On Performance Comparison Between Strong Machine Unlearning Algorithms for Logistic Regression Credit Assessment Models

Jeremy Syaloom Okey Nathanael Simbolon
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
jeremysimbolon@protonmail.com

Windy Gambetta
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
windy@itb.ac.id

***Abstract*—The enactment of the Personal Data Protection (PDP) regulation in Indonesia requires financial service institutions to erase debtors' personal data upon request. However, this is challenging to achieve when such data is implicitly stored in a trained machine learning model. In order to address this issue, machine unlearning methods have been developed to erase the influence of training data on model weights. In this research, the performance of two strong machine unlearning algorithm implementations, ε-δ Certified Removal (CR) and Projective Residual Update (PRU), was compared on the logistic regression credit risk assessment models developed in this research. Machine unlearning was performed to delete 10% of the model's training data. Our results showed that ε-δ Certified Removal yielded a model with lower L^2-distance, higher accuracy, and faster unlearning time compared to Projective Residual Update when machine unlearning was performed to delete less than 2% of the model's training data. Conversely, the opposite was observed when machine unlearning was performed to delete more or equal to 2% of the model's training data. Further research is required to explore the effect of larger training data sets with greater dimensionality on the performance of both algorithms.**

Keywords—machine unlearning, ε-δ certified removal, projective residual update, performance, credit risk assessment

I. Introduction

Credit risk assessment is a risk management practice performed by financial service institutions used to perform informed decision-making with respect to granting loans to debtors [1]. In order to accomplish this, financial service institutions have developed machine learning models based on their past credit granting decisions to identify credit-worthy debtors. The complexity of such models varies depending on the institution's needs, ranging from linear models [2] to deep learning models [3].

The enactment of the Personal Data Protection (PDP) regulation in Indonesia requires financial service institutions to erase debtors' personal data upon request [4]. This has created an urgent necessity for financial service institutions to implement an effective and efficient data removal mechanism from machine learning models. Unfortunately, deleting debtors' personal data from such models is nontrivial due to the difficulties of identifying their influence toward the model's weights.

The high computational [5] and energy [6] costs associated with retraining a machine learning model make machine unlearning a viable solution that can be pursued by financial service institutions. Machine unlearning is performed to erase the influence of training data on machine learning model weights [7]. Past publications have discussed the development of machine unlearning frameworks and algorithms that utilized model-agnostic, model-intrinsic, and data-driven approaches [8]. Past publications have also discussed the development of machine unlearning frameworks and algorithms that produced indistinguishable models (exact unlearning), models with similar parameter distributions (strong unlearning), and models with similar activation distributions (weak unlearning) compared to naively retrained model [9]. However, comparative studies between each implementation are limited, especially between those that can be applied to logistic regression credit assessment models which have been used for the last four decades [1], [2], [10].

In this research, the performance of two strong machine unlearning algorithm implementations, ε-δ Certified Removal and Projective Residual Update, was compared within the context of credit risk assessment. To this end, logistic regression credit risk assessment models to which the algorithms will be applied were developed. Subsequently, both algorithms were evaluated to comprehend the trade-off between removal size and the resulting models' layer-wise distance (L^2-distance) to naively retrained models, accuracy, and time needed to unlearn the data points given. The result is intended to provide insights for financial service institutions that need to implement data removal mechanisms for their credit risk assessment models.

II. Methodology

A. Data Acquisition

In order to develop the credit risk assessment models upon which our machine unlearning algorithm implementations would be evaluated, a debtors data set first described by Supardi and Gambetta [11] was utilized. The data set consisted of 200 data points with a 1:1 ratio of accepted and rejected credit applications. The data set features consisted of 23 factors considered by financial service institution's credit committee upon granting credit approvals, filtered to protect debtors' privacy. Prior to using the data set for model training, data preprocessing was performed, which included null handling, categorical encoding, and data normalization needed for logistic regression models.

B. Credit Risk Assessment Models Development

In this research, logistic regression credit assessment models entailing credit approval or rejection that will be used to evaluate the performance of our machine unlearning algorithm implementations were developed in Python. The model choice was based on its widespread adoption over the last four decades [1], [2], [10] and its good interpretability for fulfilling regulatory necessities [4].

This work was supported by the Institute of Research and Community Services (LPPM) of Bandung Institute of Technology (ITB) under the Research, Community Service, and Innovation (P2MI) program.

979-8-3315-0640-7/24/$31.00 ©2024 IEEE

Algorithm 1 ε-δ Certified Removal

Input: training data x;
training label y;
model weight w;
training data to be removed b;
ℓ_2-regularizer λ.
Output: updated model weight w'

for $i \leftarrow 1$ **to** $|b|$ **do**
  $\Delta \leftarrow |b_i|\lambda w + \sum_{j \in b_i} \nabla \ell\left(w^T x_j, y_j\right)$
  $H \leftarrow \sum_{j \notin b_i} \nabla^2 \ell\left(w^T x_j, y_j\right)$
  $w \leftarrow w + H^{-1}\Delta$
end for
return $w' \leftarrow w$

For experiment purposes, the debtor data set previously explained was utilized to develop three logistic regression credit assessment models with test set ratio $n \in \{0.2, 0.3, 0.4\}$. Afterward, the models were retrained with hyperparameter tuning to obtain the optimal inverse ℓ_2-regularizer C for each model and prevent overfitting. Hyperparameter tuning was done by performing a grid search through the hyperparameter space $C \in logspace(-4, 4, 101)$ using stratified 5-fold cross-validation guided by accuracy. The best hyperparameter C for each model would be used to train the optimal models.

C. ε-δ Certified Removal Implementation

ε-δ Certified Removal is a strong model-agnostic machine unlearning algorithm first described by Guo et al. [12]. This algorithm performs machine unlearning by applying a one-step Newton update to a trained model's weights. This process can be represented by the following equation:

$$w' = w - [\nabla^2 \ell(D^{\setminus D_u})]^{-1}(\nabla \ell(D')) \quad (1)$$

in which w' denotes the updated model weights after machine unlearning, w denotes the original weights of the model, $[\nabla^2 \ell(D^{\setminus D_u})]^{-1}$ denotes the inverse of the Hessian matrix of the model's loss function for the data set that does not contain the data points set for removal, and $\nabla \ell(D')$ denotes the gradient of the model's loss function for the data points set for removal. The pseudocode for ε-δ Certified Removal can be summarized in Algorithm 1. The algorithm has a computational cost of $O(k^2 d)$ in which k denotes the number of data points to be removed from the model and d denotes the data set dimension.

The ε-δ Certified Removal mechanism was chosen due to its theoretical guarantee to effectively delete training data and its compatibility with logistic regression models. For our experiment, Algorithm 1 was implemented in Python. Afterward, machine unlearning was performed using $\lambda \in \{0.01, 0.005, 0.001\}$ to delete $p(k) = 10\%$ of training data from the credit risk assessment models developed earlier.

D. Projective Residual Update Implementation

Projective Residual Update is a strong model-intrinsic machine unlearning algorithm first described by Izzo et al. [13]. This algorithm performs machine unlearning by using gradient methods to update a trained model's weights. This process can be represented by the following equation:

$$w' = w - (S^{-1^*})(\nabla \ell(D^*)) \quad (2)$$

Algorithm 2 Projective Residual Update

Input: training data x;
weighted least squares label z;
weighted least squares hat matrix H
model weight w;
amount of training data to be removed k.
Output: updated model weight w'

$\hat{y}'_1, \hat{y}'_2, \ldots, \hat{y}'_k \leftarrow synthetic\_data(x, z, H, k)$
$S^{-1} \leftarrow pseudo\_inverse\left(\sum_{i=1}^{k} x_i x_i^T\right)$
$\nabla L \leftarrow \sum_{i=1}^{k} (w^T x_i - \hat{y}'_i) x_i$
$w' \leftarrow w - fast\_mult(S^{-1}, \nabla L)$
return w'

in which w' denotes the updated model weights after machine unlearning, w denotes the original weights of the model, S^{-1^*} denotes the pseudoinverse of the projection matrix built from the features of the data points set for removal calculated using the Gram-Schmidt process and eigendecomposition, and $(\nabla \ell(D^*))$ denotes gradient of the model's loss function for the predicted synthetic data calculated using a generalization of leave-one-out residuals for linear models. The pseudocode for Projective Residual Update can be summarized in Algorithm 2. The algorithm has a computational cost of $O(kd^2)$ in which k denotes the number of data points to be removed from the model and d denotes the data set dimension.

The Projective Residual Update mechanism was chosen due to its high computational efficiency and its compatibility with logistic regression models. For our experiment, Algorithm 2 was implemented in Python. Afterward, machine unlearning was performed using $\lambda \in \{0.01, 0.005, 0.001\}$ to delete $p(k) = 10\%$ of training data from the credit risk assessment models developed earlier.

E. Experiment Design

The performance of our ε-δ Certified Removal and Projective Residual Update implementation was evaluated using the metrics L^2-distance, accuracy, and unlearn time. The L^2-distance metric was chosen because one could expect a model yielded by machine unlearning to make similar predictions when its L^2-distance compared to a naively retrained model was small [13]. The accuracy metric was chosen to corroborate the notion mentioned above. Finally, the unlearn time metric was chosen to obtain an overview of the speedup benefit each algorithm offered.

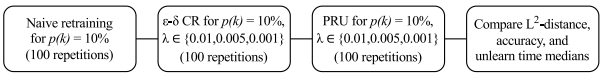

Fig. 1. Experiment scheme for machine unlearning algorithm evaluation.

To perform this evaluation, the experiment scheme illustrated with Fig. 1 was proposed for each credit assessment models with test set ratio $n \in \{0.2, 0.3, 0.4\}$. First, credit risk assessment models were naively retrained for $p(k)$ up to 10%, each with 100 repetitions. Next, machine unlearning was performed using ε-δ Certified Removal and Projective Residual Update for $p(k)$ up to 10%, each with 100 repetitions. Finally, the median of L^2-distance, accuracy, and unlearn time for each machine unlearning algorithm were analyzed.

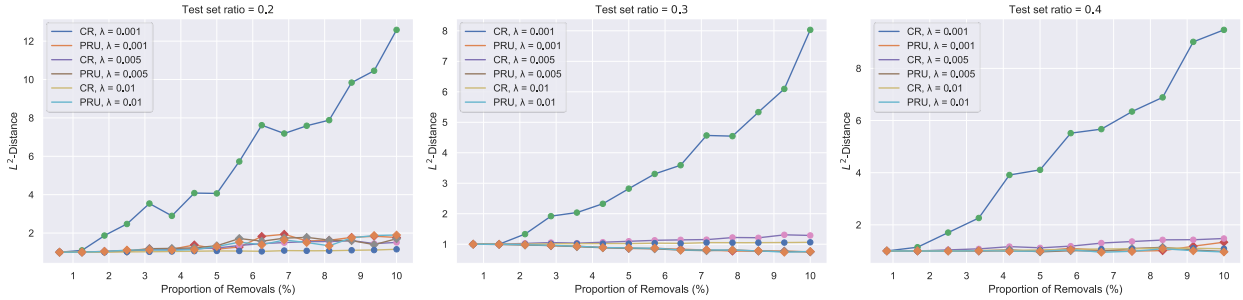

Fig. 2. Trade-off between proportion of removals and median of L^2-distance (100 repetitions). The L^2-distance is given as fraction of baseline distance $\| w^{benchmark} - w^{\setminus D_u} \|$. (a) Trade-off for model with test set ratio $n = 0.2$. (b) Trade-off for model with test set ratio $n = 0.3$. (c) Trade-off for model with test set ratio $n = 0.4$.

III. Result and Discussion

A. Benchmark Models

Three benchmark models were developed with test set ratio $n \in \{0.2, 0.3, 0.4\}$. The result of the hyperparameter tuning process and the models' performance can be found in Table 1. This result supported the suitability of logistic regression models for credit risk assessment.

TABLE I. Hyperparameter Tuning Result for Benchmark Models

Test set ratio (n)	Inverse ℓ_2-regularizer (C)	Accuracy
0.2	0.57543994	0.975
0.3	0.19054607	0.983
0.4	0.47863009	0.975

B. L^2-distance

An experiment was performed to compare the median L^2-distance of the models resulted from machine unlearning using ε-δ Certified Removal and Projective Residual Update. The result of this experiment can be found in Fig. 2.

The expected trend was observed in that the L^2-distance of the models increased given higher amount of removal k. It was also observed that ε-δ Certified Removal performed better in this metric compared to Projective Residual Update for $p(k) < 1\%$. Conversely, Projective Residual Update performed better in this metric for $p(k) \geq 1\%$. These findings were consistent with those of Izzo et al. [13], indicating that influence-based machine unlearning methods should perform better in this metric compared to Projective Residual Update for small values of k. Unfortunately, this behavior was only spotted for a single data point due to the small size of the data set used in the experiment. Regardless, ε-δ Certified Removal was concluded to be more suited for the removal of smaller values of k, whereas Projective Residual Update was more suited for larger values of k.

C. Accuracy

An experiment was performed to compare the median accuracy of the models resulted from machine unlearning using ε-δ Certified Removal and Projective Residual Update. The result of this experiment can be found in Fig. 3.

The expected trend was observed in that the accuracy of the models decreased given higher amount of removal k. It was also observed that for parameter pairs $\{\lambda = 0.001, p(k) \geq 2\%\}$ and $\{\lambda = 0.005, p(k) \geq 8\%\}$, the accuracy of the models yielded by ε-δ Certified Removal noticeably dropped after a stable accuracy value was previously maintained. These findings were consistent with those of Guo et al. [12], indicating that ε-δ Certified Removal maintains its resulting models' accuracy until a threshold value of k. The threshold k depends on the value of λ chosen during implementation. Higher values of λ are expected to allow for higher values of data removal k while maintaining the model's starting accuracy. However, this behavior failed to be spotted for $\lambda = 0.01$ due to the small size of the data set used in the experiment. Regardless, ε-δ Certified Removal was concluded to be more suited for the removal of smaller values of k, whereas Projective Residual Update was more suited for larger values of k.

D. Unlearn Time

An experiment was performed to compare the median unlearn time of the models resulted from machine unlearning using ε-δ Certified Removal and Projective Residual Update. The result of this experiment can be found in Fig. 4.

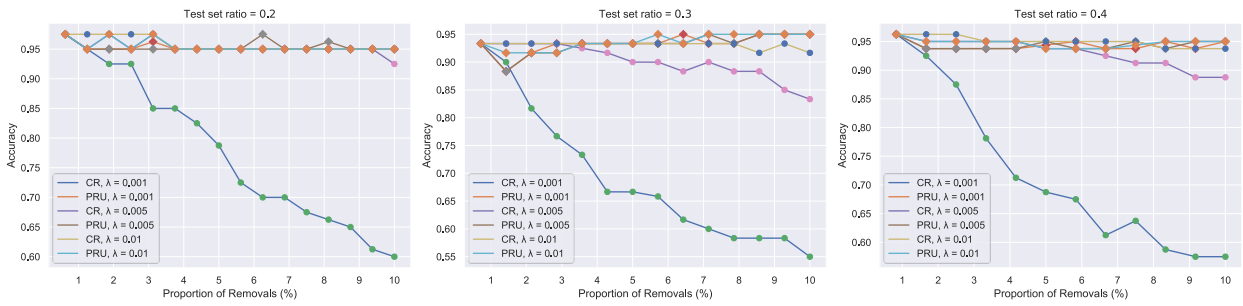

Fig. 3. Trade-off between proportion of removals and median of accuracy (100 repetitions). (a) Trade-off for model with test set ratio $n = 0.2$. (b) Trade-off for model with test set ratio $n = 0.3$. (c) Trade-off for model with test set ratio $n = 0.4$.

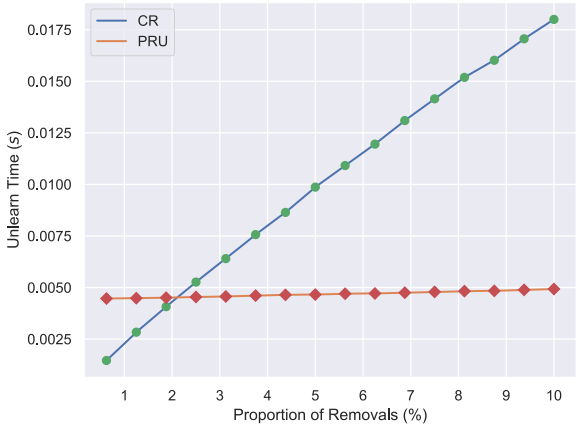

Fig. 4. Trade-off between proportion of removals and median of unlearn time (100 repetitions).

The expected trend was observed in that the unlearn time of the models increased given higher amount of removal k. It was also observed that the unlearn time of ε-δ Certified Removal increased proportionally with k, consistent with its theoretical computational cost [12]. It was also expected that the unlearn time of Projective Residual Update increased proportionally with k^2, consistent with its theoretical computational cost [13], yet only the beginning of this exponential curve was able to be observed in the result. This observation was attributed to the small size of the data set used in the experiment.

It was also noted that ε-δ Certified Removal performed better in this metric compared to Projective Residual Update for $p(k) < 2\%$. Conversely, Projective Residual Update performed better in this metric for $p(k) \geq 2\%$. Therefore, ε-δ Certified Removal was concluded to be more suited for the removal of smaller values of k if minimal unlearn time was a necessity, whereas Projective Residual Update was more suited for larger values of k and offered a more stable unlearn time.

E. Research Limitation

One potential limitation of this research is that only an empirical debtor data set with small size and low dimensionality was able to be made use of within permission. The small number of data points available affected the precision of $p(k)$ evaluated when determining the performance cutoff of both machine unlearning algorithms. Meanwhile, the low dimensionality of the data set limited the possible amount of removal k during the experiments, owing to the constraint of the Projective Residual Update algorithm [13]. Despite this limitation, the results of this research reinforce previous findings [12], [13] and provide important guidelines for selecting machine unlearning algorithms based on the number of data points to be removed. We are confident that applying our approach to other data sets should corroborate these results.

IV. Conclusion and Future Work

In this research, logistic regression credit risk assessment models were developed and a best accuracy score of 0.983 was achieved, which supported the suitability of logistic regression models for assisting financial service institutions in performing credit risk assessments of their debtors. In addition, two strong machine unlearning algorithms, ε-δ Certified Removal and Projective Residual Update, were also successfully implemented to perform machine unlearning toward logistic regression models with differing performance depending on the amount of data points to be removed k. ε-δ Certified Removal was shown to produce a model with lower L^2-distance, higher accuracy, and faster unlearn time when $p(k) < 2\%$. Conversely, Projective Residual Update produced a model with lower L^2-distance, higher accuracy, and faster unlearn time when $p(k) \geq 2\%$.

In this research, only a debtor data set of relatively small size and lesser dimensionality was able to be obtained and utilized for the experiment. Furthermore, analysis of the algorithm implementations' efficiency was focused on their computational complexity using the unlearn time metric. Further research is necessary to reproduce our findings using a data set of larger size and greater dimensionality. In addition, future research may focus on the algorithm implementations' space complexity to gain further insights into their efficiency.

References

[1] G. Sabato, "Credit Risk Scoring Models," *SSRN Electron. J.*, vol. 283, pp. 1–15, 2011, doi: 10.2139/ssrn.1546347.

[2] B. R. Marks and K. K. Raman, "State Audit Budgets and Market Assessments of Credit Risk," *J. Account. Public Policy*, vol. 5, no. 4, pp. 233–250, Dec. 1986, doi: 10.1016/0278-4254(86)90021-9.

[3] P. M. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Risks*, vol. 6, no. 2, pp. 1–20, 2018, doi: 10.3390/risks6020038.

[4] Indonesia, *Undang-Undang Republik Indonesia Nomor 27 Tahun 2022 tentang Perlindungan Data Pribadi*. Jakarta: Dewan Perwakilan Rakyat Republik Indonesia, 2022.

[5] M. Veale, R. Binns, and L. Edwards, "Algorithms that remember: Model inversion attacks and data protection law," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 376, no. 2133, 2018, doi: 10.1098/rsta.2018.0083.

[6] T. Mastelic, A. Oleksiak, H. Claussen, I. Brandic, J. M. Pierson, and A. V. Vasilakos, "Cloud computing: Survey on energy efficiency," *ACM Comput. Surv.*, vol. 47, no. 2, 2015, doi: 10.1145/2656204.

[7] Y. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," in *2015 IEEE Symposium on Security and Privacy*, May 2015, pp. 463–480, doi: 10.1109/SP.2015.35.

[8] T. T. Nguyen, T. T. Huynh, P. Le Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A Survey of Machine Unlearning," 2022, [Online]. Available: http://arxiv.org/abs/2209.02299.

[9] H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu, "Machine Unlearning: A Survey," *ACM Comput. Surv.*, vol. 56, no. 1, pp. 1–36, 2024, doi: 10.1145/3603620.

[10] Y. Yang, X. Chu, R. Pang, F. Liu, and P. Yang, "Identifying and predicting the credit risk of small and medium-sized enterprises in sustainable supply chain finance: evidence from china," *Sustain.*, vol. 13, no. 10, 2021, doi: 10.3390/su13105714.

[11] R. S. Supardi and W. Gambetta, "Knowledge-Based Credit Granting System in Indonesia with Artificial Intelligence and XAI Approach," *2023 10th Int. Conf. Adv. Informatics Concept, Theory Appl. ICAICTA 2023*, pp. 1–6, 2023, doi: 10.1109/ICAICTA59291.2023.10390466.

[12] C. Guo, T. Goldstein, A. Hannun, and L. van der Maaten, "Certified data removal from machine learning models," in *Proceedings of Machine Learning Research*, Nov. 2020, vol. 119, pp. 3832–3842, [Online]. Available: https://proceedings.mlr.press/v119/guo20c.

[13] Z. Izzo, M. A. Smart, K. Chaudhuri, and J. Zou, "Approximate Data Deletion from Machine Learning Models," in *Proceedings of Machine Learning Research*, Feb. 2021, vol. 130, pp. 2008–2016, [Online]. Available: https://proceedings.mlr.press/v130/izzo21a.html.