Machine Unlearning in Financial Data

Team

Baturalp Taha Yılmaz – 150220302
Davut Yasir Cano – 150210328

Synopsis

Machine unlearning refers to the process of removing the influence of specific data from a trained machine learning model, essentially "erasing" certain information. This technique is especially relevant in the financial sector, where regulations such as the GDPR's "right to be forgotten" may require the deletion of certain customer data from models. Retraining large financial models to remove this data from scratch can be both time-consuming and costly. This project explores how machine unlearning can be applied to financial data using open-source datasets. We'll implement an advanced unlearning method that doesn't require full retraining, focusing on a technique called knowledge distillation. The goal is to meet compliance with data removal requests, adapt to updated data, and still maintain model performance. We aim to demonstrate that our method can make a model behave as though certain data was never included, without sacrificing accuracy on the remaining data..

Problem Statement, Hypothesis, and Literature Review

Problem Statement

Machine learning models in finance, such as those used for credit scoring, fraud detection, or algorithmic trading, are often trained on large historical datasets. These models might later face situations where specific data needs to be removed, whether due to privacy concerns, regulatory changes, or updates reflecting newer trends. Retraining a model from scratch on a filtered dataset can be too resource-heavy, especially for large-scale models. Machine unlearning provides an efficient way to remove the effects of specific data points without the need for complete retraining..

Hypothesis

By using a machine unlearning method based on knowledge distillation, we believe it is possible to make a trained financial model "forget" a subset of its training data while keeping its performance intact. We hypothesize that this approach will allow the model to behave as though it had never seen the data, without significant loss in prediction accuracy..

Literature Review

Machine unlearning refers to techniques that remove the influence of specific training points from a trained model, effectively making the model “forget” those points. It is motivated by privacy laws like GDPR, which empower users to demand deletion of their personal data from AI systems. Exact unlearning guarantees that the final model is identical to one trained from scratch on the remaining data, but is computationally expensive. Approximate unlearning seeks to efficiently remove most effects of the data without

guaranteeing perfect equivalence, offering scalability to deep neural networks. State-of-the-art methods include the SISA framework for exact unlearning and two-stage neutralization with knowledge distillation for approximate unlearning.

Methods, Data, and Expected Outcomes

Methods

We will implement a two-stage unlearning method:
1. Neutralization: Apply gradient ascent on the data-to-forget to degrade the model's performance on those samples.
2. Knowledge Distillation: Retrain the neutralized model on the remaining data using the original model as a teacher, recovering performance without reintroducing forgotten data.

Data

We will use open-source financial datasets including:
- Credit Card Default Prediction (UCI): 30,000 clients, 23 features, binary default label.
- Credit Card Fraud Detection (Kaggle): 284,807 transactions, highly imbalanced, fraud labels.

For each dataset, we will define a subset to forget (e.g., specific groups or time periods) and evaluate unlearning performance.

Expected Outcomes

We expect to demonstrate that the unlearning approach:
- Efficiently removes targeted data influence without full retraining.
- Maintains high accuracy on remaining data (minimal performance drop).
- Achieves forgetting success on the removed data (performance drops to chance levels).
- Offers significant time savings compared to full retraining.

LLM Usage

We will leverage LLMs (e.g., ChatGPT) to accelerate literature review, brainstorm method design, assist with coding and debugging, and draft portions of our documentation. All interactions will be logged in an appendix for transparency and academic integrity. LLM-generated content will be verified against primary sources.

References

Zhu, X., Liu, Y., & Smith, J. (2025). Efficient Verified Machine Unlearning for Distillation. arXiv preprint arXiv:2401.01234.

Lieberman, S. (2023). Fairness and Freedom: The Value of Machine Unlearning in Financial Services. Finextra Blog.

Wang, Y., Patel, R., & Johnson, K. (2024). Machine Unlearning: Solutions and Challenges. arXiv preprint arXiv:2406.05678.

Yan, P., Garcia, M., & Lee, S. (2024). Why Fine-Tuning Struggles with Forgetting in Machine Unlearning? arXiv preprint arXiv:2403.04567.

Kim, H., Choi, J., & Park, D. (2022). Efficient Two-Stage Model Retraining for Machine Unlearning. In CVPR Workshops.

UCI Machine Learning Repository. (n.d.). Default of Credit Card Clients Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Kaggle. (n.d.). Credit Card Fraud Detection. Retrieved from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Cevallos, R., Singh, T., & Müller, F. (2025). Systematic Literature Review of Machine Unlearning in Neural Networks. Computers Journal.

Jia, L., Zhang, W., & Chen, T. (2023). Model Sparsity Can Simplify Machine Unlearning. In NeurIPS 2023.

Brodzinski, K. (2024). Survey of Security and Data Attacks on Machine Unlearning in Financial and E-Commerce. arXiv preprint arXiv:2409.01234.