Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/49/20241161

Federated Learning-based neural networks for airline passenger satisfaction prediction

Xiyu Lang[1,4], Yuxiao Liu[2] and Yang Zhang[3]

[1]Department of Computer Science, Xi'an University of Architecture and Technology, Xi'an,710055, China
[2]Department of Artificial Intelligence, Yangzhou University, Yangzhou, 225009, China
[3]Department of Computer Science, Sichuan University, Sichuan, 610065, China

[4]xylang@xauat.edu.cn

Abstract. Preserving user data privacy is paramount for airlines. However, achieving highly accurate predictions of passenger satisfaction while safeguarding the privacy of individual company's data remains a considerable challenge. To achieve this objective, we utilized a privacy-preserving machine learning technique known as Federated Learning (FL). FL allows model training on decentralized devices while ensuring data security and privacy. The FL process comprises client training processes, communication rounds, and weight aggregation. We simulate FL principles using multiple processes to enable distributed learning. Clients preprocess data and train local models, ensuring data privacy. Communication rounds involve clients downloading the global model, local training, and transmitting updates to a central server. Weight aggregation methods like Federated Averaging merge these updates, preserving data privacy. Additionally, we leverage Artificial Neural Networks (ANNs) as foundational techniques. ANNs consist of input, hidden, and output layers, with weight adjustments based on real value differences to achieve accuracy. Our approach combines FL with ANNs to demonstrate FL's potential in privacy-preserving predictive analytics. We use the "Airline Passenger Satisfaction" dataset for modeling and evaluate the impact of neural network depth and submodel quantity on prediction accuracy. Our experimental results reveal that neither the depth of neural networks nor the number of submodels significantly affects prediction accuracy. FL emerges as a promising approach to balance data privacy and prediction accuracy effectively.

Keywords: Privacy-preserving, Federated Learning, Neural Networks.

1. Introduction

In the current era of data-driven decision-making, predictive analytics has emerged as a critical tool for enterprises and industries to forecast outcomes, comprehend trends, and optimize operations. One particularly promising domain is the aviation industry, notably in forecasting passenger satisfaction with flights [1]. The aviation industry is fundamentally built upon customer satisfaction. Anticipating passenger satisfaction levels in advance can yield several pivotal advantages, encompassing enhanced service planning, efficient resource allocation, and targeted mitigation of potential issues.

DOI: 10.54254/2755-2721/49/20241161

While prior research within the field has explored aviation and customer satisfaction prediction, significant gaps remain, especially in the realms of privacy awareness and cross-domain predictive models. Conventional methods often involve aggregating data from multiple sources, resulting in data privacy breaches and increased logistical complexity. The escalating emphasis on data protection and privacy regulations further exacerbates these challenges. The research conducted by Yang et al. [2] primarily centers around centralized models that aggregate data from various airlines, airports, and customer feedback platforms. These studies indeed provide valuable insights into factors influencing passenger satisfaction. Liu [3] uses integrated algorithms such as gradient boosting and random forest to make good predictions of airline passenger satisfaction. Xiong [4] used the support vector regression model method to apply it and achieved good results. However, their applicability is constrained by the contemporary data governance landscape and the imperative to safeguard sensitive customer information.

In contrast, federated learning retains data localization and heterogeneity by training statistical models on remote devices or in isolated data centers. As a result, this enables large-scale machine learning, distributed optimization, and privacy-preserving data analysis through training within potentially extensive networks [5]. Consequently, federated learning facilitates model training on dispersed data sources without necessitating centralized data aggregation, thereby ensuring data security and user privacy.

It is important to protect privacy. Some employ federated learning to protect user data privacy while maintaining the performance of the learning model. For instance, Li et al. [6] propose a federated learning model for scenarios with a mix of sensitive and regular users within smart grids. This model employs varied fuzzy processing mechanisms catering to distinct user privacy requirements, thus enhancing privacy protection levels. However, we posit that this approach is not entirely applicable to our chosen problem. On the other hand, Li [7] presents a decentralized vertical federated learning model grounded in the neural network algorithm, suitable for scenarios where data is distributed vertically based on features and isn't reliant on a central server. This model provides a conceptual framework for addressing our problem. Leveraging the artificial neural networks as the foundational technique and using Airline Passenger Satisfaction as dataset, we construct a federated learning model to predict flight satisfaction outcomes. Simultaneously, we will undertake training for a non-federated learning model and compare its predictive accuracy with that of the federated learning model. This comparative analysis aims to validate the ability of the federated learning approach to maintain predictive accuracy while ensuring data privacy and user information security.

2. Method

2.1. Dataset preparation

The "Airline Passenger Satisfaction" dataset collected from Kaggle [8] comprises 129,880 samples and encompasses 24 features, making it a valuable resource for predictive modeling. These features encapsulate various aspects of passengers' personal information, travel details, and service ratings, collectively representing a comprehensive set of influencing factors.

Before analyzing this dataset, we performed a series of data preprocessing steps to ensure data quality and suitability. Firstly, we considered the possibility of missing values in the dataset. To address this issue, we conducted data cleaning. Additionally, for categorical features such as gender, customer type, travel type, and cabin class, we employed methods like one-hot encoding or label encoding to transform these categorical features into numeric formats, enabling the model to understand and process these features effectively. For the target variable "satisfaction," we assumed the need for label encoding, transforming it into binary labels. In this encoding, we labeled "satisfied" as 1 and "dissatisfied" as 0. In the realm of data analytics, overlooking this responsibility can result in imprecise insights and questionable decision-making, making the identification and rectification of erroneous data an ongoing concern[9].These data preprocessing steps were undertaken to ensure data consistency and usability, providing a reliable data foundation for our subsequent modeling and analysis. This also helps reduce

DOI: 10.54254/2755-2721/49/20241161

uncertainty during model training and testing, ultimately enhancing the accuracy of our predictions regarding passenger satisfaction levels [10].

2.2. Federated learning-based ANN network

2.2.1. Federated learning

Federated learning is a pioneering approach to machine learning that safeguards data privacy while enabling collaborative model training across decentralized devices or servers. In the context of our dataset, 'Airline Passenger Satisfaction,' which involves sensitive passenger information, data security is paramount. The main idea behind federated learning is to address the challenges posed by traditional centralized machine learning methods. This is achieved by allowing models to be trained locally on individual data sources, thereby reducing the need for raw data to leave its source, which is crucial when handling privacy-sensitive information.

The framework of federated learning is centered on three key components that seamlessly ensure data privacy and collaborative model development: Firstly, the process begins with local training, where individual devices or nodes, such as smartphones or edge devices, conduct model training using their data while safeguarding data privacy. This locally trained model is termed a 'local update.' Subsequently, model aggregation securely combines these local updates on a central server to create a global model update. Various aggregation techniques, such as Federated Averaging, are employed to merge these local updates while preserving privacy. Finally, the global model update is sent back to the devices, facilitating knowledge sharing while maintaining data decentralization. This iterative process continues until the model converges to an optimal state.

2.2.2. Neural network

Artificial neural networks consist of input layers, hidden layers, and output layers [11, 12]. The input layer exclusively receives information from the external environment, and it directly transfers input data to the hidden layer. In their study, Cao et al. [13] proposed that the hidden layer resides between the input layer and the output layer. Each hidden layer takes the input data from the preceding layer as input variables for a series of weighted computations, propagating the output result as input data to the subsequent layer. Ultimately, the output layer generates the final result, with each output unit corresponding to a specific classification. The network's weight is adjusted based on the deviation between the actual result value and the target value. This iterative process yields an artificial neural network model that meets accuracy requirements.

Our neural network architecture consists of four fully connected layers (fc1, fc2, fc3, and fc4), employing the Rectified Linear Unit (ReLU) activation function to introduce non-linearity. This architecture is specifically designed for classification tasks, aiming to categorize input data into one of two classes. The initial layer, fc1, acts as the input layer, receiving data with 20 features and transforming it into a higher-dimensional representation with 32 neurons. This transformation enables our network to capture intricate patterns and relationships within the data. Subsequently, fc2 serves as the first hidden layer, reducing the dimensionality from 32 to 16 neurons. This reduction aids in abstracting and simplifying the features learned by the preceding layer. Furthermore, fc3 further diminishes the dimensionality from 16 to 8 neurons, facilitating a hierarchical feature extraction process. Lastly, fc4 operates as the output layer with two neurons, which is well-suited for binary classification tasks. It furnishes our network's prediction for the input data, with each neuron representing one of the two possible classes. Throughout our architecture, introducing non-linearity, the ReLU activation function enhances the model's ability to capture complex data patterns and make accurate classifications. The architecture of our neural network is illustrated in Figure 1 below.

DOI: 10.54254/2755-2721/49/20241161

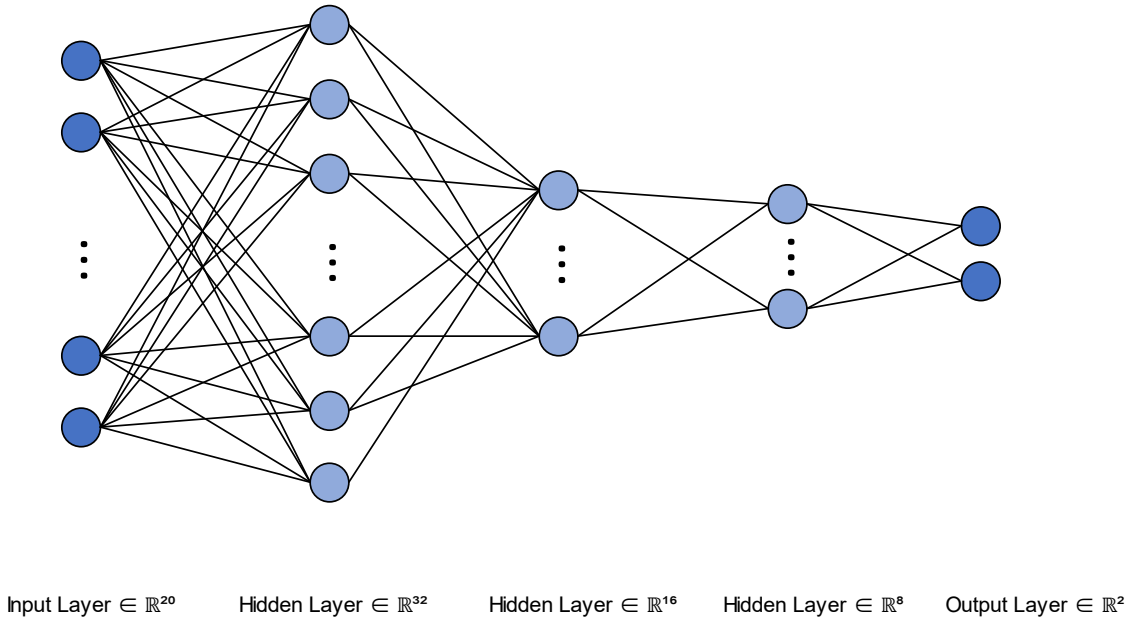

Figure 1. The structure of the proposed neural network (Photo/Picture credit : Original).

2.2.3. Federated Learning Integration with ANN

In this section, we delve into the fusion of Federated Learning (FL) with Artificial Neural Networks (ANN). Specifically, we focus on simulating FL principles using multiple processes, client training processes, and communication rounds with weight aggregation.

2.2.3.1. Simulating FL Principles Using Multiple Processes

The principles of Federated Learning can be simulated using multiple processes to achieve distributed learning. The process is as follows: 1) Global Model Initialization: A central server initializes a global model, typically a global neural network model. 2) Client Model Download: Clients, as processes, download the global model, which will be used for local training. 3) Local Training: Clients perform training with their local data, including data preprocessing, model training, and generating local model parameters. 4) Model Update Transmission: Clients transmit updates of their local model parameters back to the central server. This process iterates, simulating the iterative learning process of FL, where clients collectively improve the global model through local training.

2.2.3.2. Client Training Processes

When clients locally train their models, they undergo the following steps: 1) Data Preprocessing: Data preprocessing involves tasks such as filling missing values, assuming a zero-fill strategy, ensuring data consistency with global model requirements, and handling non-numeric features. 2) Model Update Transmission: After local training, clients calculate gradients and send them to the central server for model updates.

2.2.3.3. Communication Rounds and Weight Aggregation

FL operations occur over multiple communication rounds. In each round, clients download the current global model, perform local training, and transmit model updates to the central server. Weight aggregation is a crucial step in this process.

Weight aggregation methods, such as Federated Averaging, are employed by the central server to merge received updates. The aggregation is typically performed using mathematical formulas to ensure

DOI: 10.54254/2755-2721/49/20241161

that the central model reflects collective knowledge from all clients while preserving individual data privacy.

When it comes to weight aggregation, we can represent it using the following formulas, where θ represents model parameters (including weights and biases):

$$\Theta_{\text{global}}[\text{fc}_\text{l}] = \frac{1}{\text{N}}\sum_{\text{i}=1}^{\text{N}} \theta_{\text{local}_\text{i}}[\text{fc}_\text{l}] \quad (1)$$

In this formula, $\theta_{global}[fc_l]$represents the global model parameters for layer l, N is the number of clients, and $\theta_{local_i}[fc_l]$ represents the local model parameters of the i-th client for layer l. This formula simplifies the weight aggregation process and applies to each layer of the model, ensuring that the global model captures collective knowledge while preserving data privacy for individual clients.

2.3. Implementation details

In the experimental process, we adjusted parameters, and the improvement in model accuracy is illustrated in the Table 1, Table 2 and Table 3.

Table 1. The parameter setting of the model-1.

parameters	value
learning_rate	0.01
batch_size	2
num_communications	30
num_models	2
num_local_steps	1

Table 2. The parameter setting of the model-2.

parameters	value
learning_rate	0.1
batch_size	2
num_communications	50
num_models	2
num_local_steps	1

Table 3. The parameter setting of the model-3.

parameters	value
learning_rate	0.1
batch_size	2
num_communications	30
num_models	4
num_local_steps	2

Furthermore, optimization was performed for each model and at every communication round to facilitate the gradual convergence of the models to improved states.

The following table presents the development environment versions and hardware specifications employed in our experimental setup.

3. Result and discussion

3.1. The performance of various models

First, we change the number of neural network layers. Using a single layer of neural network with 21 input neurons and 2 output neurons resulted in an accuracy rate of 53%. Employing two layers of artificial neural networks with 21 input neurons, 21 hidden layer neurons, and 2 output neurons also yielded a correct rate of 53%. Similarly, employing three layers of artificial neural networks with 21

DOI: 10.54254/2755-2721/49/20241161

input neurons, 21 neurons in each of the two hidden layers, and 2 output neurons led to a correct rate of 53%. Likewise, utilizing four layers of artificial neural networks with 21 input neurons, 21 neurons in each of the three hidden layers, and 2 output neurons produced an accuracy rate of 53%. The experimental findings indicate that altering the number of neural network layers does not significantly impact the prediction performance for this particular problem.

Then we change the number of submodel from 1 to 5. The results show that the prediction accuracy is the highest when there is only one submodel, and when the number of submodels is higher than 1, the prediction accuracy of the model will be lower than that when the number of submodels is 1. As the number of submodels increases, the accuracy of model predictions fluctuates. The results mentioned above are illustrated in Figure 2.

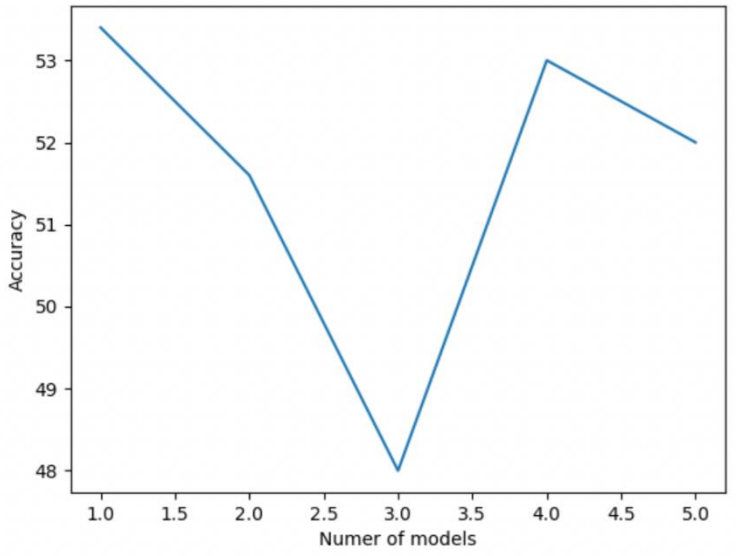

Figure 2. The accuracy of prediction performance based on different number of models (Photo/Picture credit: Original).

3.2. Discussion

The experimental parameter settings and results demonstrate that the number of neural network layers has minimal impact on prediction accuracy. Conversely, the quantity of distinct submodels does affect model accuracy. Additionally, our experiments indicate that federated learning safeguards data privacy without substantial compromise to model accuracy. The model accuracy is highest when there is only one submodel. Since increasing the number of submodels will cause inconsistency in the training data of each model, which may lead to large differences in the parameters of the trained model, the accuracy of the central model composed of multiple submodels will be slightly lower than the accuracy of the number of submodels of 1.

4. Conclusion

In this work, we proposed federated satisfaction prediction to train satisfaction prediction models using "Airline Passenger Satisfaction" dataset via federated learning to protect privacy while maintaining the prediction accuracy of the model. We combine artificial neural networks and federated learning methods and apply them to air passenger satisfaction prediction. We conducted extensive experiments to assess the proposed approach, and the results indicated that the combination of artificial neural networks and federated learning can effectively safeguard data privacy without causing a notable decrease in model accuracy. In the future, we plan to improve the accuracy of our method and apply the method to other areas.

DOI: 10.54254/2755-2721/49/20241161

Authors Contribution

All the authors contributed equally, and their names were listed in alphabetical order.

References

[1] Tsafarakis S and Pantouvakis A 2018 A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement Journal of Air Transport Management 68 61-75 ISSN 0969-6997.

[2] Yang H T and Liu X 2018 Predictive simulation of airline passenger volume based on three models Data Science 902 ISBN 978-981-13-2205-1.

[3] Liu Y 2022 Prediction of Airline Passenger Satisfaction Based on Machine Learning Pioneering with Science & Technology Monthly 35(4) 142-145.

[4] Xiong H 2021 Travel based on machine learning: Research on key technical of air passenger (Doctoral dissertation) Shanghai University of Technology.

[5] Li T Sahu A K Talwalkar A and Smith V 2020 Federated Learning: Challenges, Methods, and Future Directions IEEE Signal Processing Magazine 37(3) 50-60.

[6] Li Y and Yu H 2023 A privacy protection method for electric power customers based on a federated learning model Information Technology 379(06) 184-188.

[7] Li Z 2023 Decentralized longitudinal federated learning based on random forest (Doctoral dissertation) Jilin University.

[8] Klein T J 2019 Airline Passenger Satisfaction: What factors lead to customer satisfaction for an airline? [Online dataset] Kaggle URL: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.

[9] Xu C Ilyas I F Krishnan S and Wang J 2016 Data Cleaning: Overview and Emerging Challenges In Proceedings of the 2016 International Conference on Management of Data (SIGMOD ‘16) p 2201-2206 Association for Computing Machinery.

[10] Zheng X Wang M and Ordieres-Meré J 2018 Comparison of Data Preprocessing Approaches for Applying Deep Learning to Human Activity Recognition in the Context of Industry 4.0 Sensors 18 2146.

[11] Li X Wang J and Yang C 2023 Risk prediction in financial management of listed companies based on optimized BP neural network under digital economy Neural Computing and Applications 35(3) 2045-2058.

[12] Qiu Y Wang J Jin Z Chen H Zhang M and Guo L 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72 103323.

[13] Cao L Zhu W Wu J and Zhang C 2021 Inverse Design of Phononic Crystals by Artificial Neural Networks Chinese Journal of Theoretical and Applied Machines 53(7) 1992-1998.