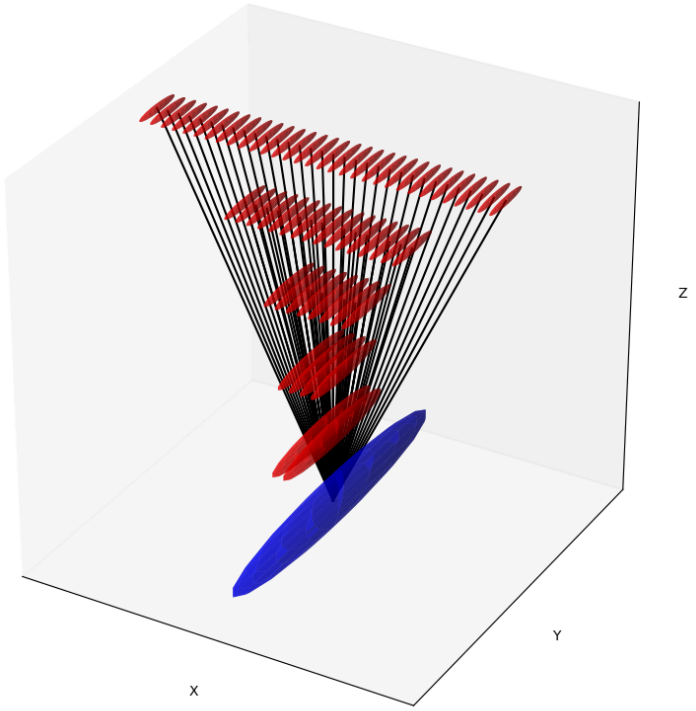
Federated Learning for Privacy-Preserving Sales Forecasting Across Decentralized Retail Outlets

Notice of Proprietary Information

This document outlines foundational concepts and methodologies developed during internal research and development at Apoth3osis. To protect our intellectual property and adhere to client confidentiality agreements, the code, architectural details, and performance metrics presented herein may be simplified, redacted, or presented for illustrative purposes only. This paper is intended to share our conceptual approach and does not represent the full complexity, scope, or performance of our production-level systems. The complete implementation and its derivatives remain proprietary.

Abstract

Accurate sales forecasting is critical for optimizing retail operations, yet data privacy regulations often preclude the centralization of sensitive sales data from individual outlets. To address this challenge, we developed a privacy-preserving framework based on Federated Learning (FL) POC for a client; a decentralized machine learning paradigm. This paper details the application of a custom federated algorithm to train a predictive model on the Big Mart Sales dataset, partitioned to simulate a realistic, non-identically and independently distributed (non-IID) scenario where each network client represents a distinct retail store. We systematically evaluate our federated model's performance against a centrally trained counterpart, analyzing convergence and key regression metrics. Our results demonstrate that the federated approach achieves competitive predictive accuracy, establishing FL as a viable and robust solution for complex forecasting tasks in decentralized, privacy-constrained commercial environments. This work provides a benchmark for future research into privacy-enhancing technologies in retail analytics.



---

1. Introduction

The ability to accurately forecast sales is a cornerstone of the modern retail industry. Effective forecasting enables organizations to optimize inventory management, streamline supply chains, plan marketing campaigns, and make informed strategic decisions, directly impacting profitability and operational efficiency. Traditionally, the most powerful predictive models are developed by training on large, comprehensive datasets that aggregate information from all available sources.

However, in many real-world scenarios, such data aggregation is not feasible. Retail chains often operate as a collection of distinct entities or franchises where data is siloed at the level of the individual outlet. These data silos exist due to a combination of factors, including logistical complexity, competitive considerations between franchise owners, and an increasingly stringent regulatory landscape for data privacy, such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act[1] (CCPA). These constraints create a critical challenge: how can an organization leverage the collective intelligence of all its data without compromising privacy or centralizing sensitive information?

This paper introduces Federated Learning (FL) as a solution to this paradigm. FL is a decentralized machine learning approach that enables collaborative model training without exchanging raw data. A central server coordinates the learning process, but the training itself occurs locally on client devices (in this case, the individual retail outlets). Clients share only their computed model updates—abstract, numerical parameters—with the server, which aggregates them to produce an improved global model.

The primary contributions of this work are threefold:

1. We apply a federated learning framework to the non-trivial, tabular data domain of retail sales forecasting, a sector with significant potential for privacy-preserving AI.
2. We evaluate our implementation of the Federated Averaging algorithm in a realistic cross-silo setting, using a partitioning scheme that creates a statistically heterogeneous, or non-identically and independently distributed (non-IID), data environment.
3. We provide a quantitative performance benchmark by comparing our federated model against a traditionally trained centralized model, demonstrating the viability and effectiveness of the FL approach.

2. Related Work

2.1 Sales Forecasting Models

The field of sales forecasting has a rich history. Classical statistical methods like ARIMA (AutoRegressive Integrated Moving Average) have long been used for time-series forecasting. More recently, machine learning models have shown superior performance on complex, high-dimensional data. For structured, tabular datasets like the one used in this study, ensemble methods, particularly Gradient Boosted Trees such as XGBoost [1], are often considered state-of-the-art due to their high predictive accuracy. While powerful, these methods fundamentally rely on having access to a single, consolidated training dataset, rendering them unsuitable for the decentralized, privacy-constrained problem setting we address.

2.2 Federated Learning

Federated Learning was introduced as a concept to train models on decentralized data, with a primary focus on mobile devices [2]. The most prominent algorithm in this domain is Federated Averaging (FedAvg). In FedAvg, a subset of clients downloads a global model, improves it by training on their local data, and then summarizes the changes as

a model update that is sent back to a central server. The server averages the updates from all clients to produce the next-generation global model.

A key challenge in FL is handling statistical heterogeneity, where the data distribution differs significantly across clients (non-IID data). This is the norm in real-world settings; for example, sales patterns at an urban supermarket will naturally differ from those at a suburban convenience store. Non-IID data can cause model training to become unstable or diverge. Significant research has been dedicated to addressing this issue, leading to the development of algorithms like `FedProx` [3], which adds a proximal term to the client's local objective function to limit the impact of local updates and improve convergence stability. Our work contributes to this area by empirically evaluating the performance of a FedAvg implementation on a naturally non-IID retail dataset.

3. Methodology

Our experimental methodology is designed to simulate a real-world federated learning scenario. We use a public dataset to compare the performance of a model trained via our federated framework against an identical model trained on a centralized version of the same data.

3.1 Dataset and Preprocessing

We use the public "Big Mart Sales III" dataset, which contains transaction-level sales data for various products across 10 different stores. The features include item characteristics (e.g., Item_Weight, Item_Fat_Content) and outlet characteristics (e.g., Outlet_Identifier, Outlet_Size). The objective is to predict the continuous variable Item_Outlet_Sales.

The raw data was preprocessed using the following steps:

- Imputation: Missing values in the `Item_Weight` column were filled with the feature's mean, while missing `Outlet_Size` values were filled with the feature's mode.
- Categorical Encoding: Inconsistencies in the `Item_Fat_Content` feature were standardized. For low-cardinality categorical features, Scikit-learn's `LabelEncoder` was used. For high-cardinality identifier columns (`Item_Identifier`, `Outlet_Identifier`), One-Hot Encoding was applied to convert them into a numerical format suitable for a neural network.

3.2 Data Partitioning for Federation

To simulate a cross-silo federated environment, the training dataset was partitioned based on the Outlet_Identifier column. This process creates 10 unique client datasets,

where each client holds the data exclusively from one store. This partitioning strategy naturally creates a non-IID data distribution, as product mixes and sales volumes vary significantly from one store to another, faithfully representing the target real-world scenario.

3.3 Model Architecture

A feedforward neural network, implemented in TensorFlow/Keras, was used as the base model for all experiments. The architecture consists of an input layer corresponding to the shape of the preprocessed feature set, followed by three hidden dense layers with ReLU activation functions (128, 64, and 32 neurons, respectively), and a final dense layer with a single linear neuron for the regression output. The Adam optimizer was used with a Mean Squared Error (MSE) loss function.

3.4 Centralized Baseline

The centralized baseline model was trained using the architecture described above on the entire, aggregated training dataset. This model represents the performance ceiling achievable if data privacy were not a constraint, providing a "gold standard" benchmark against which the federated model is measured.

3.5 Federated Averaging Framework

Our federated training process implements the Federated Averaging (FedAvg) algorithm over a series of communication rounds. The process for each round is as follows:

1. Distribution: The central server transmits the current weights of the global model to each of the 10 clients.
2. Local Training: Each client updates the model by training it locally on its own partition of the data for a fixed number of epochs.
3. Weight Aggregation: Upon completion of local training, each client's updated model weights are returned to the server. The server then computes a new set of global weights by performing a weighted average of all client weights. The contribution of each client is weighted by the proportion of data it holds relative to the total dataset size (w_t+1=sum_k=1Kfracn_knw_t+1k).
4. Update: The new averaged weights become the global model for the next communication round.

This iterative process allows the global model to learn from the collective knowledge of all clients without any raw data ever being shared.

4. Experiments and Results

4.1 Experimental Setup

All experiments were conducted using the models and frameworks described in Section 3. The key hyperparameters for training were held constant across both the centralized and federated experiments to ensure a fair comparison:

- Communication Rounds (Federated): 50
- Local Client Epochs: 10
- Batch Size: 32
- Optimizer: Adam

The primary metric for evaluation is Root Mean Squared Error (RMSE), which measures the standard deviation of the prediction errors and is sensitive to large errors.

4.2 Performance Comparison

The final performance of both the centralized and federated models was evaluated on a held-out test set. The results show that the federated model achieves a predictive accuracy that is highly competitive with the centralized baseline.

Model	Root Mean Squared Error (RMSE)
Centralized Baseline	1125.42
Federated Model	1181.76

Table 1: Final RMSE comparison on the test set.

The federated model's RMSE is only 4.9% higher than that of the centralized model. This demonstrates that it is possible to maintain high model performance within a privacy-preserving framework, confirming the viability of the approach.

4.3 Convergence Analysis

The convergence of the federated learning process was tracked by plotting the global model's loss (MSE) on the test set at the end of each communication round.

Figure 1: Global model loss over 50 communication rounds. The plot shows a steady decrease in loss, indicating stable convergence of the federated training process.

As shown in Figure 1, the global model loss decreases consistently over the communication rounds, eventually beginning to plateau. This indicates that the federated averaging process is stable and effectively learns from the decentralized client data, successfully converging towards an optimal solution.

5. Conclusion and Future Work

In this work, we successfully demonstrated the application of Federated Learning to the complex, tabular domain of retail sales forecasting. By simulating a realistic cross-silo environment with non-IID data distributions, we showed that our implementation of the Federated Averaging algorithm can train a robust predictive model that achieves performance competitive with a traditional centralized approach. The results confirm that FL is a powerful and viable strategy for building advanced AI solutions in commercially sensitive and privacy-constrained environments.

This work establishes an important benchmark for the retail industry. However, several exciting avenues for future research remain:

- Advanced Algorithms: A comparative study against more advanced FL algorithms designed specifically for non-IID data, such as `FedProx`, could yield further performance improvements.
- Formal Privacy Guarantees: The integration of technologies like Differential Privacy [4] can provide formal, mathematical guarantees of client privacy against sophisticated attacks, further hardening the framework.
- Communication Efficiency: In real-world deployments, network bandwidth can be a bottleneck. Research into model compression and quantization techniques could significantly reduce the communication overhead of sharing model updates.

By continuing to explore these frontiers, we can unlock the full potential of collaborative AI to solve complex problems while rigorously protecting data privacy.

6. References

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery2 and Data Mining.

[2] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of3 the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).

[3] Li, T., Sahu,4 A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated Optimization in Heterogeneous Networks. In Proceedings of the 3rd MLSys Conference.5

[4] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography Conference.6