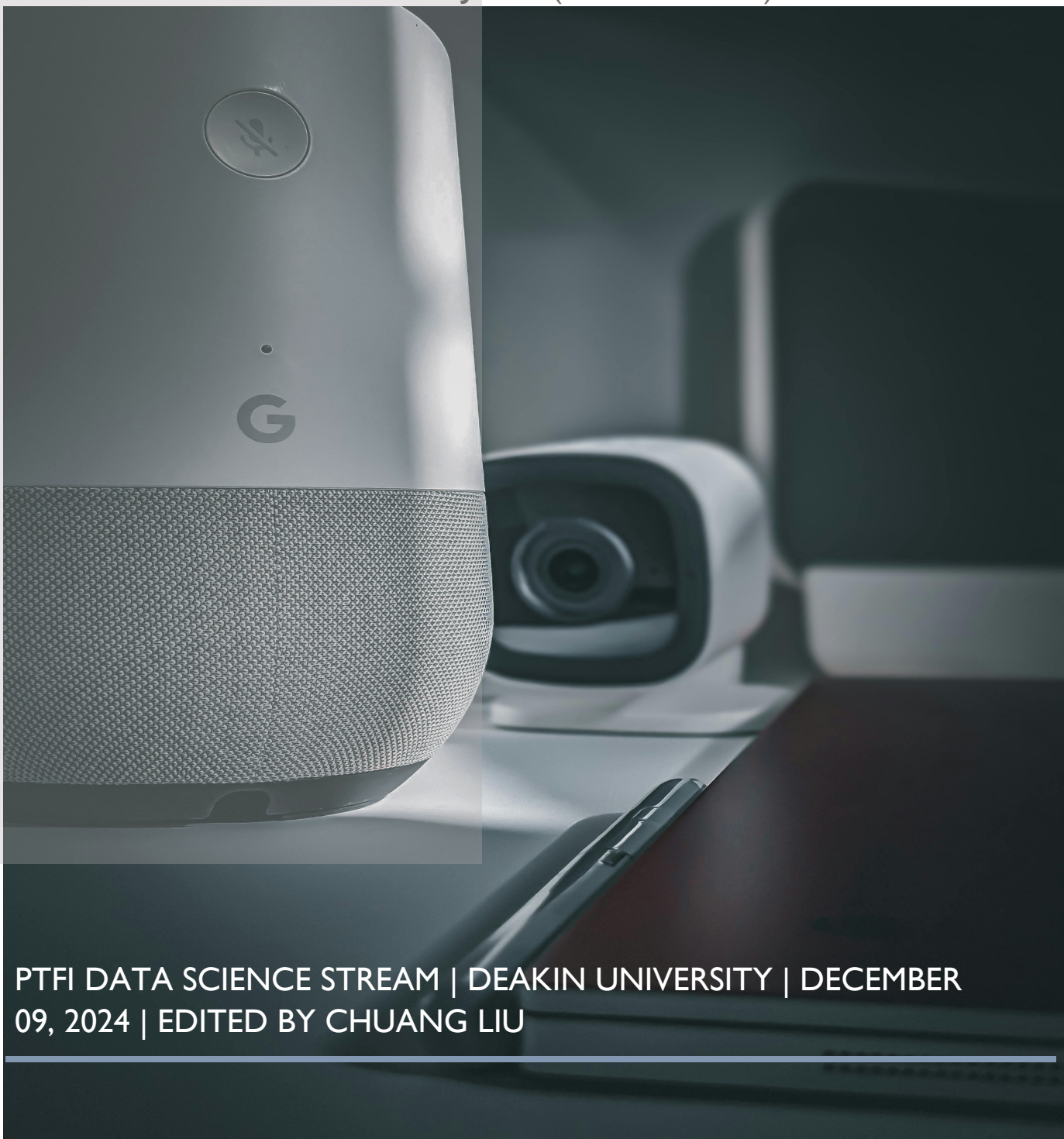
INTEGRATED FEDERATED LEARNING

PRIVACY TECHNOLOGIES FOR FINANCIAL INTELLIGENCE: A DATA BYTES COMPANY PROJECT (1ST EDITION)



PTFI DATA SCIENCE STREAM | DEAKIN UNIVERSITY | DECEMBER 09, 2024 | EDITED BY CHUANG LIU

1. BACKGROUND

In the world of finance, data is gold, but it is also a double-edged sword. There is tremendous value in this data, which can help financial institutions better understand market trends, predict risks, and even provide personalized services to customers. However, these data also contain a lot of sensitive information, such as customers' financial status, transaction records, etc., once leaked, the consequences are unimaginable. This leads to a conundrum: How to make the most of this data to improve service quality and efficiency without compromising customer privacy?

Federated learning, as an emerging distributed machine learning approach, offers an elegant solution to this conundrum. It allows multiple parties to train a model together without having to pool their data together. This way, each agency can train the model on its own data and then share only updates to the model, not the raw data. This approach not only protects the privacy of the data, but also ensures the security of the data.

Imagine that several banks want to work together to develop a new model for predicting market risk. If they follow the traditional practice, they need to pool their respective customer data on a central server, which not only involves complex data transfer and storage issues, but also may violate data protection regulations. However, with joint learning, they can train the model independently in their respective data centers and then only exchange updates of the model parameters. In this way, each bank's data remains safely on its own servers without the risk of a data breach.

In addition, joint learning can help financial institutions improve the robustness of their models. Because the data of each participant may come from different customer groups, with different characteristics and distribution. Through joint learning, models can be trained on these diverse data to better adapt to various situations and improve the accuracy of predictions.

However, the existing federated learning framework has some problems, such as the training time is too long, and it is not suitable for distributing different data. In this paper, we solve these two problems by integrating meta-learning[1], personalization layers[2] and FedAvg[3] into the federated learning framework.

At the same time, we analyze a new financial relationship (Bank Loans). Primarily, financial analysis is performed using the Loan dataset[4] generated by CTAB-GAN+[5].

The report covers as follows:

(i) Some background of Loan datasets;

(ii) Some background of Meta-Learning;

(iii) Some background of Personalization Layers;

(iv) Some background of FedAvg.

(v) Result

It is noted that there are many other methods to strengthen the federated learning framework, however, due to time constraints only the three methods selected were investigated.

DATASET: LOAN

In 2024, Jacob proposed Loan dataset, which title is Bank_Loan_modelling. This dataset is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns to better target marketing to increase the success ratio with a minimal budget. Its number of instances is 5000 and its number of attributes is 13 (including target attribute).

What is Loan's goal?

This dataset is used to research bank loan in Thera Bank. By analyzing the mathematical properties of 12 features in relation to the target, we can determine whether a Thera Bank's customer can handle a bank loan.

What is target in Loan?

Personal Loan：0, 1, Did this customer accept the personal loan offered in the last campaign?.

What are features in Loan?

Age: continuous, customer's age in completed years.

Experience: continuous, years of professional experience.

Income: continuous, annual income of the customer.

ZIPCode: continuous, Home Address ZIP code.

Family: 1, 2, 3, 4, Family size of the customer.

CCAvg: continuous, Avg. spending on credit cards per month, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

Education: 1, 2, 3, Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional.

Mortgage: continuous, value of house mortgage if any.

CreditCard: 0, 1, Does the customer use a credit card issued by UniversalBank?.

Securities Account: 0, 1, Does the customer have a securities account with the bank?

CD Account: 0, 1, Does the customer have a certificate of deposit (CD) account with the bank?.

Online: 0, 1, Does the customer use internet banking facilities?.

METHOD 1: META-LEARNING

WHAT IS META-LEARNING?

Meta-learning, also known as "learning to learn," is a machine learning approach that enables models to use previous learning experiences to accelerate the learning process for new tasks. The core idea of this approach is to design a model that can quickly adapt and learn from a small number of samples when faced with a new task. In 2020, Fallah et al. proposed an effective meta-learning approach to federated learning, which can train a good initial model in a short time.

WHAT'S THE PRINCIPLE?

The function that the client needs to optimize in Per-FedAvg is obtained after a gradient descent based on the FedAvg function. This not only preserves the benefits of federated learning (federated all client data), but also captures the differences between different users: the client can modify the initial model based on its own data to get its own model. And because the first gradient and the second gradient calculation is huge. Therefore, this paper presents an approximate method, that is, select a batch of data locally in the client, and then use this batch of data to get an unbiased estimate of the first gradient. An unbiased estimate of the quadratic gradient can also be obtained by taking a batch of data. Please refer to the article for details.

WHAT ARE THE ADVANTAGES?

The main problem of Meta-learning is how to make the model adapt to new tasks quickly under the limited data and time resources. Its advantages include: 1. Sample efficiency: Meta-learning enables the model to learn quickly on new tasks through a small number of samples, which is especially useful for scenarios where data is scarce. 2. Fast adaptability: When facing new tasks, the Meta-learning model can be adjusted quickly without starting training from scratch, which is very valuable in dynamic environments.

METHOD 2: Personalization Layers

WHAT IS Personalization Layers?

Personalization Layers is a deep feedforward neural network method for personalized training in federated learning. This method, proposed by Arivazhagan et al., solves the problem of different data distribution.

WHAT'S THE PRINCIPLE?

The personalization layer is some layer in the fixed client network model that does not undergo parameter updating and aggregation with the server model. This allows each client to retain adaptability to its own data set, eliminating the problem of varying effects due to different data distributions. Please refer to the article for details.

WHAT ARE THE ADVANTAGES?

Personalization Layers solves the problem of statistical heterogeneity of data in federated learning, especially in personalization tasks such as recommendation systems, fraud detection, etc. In these tasks, the data distribution on different user devices can vary greatly, which can seriously affect the performance of traditional federated learning algorithms. The advantages of Personalization Layers include: 1. Improved performance: By separating the base layer and the personalization layer, the model is better able to adapt to different data partitions, improving the convergence speed of the model in the global aggregation round and the final average test accuracy achieved by the customer. 2. Reduce the differences between clients: Personalization Layers can reduce the differences in test accuracy between different clients, which is important for learning fairness. 3. Adaptability: The personalization layer can adapt to the specific data and preferences of each user, so that the model can provide more personalized services for each user. 4. Flexibility: The method allows the number and structure of personalization layers to be flexibly adjusted in different application scenarios to meet different personalization needs.

METHOD 3: FedAvg

WHAT IS FedAvg?

FedAvg (Federated Averaging) is a core algorithm in federated learning. It was proposed by Brendan McMahan et al in 2017 to train machine learning models in a distributed way without centralizing data.

WHAT'S THE PRINCIPLE?

FedAvg's principle is to spread the model training process across multiple clients (such as mobile devices or distributed servers), with each client training the model using local data and then sending model parameters (such as weights) back to the central server. The server averages the model parameters sent by all clients, generates the global model, and distributes the average model to all clients, and the client uses the global model again for the next round of local training. This process is repeated until the model converges. The core of FedAvg's algorithm is that it allows the model to be trained using distributed data sources while maintaining data privacy.

WHAT ARE THE ADVANTAGES?

The main problem FedAvg solves is how to effectively train machine learning models using distributed data sources while protecting user privacy and data security. Its advantages include: 1. Data privacy protection: Since the data does not need to leave the user's device, the user's privacy is protected. 2. Reduced communication costs: FedAvg reduces the need to transfer data between the client and server, transferring only model parameters, which is especially useful in cases of large data volumes or limited network bandwidth. 3. Scalability: FedAvg can scale to thousands of clients, suitable for large-scale distributed training scenarios.

RESULT

The model we trained can be seen from the results with an accuracy of more than 95%. After training by meta-learning, an initial model with good effect can be quickly obtained for subsequent training of personalized federated learning model to improve the effect. See our demo notebook for details.

RECOMMENDATIONS

In summary, the federated learning framework that integrates meta-learning, personalization layer, and FedAvg can learn the underlying relationships of the Loan dataset in a short period of time and at low cost, resulting in an outstanding effect. Demo notebooks and models of the model can be found on GitHub below: https://github.com/liuchuang00/PTFI_Task_2

REFERENCES

[1] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized Federated Learning: A Meta-Learning Approach," arXiv preprint arXiv:2002.07948, 2020.

[2] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated Learning with Personalization Layers," arXiv preprint arXiv:1912.00818, 2019.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Artificial intelligence and statistics. PMLR, 2017, pp. 1273–1282.

[4] Sunil Jacob (2024). Bank_Loan_modelling [Dataset]. Kaggle Repository.
Available: https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling/data

[5] Z. Zhao, A. Kunar, R. Birke, H. Van der Scheer, and L. Y. Chen, "CTAB-GAN+: Enhancing Tabular Data Synthesis," Frontiers in big Data, vol. 6, p. 1296508, 2024.