# Evaluating Noise Mechanisms in Gradient Boosting for Credit Default Prediction

**Linear Regression Models Final Project**

**Dec 20, 2024**

**Liang-Po Yen, Jianwei Wang, Ronghui Miao, Kangyu Zhao, Xinrong Dong**

## Introduction

In today's data-driven financial landscape, accurately predicting credit card defaults is a crucial task for financial institutions. These institutions increasingly rely on machine learning algorithms to assess the risk of customer default and predict potential losses from default. As a result, developing efficient and robust models has become a key priority for addressing the credit default risk problem.

This project has two primary objectives: (1) predicting whether a credit card customer will default on a payment, and (2) evaluating how different privacy-preserving noise mechanisms impact model performance. To achieve these goals, Gradient Boosting is employed and applied to a combination of preprocessing and resampling techniques, including principal component analysis (PCA), synthetic minority oversampling techniques (SMOTE), k-means SMOTE, and cluster centroids. These techniques address common challenges in credit default prediction, namely high data dimensionality and class imbalance.

This study examines the roles of Laplace, Exponential, and Gaussian noise in evaluating the trade-off between privacy guarantees and predictive performance within the framework of differential privacy. It also aims to achieve a balance between these two priorities and thus provide actionable insights for financial institutions to develop privacy-aware predictive models for credit default risk.

## Data Description

The dataset *Defaults of Credit Card Client* comes from the UCI machine learning repository. It contains personal information, history of past payments, bill statements, and previous payment amounts of about 30,000 credit card holders in Taiwan from April to September 2005. The dataset includes 23 features and a binary target variable *default.payment.next.month* representing whether the credit card holder will default next month.

The client's personal information includes variables *AGE*, *GENDER*, *EDUCATION*, *MARRIAGE*, *SEX*, and *LIMIT_BAL*. *LIMIT_BAL* is defined as the amount of given credit, serving as an indicator of the client’s borrowing capability. The multi-class categorical features *GENDER*, *EDUCATION*, and *MARRIAGE* are transformed using one-hot encoding to convert each unique value into a separate binary variable, making them compatible with machine learning algorithms. These variables provide a demographic overview of the client and can influence their credit behavior. *PAY_0* through *PAY_6* indicate repayment status over

the past six months (for example, 0 indicates no delays, and positive values denote the number of months delayed). *BILL_AMT1* through *BILL_AMT6* record the bill statement amounts over the past six months. *PAY_AMT1* through *PAY_AMT6* capture the amount of previous payments during the same period.

In this dataset, 22% of credit card holders are defaulters, while 78% are non-defaulters. This imbalance implies that the classification models trained on this dataset will be biased toward the majority class (non-defaulters) and overlook the minority class (defaulters). To mitigate this problem, a variety of resampling techniques are used and evaluated based on model performance (measured by F1-score), including SMOTE, k-means SMOTE, and cluster centroids. To reduce the dimensionality of the data and mitigate the risk of overfitting, PCA is employed. In this study, the first 12 principle components, which explain more than 99% of the total variance, are considered.

## Methodology

During data preprocessing, PCA helps identify patterns in data by analyzing correlations between features. This technique finds the directions of maximal variance and projects the data onto a lower-dimensional subspace. Essentially, PCA reduces the number of features in a dataset while retaining as much variance as possible. It transforms the data into a new coordinate system defined by principal components, which are linear combinations of the original features (Jolliffe, 2002). One advantage of PCA is dimensionality reduction, which enhances computational efficiency and eliminates multicollinearity among features by generating uncorrelated principal components. However, it may lose information, especially when the number of retained principal components is small.

SMOTE is an oversampling technique that generates synthetic samples for the minority class through linear interpolation. It improves the representation of the minority class and leads to better model performance, particularly in classification tasks (Chawla, 2002). Instead of simply replicating existing observations, this technique generates artificial samples by linearly interpolating a randomly selected minority observation and one of its nearest neighbors, thereby reducing the risk of overfitting. SMOTE is more efficient than naive oversampling methods and can generate more diverse samples. However, it may introduce noise or unrealistic samples in datasets with significant overlap between classes (Chawla, 2003). k-means SMOTE simply uses k-means clustering algorithm in conjunction with SMOTE. This method reduces the risk of generating noisy samples by oversampling only in “safe” regions (areas made up of at least 50% of minority samples). It tackles both

between-class and within-class imbalances by increasing the representation of sparse minority areas. Cluster centroids undersampling replace clusters of majority samples with their corresponding centroids, ensuring that the majority class is represented by fewer but more representative samples.

Gradient Boosting is a widely used machine learning technique suitable for both regression and classification tasks, such as fraud detections. It builds a strong predictive model by sequentially combining multiple weaker models. The boosting method improves performance iteratively by training models sequentially, where each model learns to correct the errors made by its predecessor. This process is a form of iterative optimization. For example, the most common weak learners are shallow decision trees, where each new tree corrects the mistakes made by the previous ones (Li, 2020). The "gradient" in Gradient Boosting refers to gradient descent optimization, which minimizes the loss function by adding models that predict the negative gradient of the loss.

For the purpose of optimizing models, choosing the appropriate mechanism, adjusting privacy budget, and data preprocessing and feature selection are essential steps. Differential attack is an analytical method that attempts to infer private information by observing differences in outputs caused by minor changes in the input data. The core idea of a differential attack is input difference and output difference. Input difference introduces a small modification to the original data, typically by adding or removing a single record, resulting in "neighboring datasets" (Biham, 1991). Output difference observes the algorithm's output differences between these neighboring datasets and analyzes whether the output discrepancy reveals sensitive information. This method leverages the small statistical differences in the output that might disclose privacy, especially when the algorithm does not sufficiently protect the data.

To address the characteristics of differential attacks, a privacy-preserving algorithm, differential privacy, is designed to obscure the impact of input changes on output by adding noise. This ensures that even if an attacker knows the entire contents of the database and observes the output results, they cannot accurately infer the existence or specific details of an individual record. For any two "neighboring datasets" differing by a single record, because the output of a differential privacy algorithm remains nearly identical, the impact of individual records can be concealed. Differential privacy achieves privacy by adding appropriate random noise, such as Laplace noise and Gaussian noise, to statistical queries or model outputs, reducing the direct correlation between the input and output (Dwork, 2006).

The formal definition of differential privacy indicates that a randomized algorithm $M$ satisfies differential privacy, if for any two neighboring datasets and any possible output, the following inequality holds:

$$Pr[M(D) \in S] \leq e^{\epsilon} \cdot Pr[M(D') \in S] + \delta$$

$D$ and $D'$ are neighboring datasets that differ by a single record, such as adding or removing one individual's data. $S$ represents a set of possible outputs of the algorithm $M$. $\epsilon$ is privacy budget that controls the level of privacy. A smaller $\epsilon$ means that the algorithm outputs are less sensitive to changes in the input dataset, providing stronger privacy. Failure probability, $\delta$, is a small non-negative value representing the probability that the privacy guarantee might fail. The relaxation factor $e^{\epsilon}$ describes the allowable ratio of probabilities between the outputs of $M$ when applied to $D$ and $D'$.

This definition ensures that the algorithm's output is nearly indistinguishable when applied to two neighboring datasets. This guarantee is achieved by carefully adding random noise to the algorithm's output, balancing privacy and data utility. There are three widely used mechanisms to add noise: exponential mechanism, Laplace mechanism, and Gaussian mechanism. The exponential mechanism is designed for cases where the output is categorical or discrete, such as selecting the best category or item based on a utility function. The Laplace Mechanism is ideal for numerical outputs under pure $\epsilon$ differential privacy. The Gaussian Mechanism provides flexible guarantees with $\epsilon$ and $\delta$, which allows a small probability $\delta$ of failing the privacy guarantee.

The noise introduced by different mechanisms primarily affects the data distribution, the preservation of feature structure, and the impact on model performance metrics, such as accuracy and F1-score. The noise in the exponential mechanism is introduced probabilistically, and the data distribution remains largely intact. Patterns within discrete features or categorical targets are well-preserved, which allows models to fit these patterns effectively. However, its impact on classification tasks is minimal, so F1-score will not decrease. In scenarios with larger $\epsilon$, model performance is nearly identical to the case without noise. The noise in the Laplace mechanism follows a symmetric Laplace distribution, centered around the original value. For numerical data, the Laplace noise does not significantly alter the overall structure of the data, preserving its distribution characteristics. Since the data distribution is minimally disrupted, the model performance experiences limited degradation, which leaves a particularly small impact on F1-score in classification tasks. However, when $\epsilon$ is small, the model performance will gradually decline (Dwork, 2014). The

noise in the Gaussian mechanism follows a Gaussian distribution with relatively large variance. Gaussian noise exhibits stronger randomness, with wider distribution tails that may introduce extreme values or outliers. When $\epsilon$ is low, the noise intensity is higher, so it will significantly alter the overall data distribution and potentially disrupt the original feature structure. Due to the high randomness of Gaussian noise, the data signal is weakened, leading to a decrease in precision and recall in classification tasks, ultimately resulting in a lower F1-score (Canonne, 2020). Specifically, for adjusting the privacy budget, a larger $\epsilon$, weaker privacy protection, can reduce the impact on model performance, but it compromises privacy guarantees. In contrast, a smaller $\epsilon$, stronger privacy protection, requires optimizing feature processing methods to minimize the noise's effect on critical features (Qin, 2023).

Figure 1 illustrates the relationship between privacy budget and F1-score across three noise mechanisms, highlighting the foundational trade-off between privacy protection and predictive performance in the context of differential privacy. A smaller value of $\epsilon$ corresponds to stronger privacy protection by adding more noise, but this also leads to greater distortion of the original data, thereby reducing model performance (lower F1-score).

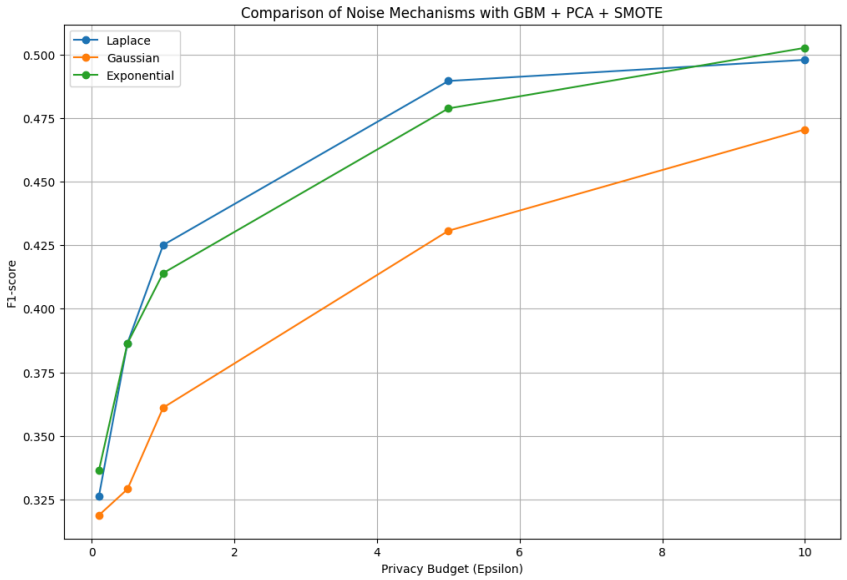


Figure 1: Relationship Between Privacy Budget and F1-score Across Three Noise Mechanisms

## Results

The confusion matrix in Figure 2 suggests that Gradient Boosting performs well in identifying default customers who actually defaulted (True Positive) but also shows the risk of incorrectly classifying non-default consumers as defaulters (False Positive). By experimenting with different combinations of preprocessing and resampling techniques on Gradient Boosting, the combination of PCA and SMOTE oversampling results in the highest

F1-score, as shown in Figure 3. PCA with oversampling methods outperform PCA alone and PCA with undersampling method.

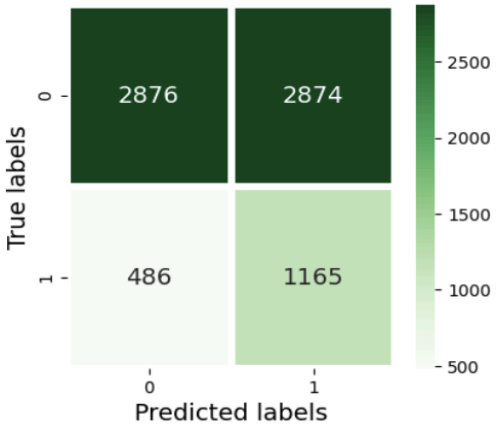


Figure 2: Confusion Matrix for Gradient Boosting Model Predictions on Credit Cred Default Classification

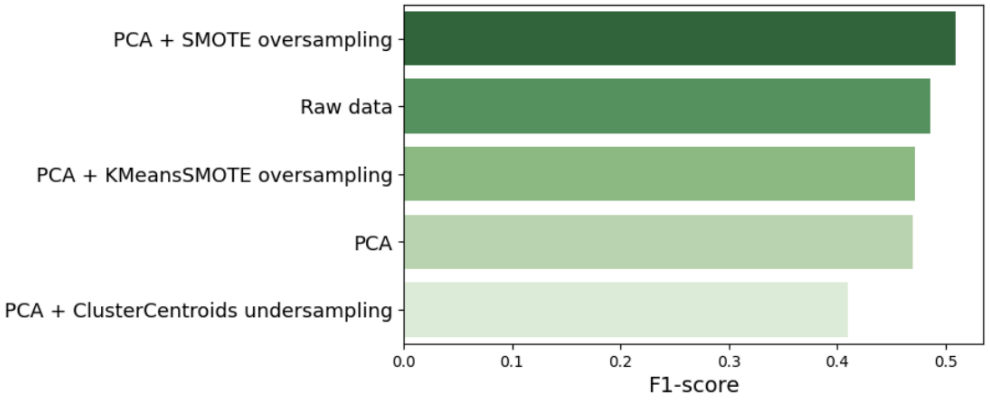


Figure 3: F1-score Comparisons of Gradient Boosting with Different Preprocessing and Resampling Techniques

Gradient Boosting with PCA and SMOTE is adopted as the model used in this study. The performance of the three noise mechanisms, namely Laplace, exponential, and Gaussian, is analyzed and compared using two-way ANOVA, Kolmogorov-Smirnov Test, and Wasserstein Distance. A number of epsilon values and the corresponding F1-scores are recorded for each noise mechanism to perform a two-way ANOVA, which examines the effects of noise mechanism and privacy budget on the model's predictive performance measured by F1-score. The null hypothesis for *Mechanism* is that the mean F1-scores are the same across all noise mechanisms. The null hypothesis for the interaction between *Mechanism* and *Epsilon* is that there is no interaction between noise mechanism and privacy budget. In other words, the effect of privacy budget on F1-score does not depend on the type

of noise mechanism. Table 1 shows that the noise mechanism (p-value = 0.135) and the interaction term (p-value = 0.905) are not statistically significant. This implies that the choice of noise mechanism has minimal independent influence on F1-scores, and its effect does not depend on privacy budget.

| Factor | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| **Mechanism** | 2 | 0.00747 | 0.003734 | 2.525 | 0.134761 |
| **Epsilon** | 1 | 0.05330 | 0.05330 | 36.040 | 0.000202 *** |
| **Mechanism:Epsilon** | 2 | 0.00030 | 0.00015 | 0.101 | 0.904887 |
| **Residuals** | 9 | 0.01331 | 0.00148 | | |

Table 1: Two-Way ANOVA Results for Effects of Noise Mechanism and Privacy Budget on F1-scores

Using Kolmogorov-Smirnov Test, the p-values of zero for all noise mechanisms indicate that the differences between the CDFs of the original data distribution and the noisy distributions generated by Laplace, Exponential, and Gaussian noise are statistically significant. That is, the addition of noise, regardless of the mechanism, significantly alters the original data. This behavior is also evident in Figure 4. However, both two-way ANOVA and Kolmogorov-Smirnov Test do not directly compare the performance of the three noise mechanisms. Wasserstein Distance, on the other hand, measures the cost of transforming one distribution into another. In Table 2, Laplace and Exponential noise have substantially smaller Wasserstein Distances than Gaussian noise, introducing less distortion or transformation to the original data distribution. These two noise mechanisms better preserve the original data and thus ensure that the crucial relationships between features and targets are not significantly disrupted by noise. Thus, Laplace and Exponential noise lead to better predictive performance while maintaining privacy protection.

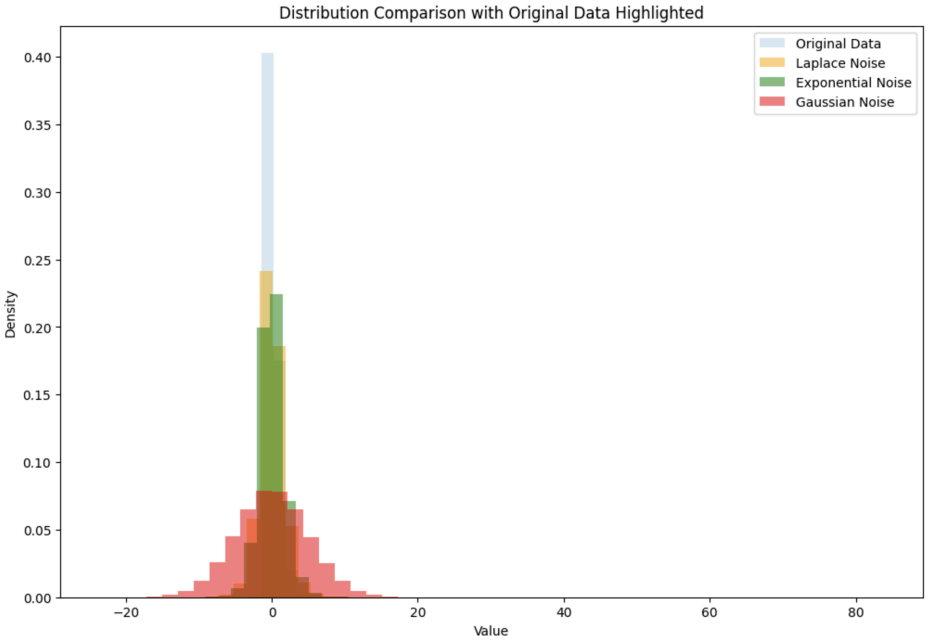


Figure 4: Comparison of Original Data and Noisy Distributions Across Noise Mechanisms

| Noise Mechanism | KS Statistic | p-value | Wasserstein Distance | KL Divergence |
|---|---|---|---|---|
| **Laplace Noise** | 0.1893 | 0.0000 | 0.6051 | 0.2118 |
| **Exponential Noise** | 0.1899 | 0.0000 | 0.5994 | 0.1430 |
| **Gaussian Noise** | 0.3916 | 0.0000 | 3.3629 | 1.0060 |

Table 2: Comparison of Noise Mechanisms Based on KS Statistics, Wasserstein Distance, and KL Divergence

## Further Development

The goal in differential privacy is to balance privacy protection and predictive performance. This study concludes that Laplace and Exponential noise preserve the original data well and support better predictive performance, while Gaussian offers flexibility in privacy guarantees but introduces more randomness (worse predictive performance). One extension of our study is to develop a weighted hybrid mechanism that incorporates the strengths of Laplace, Exponential, and Gaussian noise. The weights can be optimized based on the characteristics of the dataset (categorical/numerical) and the priority given to privacy protection or predictive performance. Another extension is to incorporate additional metrics, such as Jensen-Shannon Divergence (JSD) and Earth Mover's Distance (EMD) to provide a more nuanced comparison of noise mechanisms. JSD is a symmetric and bounded measure of the difference between distributions. EMD captures the minimal effort to transform one

distribution to another. These metrics will complement existing ones used in this study to identify the noise mechanisms that strike the best balance.

## References


Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. *Proceedings of the Seventh* International Conference on Knowledge Discovery and Data Mining (KDD), 107-114.

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer Series in Statistics. ISBN : 978-0-387-95442-4

Li, Q., Wu, Z., Wen, Z., & He, B. (2020). Privacy-Preserving Gradient Boosting Decision Trees. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 784-791.

Dwork, C. (2006). Differential privacy. In International colloquium on automata, languages, and programming (pp. 1-12). Berlin, Heidelberg: Springer Berlin Heidelberg.

Biham, E., & Shamir, A. (1991). Differential cryptanalysis of DES-like cryptosystems. Journal of CRYPTOLOGY, 4, 3-72.

Qin, S., He, J., Fang, C., & Lam, J. (2023). Differential Private Discrete Noise-Adding Mechanism: Conditions, Properties and Optimization. IEEE Transactions on Signal Processing.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211-407.

Canonne, C. L., Kamath, G., & Steinke, T. (2020). The discrete gaussian for differential privacy. Advances in Neural Information Processing Systems, 33, 15676-15688.

Yeh, I. (2009). Default of Credit Card Clients [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H.