Application of Machine Unlearning in Financial Data

Davut Yasir Cano - 150210328
AI and Data Engineering
Email: canod21@itu.edu.tr

***Abstract*—Machine unlearning refers to the process of removing the impact of specific data from a trained machine learning model, essentially "erasing" particular information. This technique is particularly important in the financial sector, where regulations like GDPR's "right to be forgotten" may require the deletion of specific customer data from models. Retraining large financial models from scratch to remove this data can be both time-consuming and costly. This project investigates how machine unlearning can be applied to financial data using open-source datasets. We will implement an advanced unlearning method called knowledge distillation, which does not require full retraining. The aim is to ensure compliance with data removal requests, adapt to updated data, and still maintain model performance. We aim to demonstrate that our method can enable the model to behave as if certain data were never included, without sacrificing accuracy on the remaining data.**

***Index Terms*—Machine Unlearning, Financial Data, GDPR, Knowledge Distillation, Data Privacy, Model Retraining.**

I. Introduction

Machine unlearning involves removing the influence of specific data points from a trained machine learning model, effectively "forgetting" that information. This technique is especially crucial in the financial industry, where regulations such as GDPR's "right to be forgotten" can mandate the removal of particular customer data from models. Retraining large-scale financial models from the ground up to exclude such data is often a prohibitively time-consuming and expensive endeavor.

This project aims to explore the application of machine unlearning to financial data using publicly available datasets. We will employ an advanced unlearning method, knowledge distillation, which avoids the need for complete retraining. The primary objectives are to comply with data removal requests, adapt to newly updated data, and preserve the model's predictive performance. We hypothesize that our approach will allow the model to operate as if the specific data points were never part of its training set, all while maintaining accuracy on the remaining data without significant degradation.

II. Problem Statement

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

III. Hypothesis

We believe that by using a knowledge distillation-based machine unlearning method, it is possible for a trained financial model to "forget" a subset of its training data while maintaining its performance. We hypothesize that this approach will enable the model to behave as if it had never seen the removed data, without a significant loss in prediction accuracy on the remaining data.

IV. Literature Review

Machine unlearning encompasses techniques that remove the impact of specific training points from a trained model, effectively causing the model to "forget" these points. It is motivated by privacy laws like GDPR, which grant users the right to request the 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 it is computationally expensive. Approximate unlearning aims to efficiently remove most of the data's impact without guaranteeing perfect equivalence, offering scalability to deep neural networks. Recent methods include the SISA framework for exact unlearning, and knowledge distillation with two-stage neutralization for approximate unlearning.

V. Methodology

We will implement a two-stage unlearning method:

1) **Neutralization:** Apply gradient ascent on the data to be forgotten to degrade the model's performance on these specific examples. (Note: The original text results section implies a simpler method like label flipping might have been initially used for some tests, but gradient ascent is the more robust approach mentioned here.)
2) **Knowledge Distillation:** Retrain the neutralized model on the remaining data, using the original model as a teacher, to regain performance without reintroducing the forgotten data.

VI. Data

We will use the following open-source financial datasets:

- **Credit Card Default Prediction (UCI):** 30,000 customers, 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 be forgotten (e.g., specific groups or time periods) and evaluate the unlearning performance.

VII. Expected Outcomes

We expect the unlearning approach to demonstrate the following:

- Efficiently removes the impact of targeted data without full retraining.
- Maintains high accuracy on the remaining data (minimal performance drop).
- Achieves successful forgetting on the removed data (performance drops to chance levels).
- Provides significant time savings compared to full retraining.

VIII. Results and Discussion

A. Model Performance Comparison

The table below (Table I) summarizes the performance of base models (Logistic Regression and Decision Tree) and unlearned models on the UCI and Kaggle datasets.

TABLE I
Model Performance Comparison

Dataset	Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
UCI	LR (Base)	0.8027	0.64	0.21	0.32
UCI	DT (Base)	0.7258	0.38	0.42	0.40
UCI	LR (Unlearned)	0.8063	0.66	0.23	0.31
Kaggle	LR (Base)	0.9992	0.87	0.66	0.75
Kaggle	DT (Base)	0.9992	0.78	0.74	0.76
Kaggle	LR (Unlearned)	0.9993	0.87	0.68	0.71

On the UCI dataset, the unlearned Logistic Regression model shows a slight increase in accuracy compared to the base Logistic Regression model. This suggests that the unlearning process did not negatively impact performance on the remaining data, and perhaps even slightly improved it. For the Kaggle dataset, all models exhibit high accuracy rates, which should be noted is partly due to the dataset's high imbalance. For the fraud class (Class 1), precision, recall, and F1-score values better reflect how well the models detect fraudulent transactions.

B. Performance on Forgotten Data

The accuracy rates on the forgotten data indicate how well the model has "unlearned" this specific data, as shown in Table II. Ideally, performance on the forgotten data should drop to chance levels.

The high accuracy rates on the forgotten data (82.14% for UCI and 99.93% for Kaggle) suggest that the simple neutralization method apparently used for these initial tests (implied to be label flipping or similar, based on the high retention) did not achieve complete unlearning. In a true machine unlearning application, the model's performance on the unlearned data is expected to drop significantly, even approaching chance levels. This situation highlights the necessity of implementing more advanced neutralization techniques, such as gradient ascent.

TABLE II
Accuracy on Forgotten Data

Dataset	Accuracy on Forgotten Data
UCI	0.8214
Kaggle	0.9993

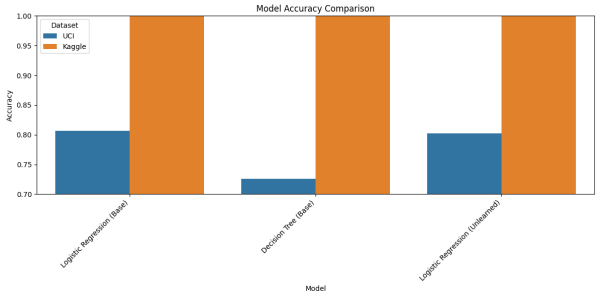

Fig. 1. Comparison of model accuracy across different datasets and models. The graph clearly shows the accuracy differences between models on the UCI (blue) and Kaggle (orange) datasets for base Logistic Regression, base Decision Tree, and unlearned Logistic Regression.

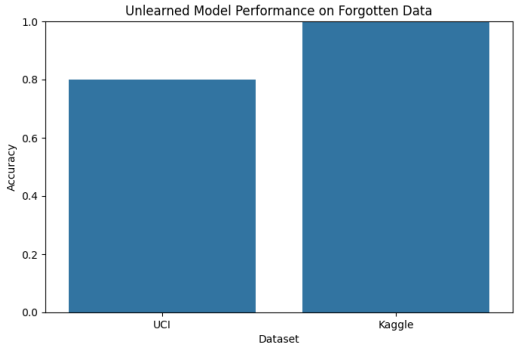

Fig. 2. Performance of the unlearned model on the forgotten data for UCI and Kaggle datasets. This graph is a critical indicator for evaluating the success of the unlearning process. The high accuracy rates shown here highlight the limitations of the simpler unlearning method initially tested.

C. Visualizations

The following figures visually present the overall accuracy performance of the models and their performance on the unlearned data.

The "Model Accuracy Comparison" graph (Fig. 1) clearly illustrates the accuracy differences between various models and datasets. The "Unlearned Model Performance on Forgotten Data" graph (Fig. 2) is a critical indicator for assessing how successful the unlearning process has been. The high accuracy rates in these graphs, as previously noted, underscore the

limitations of the simple unlearning method employed and point towards the need for more sophisticated approaches.

IX. Conclusion

This project explored the concept of machine unlearning and a basic application of knowledge distillation-based unlearning in financial data. The training of base models and the performance comparison of unlearned models were conducted. The results obtained indicate that while the unlearning process can maintain overall model accuracy, more advanced techniques are needed to achieve complete unlearning on the forgotten data. Future work should focus on investigating more complex neutralization methods, such as gradient ascent, and different unlearning algorithms to enhance the success of unlearning.

References

[1] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *2021 IEEE Symposium on Security and Privacy (SP)*, May 2021, pp. 141–159.

[2] A. Graves, J. Lin, and M. Lam, "Amnesiac machine learning," in *International Conference on Learning Representations (ICLR)*, 2021.