Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241092

Federated learning-based neural network for hotel cancellation prediction

Yifei Hang

The Department of Applied & Computational Mathematical Sciences, University of Washington, Seattle, WA, 98105, the United States

yhang2@uw.edu

Abstract. Hotel reservations have become a prevalent choice for customers. However, cancellations of these reservations present a significant challenge for hotels, potentially resulting in financial losses and a decline in customer satisfaction. To address the issue of improper management of cancellations and minimize losses, machine learning can be employed to analyze and predict cancellations based on customer information. In cooperative scenarios where hotels collaborate to train a unified model, traditional algorithms that aggregate all data raise concerns about the protection of sensitive customer information. In this context, federated learning emerges as an effective solution to ensure the privacy protection of customers while achieving the desired predictive outcomes. Thus, to protect customer privacy while preserving performance of the federated learning model in comparison to the non-federated version, this work proposed to implement both federated learning and non-federated algorithms based on neural network to make predictions of cancellation based on multiple factors. The federated learning approach achieved a final testing accuracy of 76.64%. Although this accuracy was about 9% lower than the non-federated case, its loss was over half the loss of non-federated, and its testing accuracy was similar to the training accuracy, while the non-federated algorithm's testing accuracy was approximately 4% lower than the training one. Such results indicate that although accuracy was relatively lower, the federated learning approach prevented the overfitting problem in the non-federated case, while the data privacy problem was resolved.

Keywords: Federated Learning, Customer Prediction, Hotel Reservation, Cancellation Analysis.

1. Introduction

In the rapidly evolving landscape of data collection and communication technology, coupled with the demand for accessible event organization methods in the fast-paced society, online reservation of hotels becomes a ubiquitous choice. Park et al. explored factors that influence customers' willingness to use online booking websites, and ease to use appears to be the most influential factor, indicating the convenience of online reservation brought to customers [1]. Additionally, the hotels that received the reservations can also adequately prepare their service to meet the customers' expectations.

However, due to external factors such as weather conditions, politics, or economics, reservations are occasionally followed by cancellations, which causes significant losses, up to approximately 20 percent [2, 3]. Yet even when the hotels try avoiding vacant rooms with overbooking, occasions when there are fewer cancellations than expected brings long-lasting negative effects with downgrading service and trivial positive effect with upgrading [4].

In order to mitigate these financial setbacks, the analysis and prediction of customer behavior assume paramount importance. Machine learning, a widely employed technique across industries, serves as a powerful tool for conducting such assessments. Take the telecommunication realm, which is a core and highly competitive industry in developed countries, as an example, machine learning can be used to predict customer churn, so that companies can provide personalized services to retain customers who are likely to leave [5]. Also, in the financial sector, credit risk assessment, or the evaluation of whether a customer will fail to meet their financial obligations, had been essential yet demanding in recent academic and business organizations. Such assessment can be improved in accuracy and efficiency using various machine learning algorithms [6, 7]. Moreover, handling customer complaints plays a fundamental role in business. Machine learning algorithms can help analyze these complaints to prioritize crucial problems, ameliorate complaint management strategies, and therefore enhance customer satisfaction and business profitability [8]. Similarly, customer prediction in hotel reservation can also provide those who are likely to cancel with individualized discounts or additional services. Moreover, with a relatively accurate prediction on number of cancellations, the hotels can plan overbooking accordingly and restrain any unfavorable consequence.

While the accuracy of prediction is pivotal, another essential aspect should be also considered, which is customer data privacy. Information used for hotel reservation are highly personal, and when several hotels wish to collaborate to train a customer prediction model, they would not want to share users' sensitive information with the other hotels. Therefore, traditional machine learning algorithms that aggregate all data are not feasible. To overcome this privacy challenge, federated learning can be applied, which enables privacy protection through local model training and parameter exchange, so that information can be shared amongst multiple firms with security and equitability [9].

To produce customer prediction with promising accuracy and data security, this paper thus aims to employ federated learning algorithms for investigation. The dataset used for analysis is acquired from Kaggle, and Federated Averaging model is built based on neural network. Moreover, the non-federated learning version of neural network is trained to compare the accuracy differences and ensure the preservation of accuracy using federated learning.

2. Method

2.1. Data Preparation

The dataset for this research is acquired from a hotel reservation dataset on Kaggle [10]. The dataset contains 36,275 rows of data, each representing a customer's reservation detail. It has 18 features, including the composition of the guests, the length and date of stay, required service from customer, etc., and 1 label, booking_status, categorized into Not_Canceled that consists 67% of the data and Canceled that consists 33%.

The preprocessing implemented in this study was consisted of the following steps. First, the label "Booking_ID" was removed from the dataset, as it was unique in each row and was thus a trivial factor to the final prediction. Subsequently, whether the dataset contains null values was examined to ensure that there were no missing values that would hamper the functionality of machine learning algorithms. Fortunately, this dataset contained no null values, which was ideal as a machine learning dataset. Following that, values in categorical features were converted into numerical ones through assigning each category a different integer value and mapping them accordingly into the data columns. For instance, the feature "type_of_meal_plan" contained 4 categories, "Not Selected" and meal plans 1 through 3. After conversion, "Not Selected" was mapped as 0, and the meal plans were 1 through 3 accordingly. Finally, the dataset was split randomly into the training dataset, which was 80% of the original dataset, and the testing dataset that consists 20% of the original dataset.

DOI: 10.54254/2755-2721/45/20241092

2.2. Proposed approach - Federated learning-based vs non-federated neural network

This research used the base method of neural network and applied federated learning shown in Figure 1 to simulate the collaborative learning among hotels while ensuring data privacy. Federated learning is realized through randomly separating the train dataset into n subsets of equal sizes, where n is the number of clients, or number of different hotels if put into context. The data subsets are then trained separately with the same initial machine learning model for k local updates, and after training, the updated parameters of the trained local models are aggregated without collecting the data. The above steps are then repeated for a few more iterations to improve the global model. In this research, Federated Averaging (FedAvg) was used, which took the average of all local models to produce the global model.

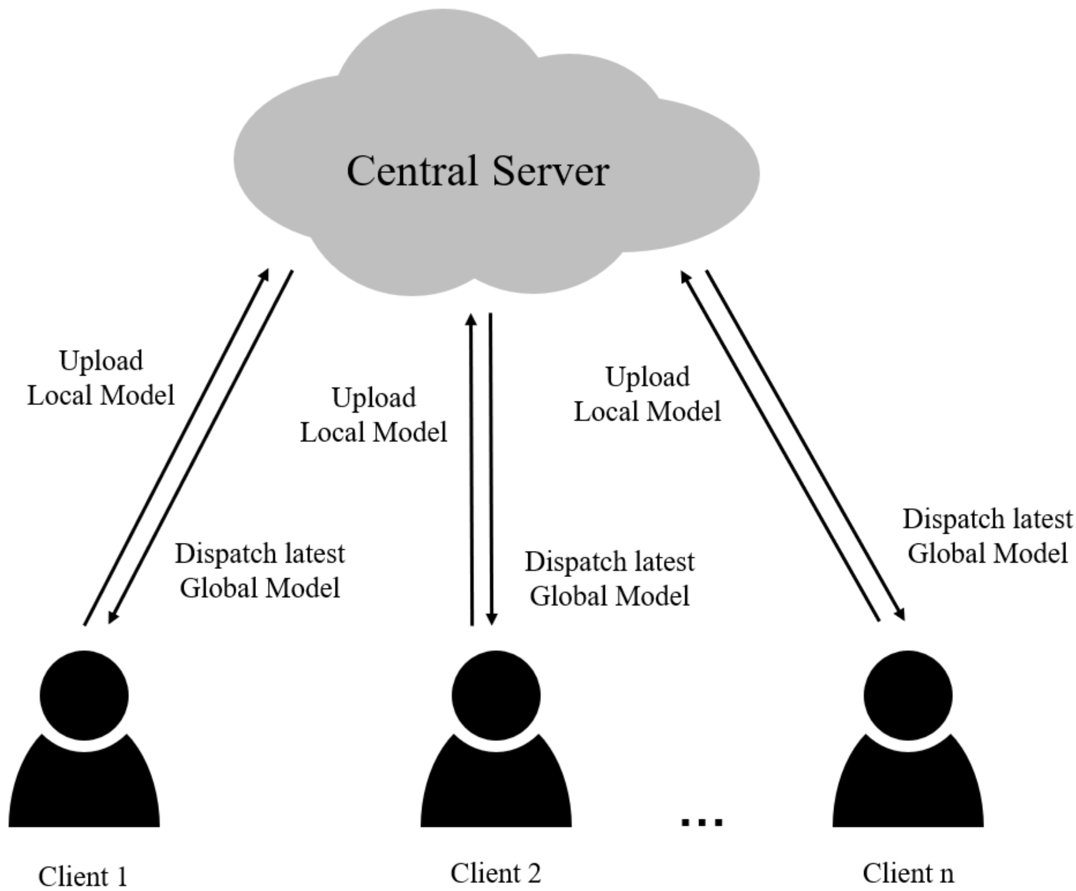

Figure 1. Federated Learning Model Structure (Photo/Picture credit: Original).

To apply neural network for each local model update, network consisted of 5 layers was built, with an input layer that takes in the 17 feature values of one data sample, 3 hidden layers that process the values and extract the data patterns, and an output layer that yields the predicted label value. The three hidden layers used rectified linear activation function (ReLU), while the output layer used Sigmond to produce a result ranging from 0 to 1 to ensure binary results. A detailed structure of the network is shown in Figure 2. The output was compared with the actual label and loss was calculated using the assigned loss function. Then, the gradient was calculated from loss, and the parameters were updated using optimizers.

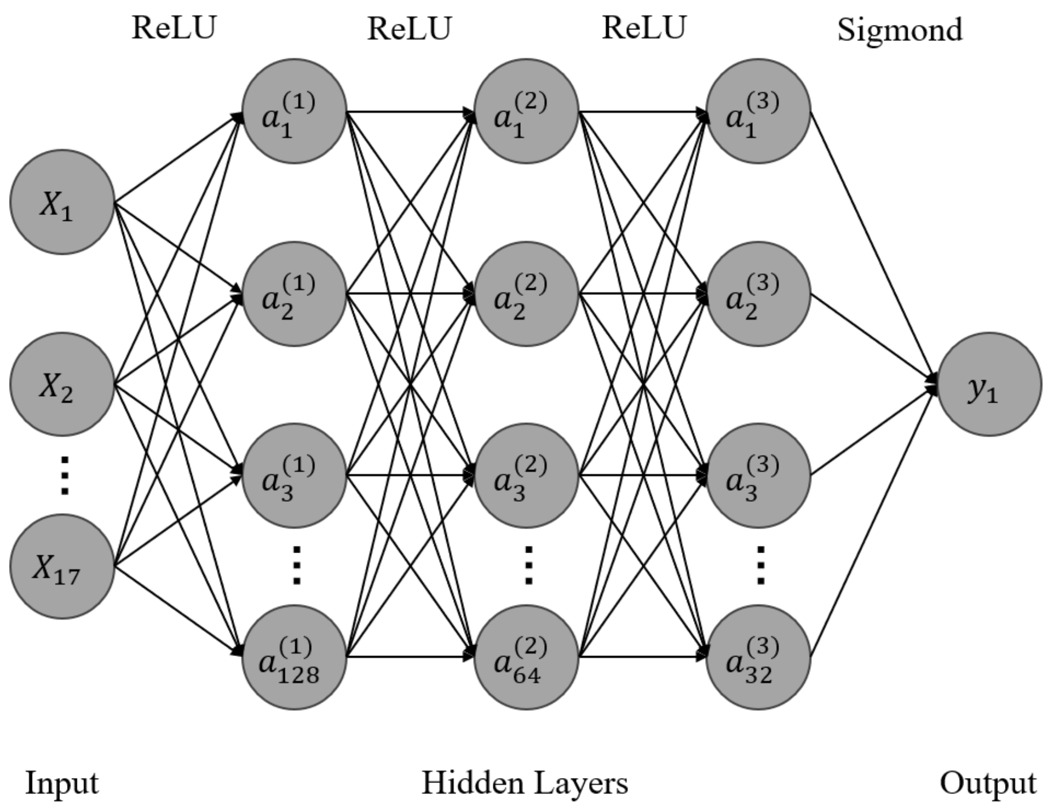

Figure 2. Neural Network Model Structure (Photo/Picture credit: Original).

In addition to the federated learning-based version, a non-federated neural network was also implemented using the same neural network, loss function, and optimizer. The functionality of this approach is to compare the loss and accuracy with the federated learning approach to ensure that accuracy was not severely compromised as introduced data protection.

2.3. Implementation details

PyTorch was employed to build and work with neural network. The loss function, Binary Cross Entropy loss function, was used for calculating the loss in binary classification, while the Stochastic Gradient Descent (SGD) optimizer was utilized for parameter updates. The federated learning process was enabled using the multiprocessing pool, which execute functions in parallel across multiple processes, and was thus ideal for simulating the local model updating process in federated learning.

This research used a batch size of 32, 10 communicating iterations, 4 clients, and 4 local updates. The learning rate was tuned by firstly setting the rate as powers of 0.1, and the optimum accuracy was achieved between 1×10^{-5} and 1×10^{-6}. After that, more precise tuning of the learning rate is applied by gradually decrease the learning rate from 1×10^{-5} by 1×10^{-6}, and the learning rate that produced the highest accuracy was 7×10^{-6}.

For the non-federated approach, batch size was the same as the federated version, and the number of iterations was 40, calculated through multiplying the number of communicating iteration and local updates of federated learning approach. The learning rate differed due to the different nature of the algorithms, and the learning rate chosen for this approach was 0.1.

DOI: 10.54254/2755-2721/45/20241092

3. Results and discussion

From the data shown in Table 1, both federated and non-federated approach produced a relatively high accuracy, with the non-federated approach approximately 9% greater than the federated approach. However, the loss of the non-federated approach was over twice the loss of the federated. The comparatively lower accuracy of the federated learning approach while having less loss may be because of the possibly biased or noisy data distribution when separating the train dataset into subsets. Although the datasets were separated randomly, such randomness does not promise the even distribution of data or that the data subset for each client is equally noisy. Another possibility would be that a client's local update led its result to the local minimum. If a client with such noisy or inaccurate data contributes significantly to the global model, it can negatively impact the global model's performance, and therefore produce a less accurate outcome.

Table 1. Performance of federated and non-federated neural network.

	Training Accuracy	Testing Loss	Testing Accuracy
Federated Learning Neural Network	76.57%	117.5226	76.64%
Non-Federated Learning Network	89.75%	276.3813	85.71%

While the federated approach may exhibit slightly lower accuracy due to algorithmic constraints, it offers a compelling blend of effectiveness and privacy preservation. With federated learning, as the data subsets are trained locally, minimal data transmission occurs during application—only model parameters are aggregated without revealing any local data. Furthermore, parallel training enhances time efficiency; when clients work concurrently with smaller data portions, training time can be significantly reduced, especially if the computational capabilities of clients are relatively comparable. Additionally, by keeping each client's data local and preventing the global model from accessing individual data, federated learning ensures the preservation of data privacy within the learning framework. Moreover, it is worth noticing that for the non-federated neural network approach, the testing accuracy was about 4% lower than the training accuracy, which is an indication of overfitting. In the federated learning case, on the other hand, the testing accuracy is about the same as, or even slightly better than, the training accuracy. Through implementing federated learning, the model successfully prevented overfitting that occurred in the non-federated case. Therefore, the federated learning approach not only ensured efficiency and privacy but also avoided overfitting, and it is worthy to compromise accuracy for these benefits.

4. Conclusion

In this work, the use of federated learning machine learning algorithm was proposed to mitigate the hotel reservation cancellation problems while preserving customer data privacy. Through developing FedAvg and non-federated neural network models and tuning the hyperparameters, hotel cancellation models were built to analyze cancellations based on various factors and predictions were made, and the performance of the two algorithms were compared. The results demonstrated that although the federated learning approach compromised slightly in prediction accuracy, its promising performance in efficiency and privacy protection as well as its avoidance of overfitting made such compromise worthy. Future research could try other data preprocessing methods, use other sophisticated federated learning algorithms, or add more data privacy protection algorithms to enhance the performance and privacy protectability of the federated learning model.

References

[1] Park Y A and Gretzel U and Sirakaya-Turk E 2007 Measuring Web Site Quality for Online Travel Agencies (Journal of Travel & Tourism Marketing vol 23) pp 15–30

[2] Xiao Y and Tian X and Xiao M 2020 Tourism Traffic Demand Prediction Using Google Trends Based on EEMD-DBN (Engineering vol 12) pp 194–215

[3] Sierag D D and Koole G M and van der Mei R D and van der Rest J I and Zwart B 2015 Revenue management under customer choice behaviour with cancellations and overbooking (European Journal of Operational Research vol 246) pp 170–185

[4] Wangenheim F and Bayón T 2007 *Behavioral Consequences of Overbooking Service Capacity* (*Journal of Marketing* vol 71) pp 36–47

[5] Ahmad A K and Jafar A and Aljoumaa K 2019 *Customer churn prediction in telecom using machine learning in big data platform* (*Journal of Big Data* vol 6) pp 1–24

[6] Leo M and Sharma S and Maddulety K 2019 *Machine Learning in Banking Risk Management: A Literature Review* (*Risks* vol 7) p 29

[7] Alarifi G and Rahman M F and Hossain M S 2023 *Prediction and Analysis of Customer Complaints Using Machine Learning Techniques* (*International Journal of e-Business Research* vol 19) pp 1–25

[8] Yang Q and Liu Y and Chen T and Tong Y 2019 *Federated Machine Learning: Concept and Applications* (*ACM Transactions on Intelligent Systems and Technology* vol 10) pp 1–19

[9] Yu Q et al 2020 Clustering Analysis for Silent Telecom Customers Based on K-means++ *In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* Vol 1 pp 1023-1027 IEEE

[10] Kaggle 2023 Hotel reservations classification dataset https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset