|  |
| --- |
| Flower Federated Learning experiment |
| PRIVACY TECHNOLOGIES FOR FINANCIAL INTELLIGENCE: A Data Bytes Company Project |
| PTFI AI (Artificial Intelligence) team | DEAKIN UNIVERSITY | April 7, 2024 |

# Introduction

# The classical approach to machine learning is that all the data will be collected on a central server where the model resides. This approach does not allow for collaboration among parties due to the sensitive nature of data. The Privacy Technologies for Financial Intelligence project aims to assist with preventing and understanding financial crime through utilising financial technology and exploring prominent privacy technologies[[1]](#footnote-1). The classical approach does not work for real world financial cases.

Federated Learning will reverse the classical machine learning approach[[2]](#footnote-2). It allows for model training across multiple decentralized devices without sharing sensitive data. Instead of sending all the data to a central server for training a global model, the model will be sent to each device server. The devices will train the model in their own server on their own data and share only model parameters with the central server or other devices.

**Use-Case: Federated Learning**

A common scenario which financial institutions (FI) face is attempting to identify fraudulent transactions within their transaction datasets. Generally, the data used for modelling is limited to an individual FI’s data which can lead to lower accuracy models and problems with generalization to real world data. Federated learning offers a solution to this scenario where FI’s can collaborate without the concerns or privacy issues, improve model performance through training on much more data and by leveraging adaptive learning to tackle fraudulent transactions more swiftly and efficiently.

In this experiment, we will aim to apply a federated learning algorithm to multiple synthetic bank transaction datasets. The purpose will be to attempt to train the data from local nodes (datasets) and to generate global model parameters which will then be applied to new datasets to create predictions.

The aim is to create an iterative learning approach which utilizes data from a range of data sources i.e. banks instead of a single bank. The focus will be on flagging suspicious transactions which might be fraudulent so that they can be investigated further.

**Method: Federated Learning**

There are many ways to apply federated learning, however, the experiment will focus on using Flower Federated Learning Framework in python. The main reasons for choosing Flower are ease of use, code flexibility, technical resources/documentation and efficient performance profile.

The experiment notebook was adapted from an existing Flower Federated Learning simulation example which can be located via <https://colab.research.google.com/github/adap/flower/blob/main/examples/simulation-tensorflow/sim.ipynb>3

The main reasons for choosing Google Colab to run the simulation were accessibility making it easy to get started, GPU support and integration with flower frameworks.

The notebook contains the following sections:

1. **Install and import necessary libraries**
2. **Data pre-processing:** in this experiment, the Anti Money Laundering Transaction Data (SAML-D) located in Kaggle was used. This dataset was also used for synthetic data generation task. The dataset was rebalanced due to a large class imbalance where fraud transactions only represented 0.1% of all transactions. The majority class i.e. non-fraudulent transactions were down sampled with a ratio of 3 non-fraudulent transactions to 1 fraudulent transaction. Not all features were used due to lack of variability and lack of modelling value i.e. date, time, bank location. Data was scaled but not transformed (this was considered after the experiment was completed).
3. **Flower Simulation:** applied federated learning on partitions of the training dataset. A simple sequential deep learning model was used across 10 clients (partitions) and enabled with GPU. The plots show the model improved with the first few clients then stabilized around the 78% accuracy mark,
4. **Training without federated learning:** a subset of the training data (20%) was trained without federated learning to determine if there was a difference in accuracy between the two approaches.
5. **Evaluation of results:** also covered further in this report.

As shown, the method deviates from the original proposed structure of the experiment. The notebook does not generate synthetic data or apply the model parameters to unseen datasets (although the test dataset could be considered a small example of this). The reasons for the alterations in approach were due to time constraints and more research required for adapting the code for this specific use case (fraud detection).

**ResultS**

The notebook successfully performed a federated learning simulation using flower and a standard local model without federated learning.

The results show that the accuracy using federated learning was around 78% compared with 76% without using federated learning. This indicates that for this dataset, federated learning did not improve the accuracy of the predictions. There might be some valid reasons for this such as:

* Class imbalance issues
* Reduced feature input/simple dataset
* Other metrics might be more useful
* Model only trained with one epoch and with basic model.

Given the results, there is scope to expand upon this experiment to incorporate multiple datasets (synthetic or real) instead of partitioning one large dataset, attempt different model types such as logistic regression or gradient boosted trees and customise the strategies and client side execution through config dictionaries.

For future experiments with larger datasets and more complex modelling, it would be advisable to consider running the experiment locally i.e. not on a cloud environment such as Google Colab. The reasons for this include slow processing times, high memory usage and data privacy concern with uploading data.

However, for smaller scale experiments, the use of Google Colab is encouraged due to it’s simplicity and explainability. The choice of environment will be dependent on the scope of the project work.

**REFERENCEs**

1. DataBytes Organisation. (2024). Privacy Technologies for Financial Intelligence. [GitHub Repository]. Retrieved from <https://github.com/DataBytes-Organisation/Privacy-Technologies-for-Financial-Intelligence>
2. FlowerAI. (2024). Tutorial Series: What is Federated Learning? [Webpage]. Retrieved from <https://flower.ai/docs/framework/tutorial-series-what-is-federated-learning.html>
3. Adap. (2024). Simulation TensorFlow with Flower. [Google Colab Notebook]. Retrieved from <https://colab.research.google.com/github/adap/flower/blob/main/examples/simulation-tensorflow/sim.ipynb>

1. https://github.com/DataBytes-Organisation/Privacy-Technologies-for-Financial-Intelligence [↑](#footnote-ref-1)
2. https://flower.ai/docs/framework/tutorial-series-what-is-federated-learning.html [↑](#footnote-ref-2)