an approach that effectively captures the interactions between users and items and organizes features into a flat and hierarchical structure to enhance recommendation performance. Tang et al. (2023) showed that the accuracy and performance of user behavior prediction could be improved by graph propagation of the user node feature matrix and the relational network adjacency matrix, fusing the interaction information between different features. Similarly, Wilson et al. (2021) used feature interactions in a classification prediction model and proposed the Wide & Deep model, which has better prediction results.

Feature interactions can be used not only in deep learning models but also in traditional machine learning models. For example, Konstantinov and Utkin (2021) proposed Gradient Boosting Machines (GBMs) model, which can improve model performance by iteratively adding feature interaction terms. In addition, feature interactions can be used in traditional linear models. For example, Yan et al. (2020) proposed an improved Factorization Machines (FM) model that can solve the problem of high-dimensional sparse data by introducing feature interaction terms. In addition to improving model performance, feature interactions can also improve the interpretability of a model. For example, in an ad click-through prediction task, feature interactions can help to determine the relationships between different features and thus better understand the prediction results (Deng et al., 2020). Feature interactions can also help to discover the interactions between different features and thus better understand the mechanisms behind the prediction results (Li et al., 2022). Therefore, feature interaction is of high practical value in the field of machine learning.

2.3. Model interpretability analysis

With the increasing application of artificial intelligence, model interpretability analysis techniques are gaining more and more attention (Rajapaksha et al., 2020). Model interpretability refers to the process of explaining the inner workings of a model and the decision rules that enable humans to understand the reasons for the model's results.

Currently, common model interpretability techniques include feature importance analysis, decision tree interpretation, LIME, SHAP, and DeepLIFT. The application scenarios of model interpretability technology are very broad, including finance, healthcare, e-commerce, smart manufacturing, and many other fields (Barredo Arrieta et al., 2020). In the financial sector, model interpretability techniques can help banks and financial institutions understand the decision rules and results of models for credit assessment and other operations so that they can better control risk and approve loans (de Lange et al., 2022). In the medical field, model interpretability technology can help doctors and medical institutions understand the decision rules of models for disease diagnosis and drug treatment, improving the accuracy of diagnosis and treatment effectiveness (Yang et al., 2022). In the e-commerce sector, model-interpretable technologies can help companies understand the patterns and factors of consumer buying behavior and optimize product recommendations and sales strategies (Zou and Pang, 2022). In the field of intelligent manufacturing, model interpretability technology can help engineers and manufacturing companies understand the decision rules of models for quality control and fault detection in the production process, and improve production efficiency and product quality (Molnar et al., 2020).

In general, research on the interpretability of predictive models significantly impacts various application scenarios. These studies offer valuable insights and methods to improve the interpretability of predictive models. Analyzing the significance of model features and utilizing diverse interpretable methods can enhance the interpretability of predictive models, rendering the model's predictions more understandable and accepted. Moving forward, research into the interpretability of predictive models will persist, yielding novel interpretability analysis methods and use cases.

This paper presents a review of research regarding passenger satisfaction following flight delays, comprehensively examining the influence of in-flight services on passenger contentment and related studies. In terms of satisfaction prediction methods, traditional methods and machine learning-based approaches, including deep learning, are widely discussed. Traditional methods heavily rely on statistical analysis and surveys, which can partially reveal key satisfaction factors but lack flexibility and accuracy. On the other hand, machine learning-based approaches such as the Wide & Deep model can offer more precise predictions of passenger satisfaction by considering feature interactions. However, these approaches pose challenges in interpreting prediction results since they often function as black-box models, making it difficult to understand their decision-making processes and underlying mechanisms. To enhance model interpretability, researchers have extensively studied methods to interpret predictive models. Approaches such as LIME, DeepLIFT, and LORE furnish avenues to elucidate prediction outcomes. These methods use local explanations, feature importance analysis, and rule-based approaches to help users understand the model's prediction process and outcomes.

In light of this, the objective of this paper is to develop a satisfaction prediction model grounded in feature interactions and to conduct model interpretability analysis through the application of the DeepLIFT algorithm. Simultaneously, corresponding service enhancement strategies will be proposed based on the analysis results, aiming to assist airlines in targeted efforts to enhance passenger satisfaction.

3. Data preprocessing and character analysis

3.1. Data collection

The airline passenger satisfaction dataset on Kaggle contains 24 features including passenger ID, gender, customer type, age, travel type, class of cabin, flight distance, departure delay (minutes), arrival delay (minutes), departure airport, arrival airport, booking type, online booking convenience, gate location, food and beverage, online boarding, seat comfort, in-flight entertainment, in-flight service, leg room, baggage handling, check-in service, cleanliness, country of departure and country of arrival. There are 129,880 data items in this dataset with no missing values. The passenger ratings for service span from 0 to 5 (see Table 1).

The subject of this paper is the prediction of passenger satisfaction with in-flight service after delays. The dataset contains journey characteristics such as delayed departure times and delayed arrival times of flights as well as passenger ratings of various aspects of in-flight service. These characteristics are all related to satisfaction with in-flight service after delays. For example, delayed departures and arrivals may lead to poorer passenger moods, which may affect their ratings of service. In contrast, ratings of various aspects of in-flight service in the context of continued travel on a delayed flight directly reflect passenger satisfaction. Therefore, all these characteristics can be used to predict passenger satisfaction with in-flight service after a delay.

Table 1
Dataset of features and attributes.

check DT

Features	No.	Field name	Attributes
Data ID	–	ID	String
Passenger	0	Gender	String
Characteristics	1	Customer type	String
	2	Age	Numerical
	3	Type of travel	String
Trip features	4	Class	String
	5	Flight distance	Numerical
Service score	6	Inflight Wi-Fi service	Numerical
	7	Departure/Arrival time convenient	Numerical
	8	Ease of Online booking	Numerical
	9	Gate location	Numerical
	10	Food and drink	Numerical
	11	Online boarding	Numerical
	12	Seat comfort	Numerical
	13	Inflight entertainment	Numerical
	14	On-board service	Numerical
	15	Leg room service	Numerical
	16	Baggage handling	Numerical
	17	Check in service	Numerical
	18	Inflight service	Numerical
	19	Cleanliness	Numerical
Delayed characteristics	20	Departure delay in minutes	Numerical
	21	Arrival delay in minutes	Numerical
Satisfaction results	22	Satisfaction	String

→ 0→5

3.2. Data preprocessing

3.2.1. Data cleaning

As the ID of a recorded passenger is unique and will not appear in subsequent studies, the ID column at the beginning of the data set is removed first. As the ID columns are removed, the data will then need to be converted into vector form using the matrix method and then stitched together to form the required data matrix.

In the dataset, the feature Age has outliers. Specifically, there are negative values and oversized values for the Age feature. These outliers may be caused by data entry errors or incorrect information provided by the passenger. Statistically, the percentage of negative values is found to be 0.36%; the percentage of passengers with an age greater than or equal to 100 years is 0.14%. Therefore, the proportion of outliers is relatively small and we remove the outlier samples for data cleaning.

3.2.2. Data pre-processing

In order to facilitate the subsequent fit of the data to the model, several features and data types in the dataset need to be adapted and processed to change the structure of certain elements to make the data easier to use.

Initial processing of the string type data to convert it to a numeric type with five features, including gender, loyalty, trip category, cabin category, and satisfaction results. Further, the gender is set as male = 0 and female = 1 respectively; the loyal customer characteristics correspond to Loyal customer = 0 and Disloyal customer = 1; the trip category is set as Business travel = 0 and Personal travel = 1; the cabin class is set as Eco = 0, Eco Plus = 1, and Business = 2; satisfaction results are 0 for satisfied and 1 for neutral or unsatisfied. After that, the proportion of satisfied passengers and the number of dissatisfied passengers is not significantly different, which is close to the balance, and the classification can be regarded as a balanced classification, which is conducive to the subsequent model training.

?

3.2.3. Normalization

Min-max normalization is applied to normalize the three numeric types of data, flight distance, departure delay minutes, and arrival delay minutes in the [0,1] interval. This effectively reduces the impact of attribute data with larger magnitudes on the model predictions or classification results.

3.2.4. Data filtering based on delayed scenarios

Departure delay or arrival delay both belong to the flight delay scenario, so the target data can be filtered using the judgment of departure delay in minutes $\neq 0$ or arrival delay in minutes $\neq 0$. The training and test sets are divided using the ten-fold cross-validation method.

?

3.3. Feature analysis

3.3.1. Pearson correlation analysis

To explore the relationships between passenger-related features, we use heat maps to display the correlation coefficient matrix. Fig. 1 shows the correlation coefficient matrix among the 23 dataset features, including passenger features, journey features, service ratings, delay features, and satisfaction outcome features, after data processing.

Based on the heat map shown in Fig. 1, we can conclude that there are varying degrees of correlation between the different attribute features. This indicates that these characteristics are not independent of

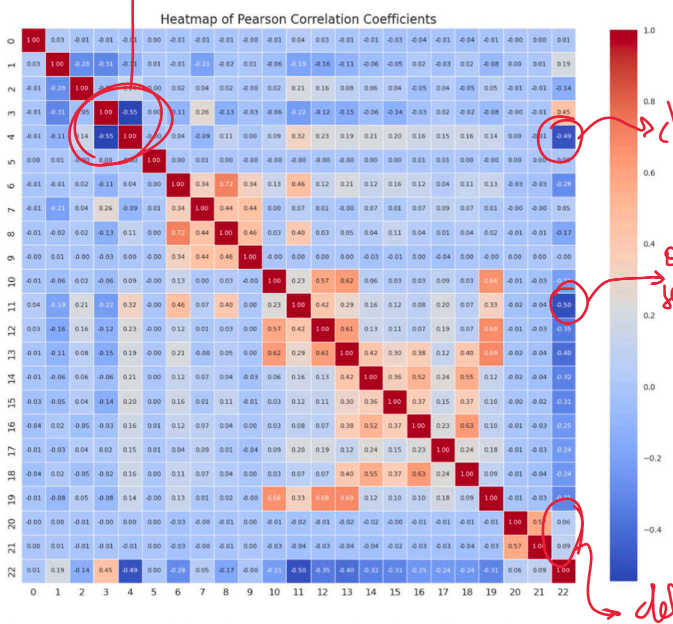


Fig. 1. Heat map of the characteristics of each attribute of passenger satisfaction with in-flight service.

each other, but that there are relationships that influence each other, and that there are many such characteristic correlations. For example, in terms of passenger characteristics, there is a strong negative correlation between the Type of Travel and Class in the trip characteristics. This suggests that there is a link between passenger purpose of travel and cabin class, with business trips more likely in the premium class category, while leisure trips are to be likely in the economy class category. Therefore, the correlation between passenger characteristics and journey characteristics should be fully considered in a satisfaction prediction model to ensure that the model more accurately captures the impact of passenger travel purpose and cabin category on satisfaction. As for another example, from a service rating perspective, a significant positive correlation is shown between Inflight Wi-Fi service and Ease of Online booking. This can be interpreted to mean that the provision of good in-flight Wi-Fi service by the airline has a positive impact on the convenience of the passenger's online booking experience.

We analyzed the correlations between passenger characteristics, journey characteristics, service ratings, delay characteristics, and satisfaction outcome characteristics through heat maps, and it can be seen that there is a significant interaction between the different characteristic attributes. This interplay is critical to the construction and performance evaluation of satisfaction prediction models. Therefore, selecting algorithmic models that enable feature interactions within the model can better understand and exploit the correlations between features and improve the predictive performance and interpretation of the model, resulting in more accurate predictions of passenger satisfaction.

3.3.2. Passenger characteristics analysis

Some studies have focused on the impact of individual passenger characteristics on satisfaction. For instance, research (Walia et al., 2021) indicates significant differences in passenger responses to delayed flights based on factors such as age, gender, occupation, and travel purpose. Therefore, during the preliminary data exploration and analysis, it is essential to categorically explore passengers, studying common features within each group. This aids in better understanding the needs and preferences of different passenger types, subsequently targeting strategies to enhance their satisfaction, thereby offering valuable insights and support for airline operational management.

This paper employs a combination of Principal Component Analysis (PCA) and K-means clustering methods to partition and analyze passenger characteristics. PCA, a statistical technique, helps identify key variables within high-dimensional datasets, explaining observed differences and simplifying analysis and visualization without significant information loss. By integrating PCA with the K-means algorithm, effective clustering analysis can be conducted while reducing data dimensionality. The clustering results are illustrated in Fig. 2.

?

From the visualized clustering results, it is evident that a substantial amount of passenger data has been effectively divided into four categories with distinct boundaries through the integration of Principal Component Analysis and K-means clustering. This suggests a certain relationship between the various characteristics and service evaluation scores of passengers on delayed flights and their final satisfaction evaluations. Similar high-scoring features are clustered together.

To further analyze multi-feature behaviors, this paper combines the previous clustering results and employs Principal Component Analysis, integration, and clustering techniques. By extracting the most significant principal components, features with substantial information are identified from various attributes. Within the PCA analysis, the top five contributing dimensions are selected and processed in descending order. These dimensions are online boarding, seat comfort, cleanliness, inflight entertainment, and age. Then, 100 data points from each category are randomly selected and the corresponding feature-matching maps are generated as shown in Fig. 3. It provides a clearer visualization of user similarity features, helping us better understand the attributes of each cluster. Since the dataset is sourced from an airline company, we can further synthesize information across feature dimensions to summarize the characteristics of each passenger category, as well as the varying degrees of demand for the attributes of economy, efficiency, and comfort.

As a result, the four passenger categories are respectively denoted as A, B, C, and D types. Category A passengers are predominantly youth, with their satisfaction evaluations for inflight entertainment and cleanliness mainly concentrated between 4 and 5, while satisfaction with online boarding and seat comfort mostly falls within 1–3. Category B passengers are middle-aged and elderly, with their evaluations for seat comfort mainly ranging from 3 to 4, generally indicating moderate satisfaction. However, the other three features - entertainment service, online boarding convenience, and inflight environment - are generally expected to meet higher standards. Category C passengers are mainly adolescents and youth, with their evaluations for all four dimensions generally concentrated between 1 and 3, indicating their higher requirements for convenience, comfort, cleanliness, and entertainment. Category D passengers are middle-aged, with generally higher evaluations for various onboard services post-delay. However, their satisfaction with online boarding is relatively concentrated between 2 and 4,

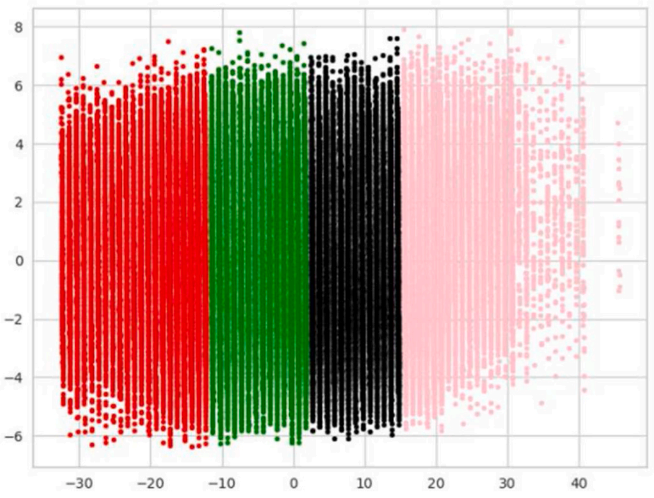
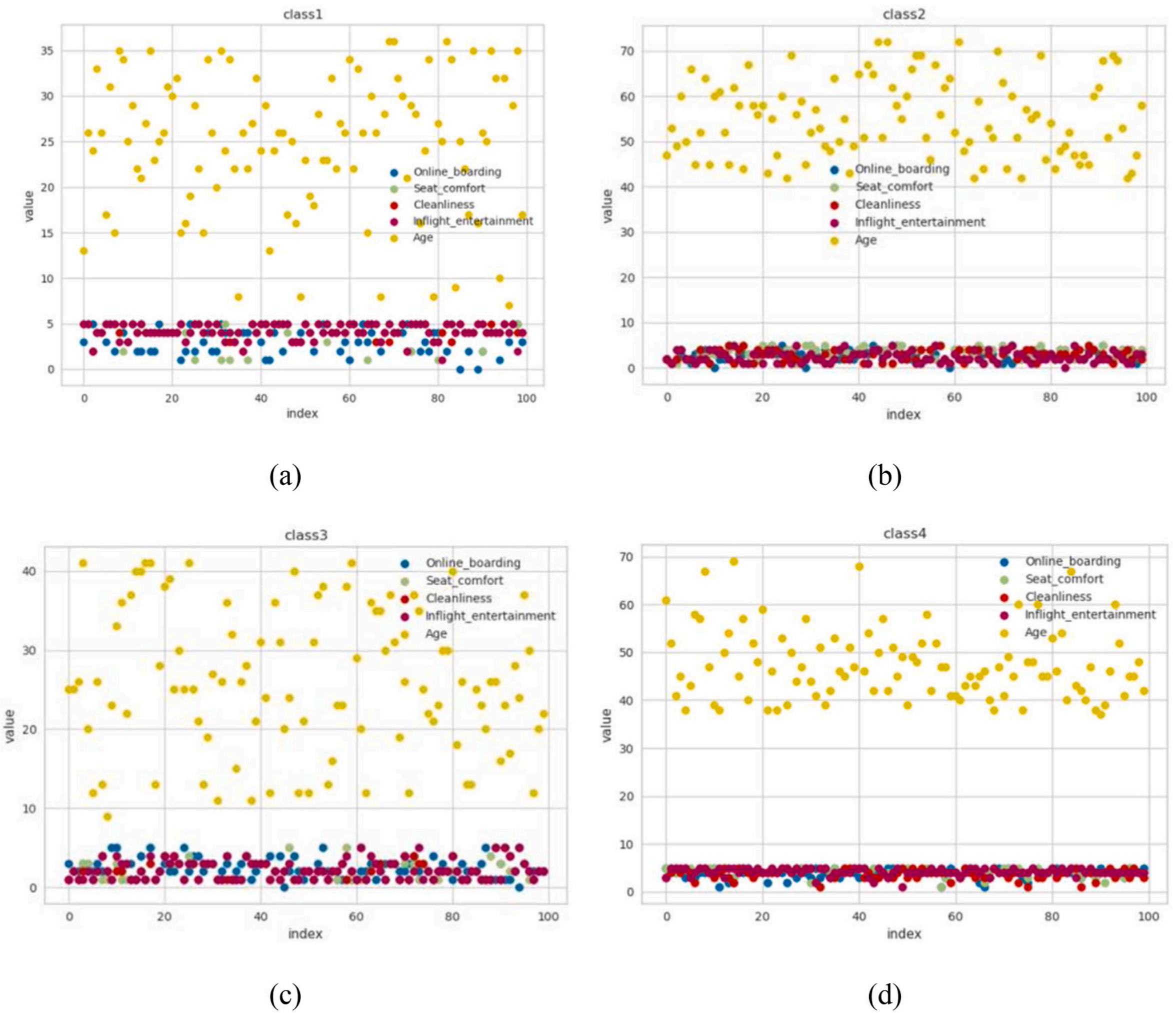


Fig. 2. PCA-based K-means clustering results.



(a)

(b)

(c)

(d)

Fig. 3. Feature matching plot for four categories of passengers based on PCA.

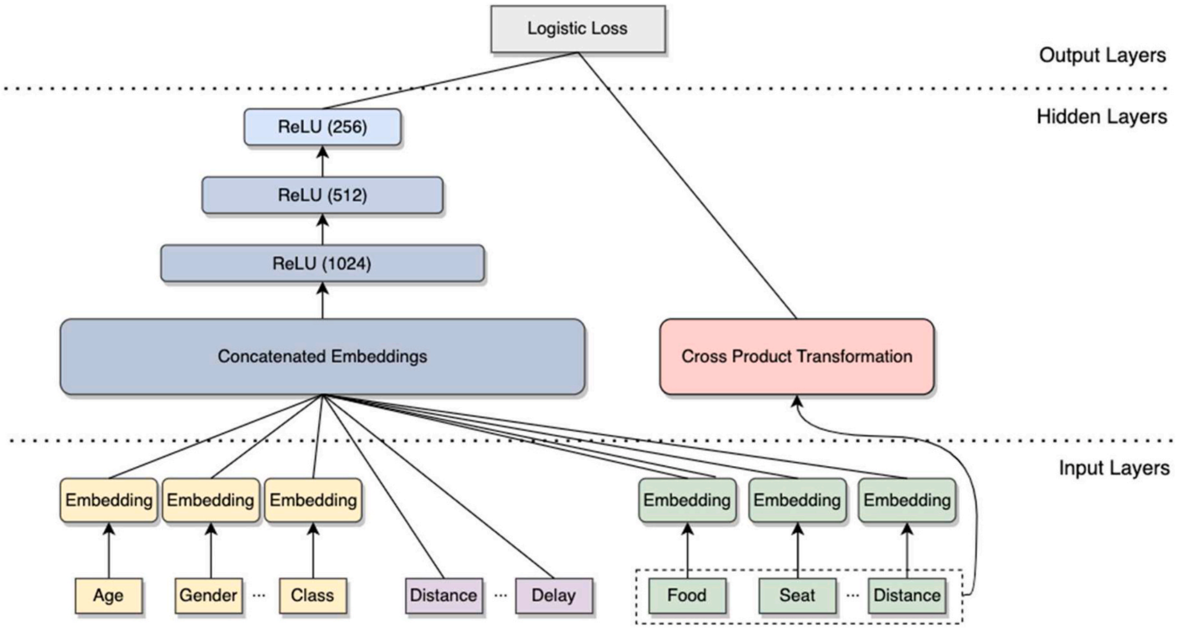


Fig. 4. Based on Wide & Deep's satisfaction prediction network structure.

and they demonstrate lower demands for comfort and entertainment aspects.

4. A satisfaction prediction model based on feature interactions

4.1. Model construction

Based on a large number of features, the Wide & Deep model with better fitting ability and higher prediction accuracy is used for predicting passenger satisfaction.

(1) Input layer: As shown in Fig. 4, the input layer of the model mainly takes passenger features, journey features, and rating features of in-flight services after delay as input features. Among them, categorical features are converted into their respective representation vectors through an embedding layer. The embedding layer for each feature input is a single-layer MLP, and the parameters are not shared.

(2) Hidden layer: The hidden layer receives the representation vectors from the input layer. In the left Deep side, all the input vectors are concatenated before being fed into a three-layer MLP network. The vector dimensions of each MLP layer are 1024, 512, and 256, and the ReLU function is used as the activation function for non-linear transformation.

(3) Output Layer: The output layer combines the outputs from the Deep and Wide sides of the hidden layer to obtain the final satisfaction prediction result of the model. At the same time, the model adopts a joint training strategy to train the left Deep model and the right Wide model simultaneously, and the weighted sum of the results from the two models is used as the final prediction result.

4.2. Model optimization

4.2.1. Hyperparameter tuning

In machine learning models, there are two types of parameters: one type can be learned from the data, known as parameters; the other type cannot be estimated from the data and must be set through experience, known as hyperparameters. These parameters need to be set before the model learning process begins. This paper employs the Hyperopt tool to implement Bayesian optimization for hyperparameters. The basic steps are as follows:

(1) Define an objective function for Hyperopt based on the model's loss function. The goal is to optimize hyperparameters with respect to this objective function.

(2) Define the search space for hyperparameters, specifying the search range for each hyperparameter.

(3) Specify the search algorithm; in this case, the default is to use the random optimization algorithm.

(4) Execute the Hyperopt function, obtaining a set of hyperparameters that optimize the objective function.

In Bayesian optimization, the model parameters are treated as random variables whose prior probability distributions represent the initial estimates of the parameters. In contrast, the posterior probability distributions represent the estimates of the parameters as they are continuously updated. Tables 2 and 3 show the Hyperopt hyperparameter search space settings and search results respectively.

Table 2
Hyper reference search space settings.

Parameters	Search space
Number of hidden layers	[3, 4, 5, 6, 7]
Number of hidden units	[64, 128, 256, 512]
dropout	[0, 0.05, 0.1, 0.3, 0.5]
Learning Rates	[1e-5, 1e-4, 1e-3, 1e-2, 1e-1]
Batch size	[64, 128, 256, 512]

4.2.2. Model training process

Tensor Board is used to monitor the performance details of the model. We record the curves of the loss function and the AUC of the model prediction as the model learning process progressed to ensure that the entire model training process is normal, as shown in Fig. 5. Where orange, red, and blue denote the training set, validation set, and test set in that order.

The three curves in Fig. 5 exhibit similar trends, gradually flattening as the number of training iterations increases. This indicates that the model's predictive accuracy on the training, validation, and test sets has stabilized with increasing training iterations, indicating convergence. Additionally, the relative positions of the three curves remain consistent, with the AUC highest on the training set and lowest on the test set, as expected.

4.2.3. Over-fitting problem improvement strategies

Two common regularization methods, dropout and batch normalization, are used to mitigate overfitting. The dropout forces the network not to depend on any input features by randomly setting the output of some neurons to zero during training, thus avoiding overfitting. Batch normalization is performed on each mini batch to make the network more stable, thus avoiding the problem of gradient disappearance or gradient explosion, and also alleviating overfitting. Regarding the prediction model in this paper, the specific methods are as follows:

(1) For the LR model on the Wide side of Wide & Deep, we introduce an L1 regularization term $c||\omega||1$ into its objective function to constrain complexity.

(2) In the Wide & Deep model, we incorporate dropout and batch normalization to optimize the MLP model on the Deep side.

4.3. Baseline model and assessment indicators

Considering the final need for model interpretability analysis of the prediction, it is desired to select some machine learning models with inherent interpretability and uninterpretable integrated models. As benchmark models, MLP, SVM, decision tree, logistic regression, and random forest are chosen.

Considering the relatively balanced proportion of positive and negative samples in the dataset, the accuracy rate and F1 value as one of the evaluation indexes are chosen. In addition, considering the cost of misclassifying positive samples and misclassifying negative samples may be different, the accuracy rate and recall rate as one of the evaluation indexes are chosen. The Area Under Curve (AUC) is a common index for evaluating the performance of binary classification models, which can take into account the performance of the model under different thresholds and is also useful for cases where the proportion of positive and negative samples is balanced.

4.4. Analysis of experimental results

4.4.1. Experimental environment setup

The general parameters and baseline hyperparameter are set as follows: In the MLP, the number of hidden units is 256, the number of hidden layers is 5, the dropout rate is 0.5, the learning rate is 0.01, and

Table 3
Hyperparameter search results.

Number of hidden layers	Number of hidden units	Dropout	Learning Rates	Batch size
4	256	0.3	1e-4	128

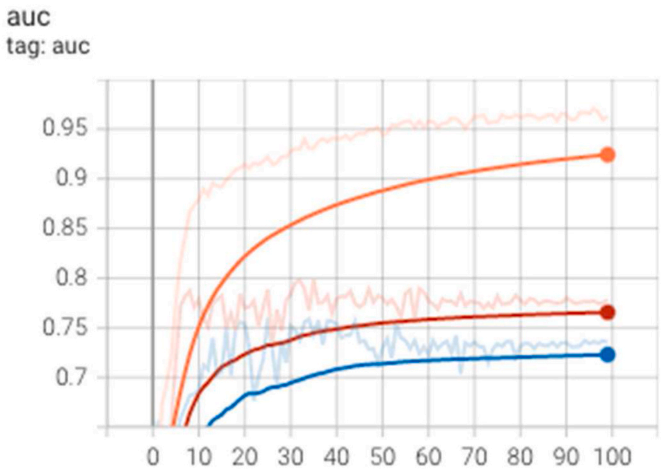


Fig. 5. Preliminary results of the predictive model.

the batch size is 128. In the SVM, the penalty parameter C is 1, the kernel function is RBF, and the maximum number of iterations is 100. In the Random Forest, the maximum depth of decision trees is 30, the minimum number of samples required to split an internal node is 2, the minimum number of samples required to be at a leaf node is 1, and the maximum number of leaf nodes is unlimited. In the Logistic regression, the regularization is L1, the maximum number of iterations is 100, and the tolerance for stopping criteria is 1e-4. In the Sample size and dataset splitting, the dataset contains 129880 samples, which are split into training, validation, and test sets in an 8:1:1 ratio using 10-fold cross-validation.

4.4.2. *Performance comparison experiment*

We compare the performance of five benchmark models (MLP, SVM, decision tree, logistic regression, random forest) with the proposed Wide & Deep model. We use accuracy, precision, recall, and f1 values as evaluation indexes. The dataset is randomly divided into a training set, a validation set, and a test set. Then the baseline model is trained and evaluated by ten-fold cross-validation. As for the Wide & Deep model, we trained on the training set, tuned on the validation set, and finally evaluated on the test set.

Table 4 shows that the prediction model Wide & Deep performs the best, with accuracy, precision, recall, f1 value, and AUC reaching 0.929, 0.945, 0.913, 0.929, and 0.935, respectively. The random forest performs the second best, with accuracy, precision, recall, f1 value, and AUC reaching 0.898, 0.921, and 0.871, respectively. In contrast, the decision tree performs worst. In addition, between the benchmark models, the MLP and logistic regression algorithms perform well in terms of accuracy and AUC but fall slightly short in other indexes. The SVM algorithm performs moderately well, and the decision tree algorithm performs worst in all indexes. In all, the prediction model Wide & Deep can be used as an effective prediction model for the analysis of airline passenger satisfaction.

Table 4
Performance comparison results.

Method	Accuracy	Precision	Recall	F1-Score	AUC
MLP	0.853 ± 0.012	0.874 ± 0.015	0.827 ± 0.018	0.849 ± 0.013	0.872 ± 0.007
SVM	0.826 ± 0.016	0.848 ± 0.012	0.794 ± 0.021	0.819 ± 0.017	0.863 ± 0.009
Decision tree	0.784 ± 0.022	0.811 ± 0.018	0.739 ± 0.026	0.776 ± 0.021	0.814 ± 0.012
Logistic regression	0.869 ± 0.011	0.891 ± 0.014	0.846 ± 0.017	0.869 ± 0.012	0.846 ± 0.006
Random forest	0.898 ± 0.008	0.921 ± 0.009	0.871 ± 0.015	0.896 ± 0.009	0.879 ± 0.003
Wide & Deep	0.929 ± 0.006	0.945 ± 0.007	0.913 ± 0.009	0.929 ± 0.006	0.935 ± 0.002

4.4.3. *Model ablation experiment*

To investigate the impact of the different components and parameters of the Wide & Deep model, we conducted a model ablation experiment. The AUC values for the Wide side only, Deep side only, without Embedding layer, Deep side MLP layer number 1 and layer number 2, and with random superparameter settings are compared in Table 5.

It shows that the AUC of the model containing only the Wide side is 90.96 ± 0.28 and the AUC of the model containing only the Deep side is 92.13 ± 0.21, which indicates that both the Wide & Deep sides contribute to the prediction effectiveness of the model. In the model without the Embedding layer, the AUC is 69.43 ± 0.78, which is relatively low, while the performance of the model containing only the Deep side and the model containing only the Wide side are both better than the model without the Embedding layer, indicating that the Embedding layer played an important role in the improvement of the model's prediction. With the number of MLP layers on the Deep side, the AUCs are 91.36 ± 0.27 and 91.72 ± 0.23, indicating that increasing the number of MLP layers could improve the prediction effect of the model, and the model with MLP layer 2 performs slightly better than the model with MLP layer 1.

Meanwhile, it is found that the performance of the model with the random superparameter setting is much lower than the performance of the other models. We also used experiments with random superparameter settings and obtained a relatively low AUC of 75.28 ± 1.39, indicating that proper superparameter settings are important for the improvement of the model prediction. The AUC of the final Wide & Deep model is 93.57 ± 0.12, which is much higher than the results of other ablation experiments, indicating that the proposed Wide & Deep model is very effective in predicting passenger satisfaction.

5. Model interpretability analysis

The DeepLIFT algorithm is applied to perform the interpretability analysis of the prediction results. We explore the importance of the features on which the model is based in predicting post-delay passenger satisfaction, understand the extent to which different features contribute to the prediction results, and thus identify the features that have a greater impact on the prediction.

5.1. *DeepLIFT-based feature importance*

The DeepLIFT algorithm is a widely used method for the explanatory analysis of deep learning model, which can interpret globally and relatively locally, either for the overall model behavior or for different classes of samples. It can obtain an analysis of the feature importance of the whole model over all samples by calculating the degree of contribution of each feature to the model output, as shown in Fig. 6.

The horizontal axis represents the indexes of the 22 features and the vertical axis represents the feature importance score. It shows that for passengers who continue to travel on delayed flights, Inflight Wi-Fi service is the most important feature in terms of ultimate passenger satisfaction, followed by Flight Distance and Type of Travel. The ratings

Table 5
Experimental results of model ablation experiments.

Operation	AUC
Wide side only	90.96 ± 0.28
Includes Deep side only	92.13 ± 0.21
Embedding layer not included	69.43 ± 0.78
Deep side MLP with 1 layer	91.36 ± 0.27
Deep side MLP with 2 layers	91.72 ± 0.23
Random hyperparameter setting	75.28 ± 1.39
Wide & Deep	**93.57 ± 0.12**

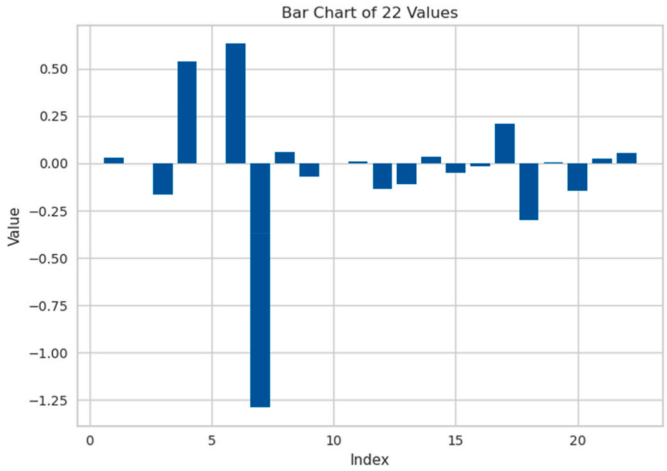


Fig. 6. Feature contribution score based on DeepLIFT algorithm.

of check-in service and baggage handling also have a significant impact on the prediction of the satisfaction model.

The characteristic Importance value of −1.3 for the feature Inflight Wi-Fi service in predicting passenger satisfaction means that for passengers on delayed flights, low quality of Wi-Fi service may have a negative impact on their satisfaction. It is because passengers need to spend more time on board, where the poor-quality in-flight Wi-Fi service is, which would result in dissatisfaction with the passenger experience and thus affect their satisfaction with the airline.

lFight distance, as another feature, has a significance value of 0.65 in predicting passenger satisfaction, which has a greater impact on the prediction. This may be due to the differences in service quality and comfort between long-haul and short-haul flights. Long-haul flights typically offer more services and amenities such as meals and entertainment, whereas short-haul flights may be more simplified, and therefore passengers may be more dissatisfied with delays on long-haul flights due to higher expectations and concerns about long-haul flights. In addition, whether a passenger is traveling for business or personal purposes has a significance value of 0.55 in prediction, meaning that the purpose of the passenger's flight contributes to the predicted outcome.

As we explore the impact of inflight service on satisfaction in more depth, the importance of other characteristics for inflight service items after excluding characteristics related to the established fact of flight delays, passenger characteristics, journey characteristics, and non-inflight service characteristics are analyzed, we see that with the exception of inflight Wi-Fi service, baggage handling is rated as a more important service feature in predicting satisfaction, followed by the cleanliness and seat comfort, indicating that improving in-flight services, in particular baggage handling, should be of concern in terms of improving passenger satisfaction.

6. Conclusion and future research directions

6.1. Conclusion

Flight delays not only cost airlines but also negatively impact passengers' travel plans and experiences. This is especially true for those passengers who insist on taking the delayed flight. Their satisfaction will be directly affected by the in-flight service of the delayed flight. By reviewing the research status at home and abroad, it is found that there are only a few passenger satisfaction prediction studies based on machine learning and deep learning technology, and such studies still have shortcomings in the interpretability of the model. Therefore, this study adopts the Wide & Deep model based on feature interaction and model interpretability analysis based on the DeepLIFT algorithm, which provides a new method and tool for passenger satisfaction prediction. The following are the main conclusions of this study:

In the binary classification prediction task of researching on-board service satisfaction after delay, this paper chooses the Wide & Deep model that can internally realize the interaction of explicit features and the implicit feature interaction of neural networks to predict the model The model is built, and the model is improved through hyperparameter tuning, overfitting mitigation, and learning rate adjustment. In the performance comparison experiment, this paper uses the public data set of airline passenger satisfaction to predict the empirical analysis, comprehensively compares the benchmark model with the satisfaction prediction model in this paper, and verifies that the Wide & Deep model based on feature interaction can improve the performance of passengers after delays. Satisfaction prediction has the best accuracy and predictive performance. Then, in the model ablation experiment, the influence of each component of the Wide & Deep model on the final performance was independently evaluated. The results proved that factors such as the Wide side, the Deep side, the Embedding layer, the number of MLP layers on the Deep side, and the settings of hyperparameters all affect the model. The prediction effect has a non-negligible contribution.

6.2. Future research directions

First, the publicly available dataset used in this paper is based on data from one airline, and the sample data may have certain limitations and specificity. We recognize that empirical analysis can increase confidence in the model results. In future work, a data set will be collected by actual survey or other means, and the model's characteristic importance results will be validated. It will further support the reliability and interpretability of our study. Future research may consider introducing data from more airlines, including different regions and types of airlines, to increase the sample diversity and generalization of the study. In addition, future research could consider further exploring passengers' psychological characteristics and behavioral preferences. For example, factors such as passengers' emotional state and travel experience may have a significant impact on satisfaction, and consideration could be given to incorporating these factors into the model to accurately predict passenger satisfaction. As for service quality, the focus of the future will be to compare the services provided by airlines after delays with those provided under normal circumstances, with the aim of conducting a more targeted assessment of service effectiveness.

Second, attempts can be made to incorporate other types of data into the model to improve its accuracy and usefulness. This study considered passengers' personal information and flight attributes, but airline passenger satisfaction may also be influenced by other external factors, such as weather and airfares. Future research can make use of external data sources such as weather data and airfare data and add these factors to the satisfaction prediction model through methods such as data fusion or feature engineering, in order to improve the prediction accuracy and practicality of the model. Additionally, it is interesting to investigate whether real-time updates of flight information during the flight can impact passenger satisfaction and the extent of its influence.

Third, future attempts to compare other machine learning models or deep learning models can be considered, and model fusion experiments can be attempted to combine the prediction results of multiple models, thereby improving the overall performance of the model.

Last, in this study, due to the length of the paper, we only used the DeepLIFT algorithm to analyze the interpretability of the airline passenger satisfaction prediction results. Although the DeepLIFT algorithm satisfies the interpretability requirements of this paper to a certain extent, in order to better understand the decision basis and prediction results of the model, we may also consider using other interpretable analysis methods, such as LIME and SHAP, to conduct more comprehensive and in-depth interpretative analysis of the model in the future.

Funding

The first two authors of this research were partially funded by [National Natural Science Foundation of China] grant number [71901218].

Authors' contributions

Conceptualization, C.S.; Methodology, C.S.; Software, X.M.; Formal Analysis, X.M.; Data Curation, X.M.; Writing—Original Draft Preparation, X.M. and C.S.; Writing—Review and Editing, C.A. and J.Z.; Supervision, J.Z.; Funding Acquisition, C.S.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of competing interest

The authors declare that they have no competing interests.

Acknowledgements

The first two authors of this research were partially funded by [National Natural Science Foundation of China] grant number [71901218].

References

Almuqren, L., Cristea, A.I., 2023. Predicting STC customers' satisfaction using Twitter. IEEE Trans. Comput. Social Syst. 10 (1), 204–210.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., et al., 2020. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

CAAC (Civil Aviation Administration of China), 2023. Monthly Consumer Complaint Report [EB/OL]. URL: http://www.caac.gov.cn/XXGK/XXGK/TJSJ/index_1217.html.

Cheng, H., Koc, L., Harmsen, J., et al., 2016. Wide & Deep Learning for Recommender Systems. In: DLRS 2016: Proceedings of the 1st Workshop on Deep Learning for Recommender, pp. 7–10.

de Lange, P., Melsom, B., Bakke Vennerød, C., et al., 2022. Explainable AI for credit assessment in banks. J. Risk Financ. Manag. 15 (12), 556.

Deng, W., Pan, J., Zhou, T., et al., 2020. DeepLight: deep lightweight feature interactions for accelerating CTR predictions in ad serving. In: CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management 922-930.

Du, Y., Zhang, L., 2020. The current situation and countermeasures of flight delay in China. China Sci. Technol. Inform. 21, 35–37 (In Chinese).

Giao, H.N.K., Vuong, B.N., 2021. The Impact of Service Quality on Passenger Loyalty and The Mediating Roles of Relationship Quality: A Study of Domestic Flights with Vietnamese Low-Cost Airlines. Transport. Res. Procedia 56, 88–95.

Konstantinov, A.V., Utkin, L.V., 2021. Interpretable machine learning with an ensemble of gradient boosting machines. Knowl. Base Syst. 222, 106993.

Li, S., Cui, Z., Pei, Y., 2022. A dual adaptive interaction click-through rate prediction based on attention logarithmic interaction network. Entropy 24 (12), 1831.

Liu, Y., Yin, M., Hansen, M., 2019. Economic costs of air cargo flight delays related to late package deliveries. Transport. Res. Part E 125, 388–401.

Molnar, C., Casalicchio, G., Bischl, B., 2020. Interpretable machine learning–a brief history, state-of-the-art and challenges. Commun. Computer Inform. Sci. 1323.

Park, J., Robertson, R., Wu, C., 2004. The effect of airline service quality on passengers' behavioral intentions: a Korean case study. J. Air Transport. Manag. 10 (6), 435–439.

Pei, W., Yang, J., Sun, Z., et al., 2017. Interacting attention-gated recurrent networks for recommendation. In: Conference on Information and Knowledge Management, pp. 1459–1468.

Rajapaksha, D., Bergmeir, C., Buntine, W., 2020. LoRMIkA: local rule-based model interpretability with k-optimal associations. Inf. Sci. 540, 221–241.

Ryu, Y.K., Park, J.W., Lee, S., 2019. The effect of the in-flight meal on the in-flight service satisfaction, airline image, price sensitivity, and Re-use intention. Aviation Manag. Soc. Korea 17, 45–61.

Tang, S.K., Cheng, F., Zhang, D.M., 2023. Method for User Behavior Prediction Based on Feature Interaction and Graph Neural Network. China, CN 112465226 B[P].

Tompson, J., Jain, A., LeCun, Y., Bregler, C., 2014. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. NIPS 1799-1807.

Walia, S., Sharma, D., Mathur, A., 2021. The impact of service quality on passenger satisfaction and loyalty in the Indian aviation industry. Int. J. Hospital. Tourism Syst. 14 (2), 137–143.

Wang, H., Wang, N., Yeung, D., 2015. Collaborative deep learning for recommender systems. In: Proc. KDD, pp. 1235–1244.

Wilson, C.M., Fridley, B.L., Conejo-Garcia, J.R., et al., 2021. Wide & Deep learning for automatic cell type identification. Comput. Struct. Biotechnol. J. 19, 1052–1062.

Yan, C., Chen, Y., Wan, Y., et al., 2020. Modeling low- and high-order feature interactions with FM and self-attention network. Appl. Intell. 51, 3189–3201.

Yang, F.C., Zheng, S., Li, J., 2022. Interpretable machine learning methods applied in disease risk prediction: a case study of sepsis mortality risk predication. J. Capital Med. Univ. 43 (4), 610–617.

Zou, W.J., Pang, T.J., 2022. A hybrid data clustering recommendation algorithm for users and rating information. J. Taiyuan Normal Univ. (Nat. Sci. Ed.) 21 (2), 30–35.