ISSN: 2834-7706

Research Article

Journal of Mathematical Techniques and Computational Mathematics

Augmented Differential Privacy Framework for Data Analytics

P. H. Anantha Desik[1*] and Sumiran Naman[2]

[1]*Research & Innovation Head (CEG), TCS, India*

[2]*Research Engineer (CEG), TCS, India*

[*]Corresponding Author
P. H. Anantha Desik, Research & Innovation Head (CEG), TCS, India

Submitted: 2023, Dec 11; **Accepted:** 2023, Dec 27; **Published:** 2024, Jan 03

Citation: Desik, P. H. A., Naman, S. (2024). Augmented Differential Privacy Framework for Data Analytics. *J Math Techniques Comput Math, 3*(1), 01-09.

Abstract

Differential privacy has emerged as a popular privacy framework for providing privacy preserving noisy query answers based on statistical properties of databases. It guarantees that the distribution of noisy query answers changes very little with the addition or deletion of any tuple. Differential enjoys popular reputation that providing privacy without building any assumptions about the data and protecting against attackers who know all but one record. Differential privacy is a relatively new field of research. Most users have a limited experience in managing differential privacy parameters and achieving a suitable level of privacy without affecting the quality of the analysis. A vast majority of users is still learning how to effectively apply differential privacy in practice. In this paper, we discussed: on the proposed augmented framework which enables the differential privacy data of any given query, the various differential privacy techniques, metrics for the privacy & utility tradeoff of the data and efficacy of the framework. Discussed state of the art of different differential privacy techniques defined in the framework Laplace, Laplace bounded, Randomized response and Exponential for different data types. The augmented framework consists of three parts one on privacy parameter inputs to control interactively and iteratively on the querying the data , the various differential privacy techniques, the metrics to measure privacy and utility threshold which allows the data analyst to evaluate the accuracy of the privacy safe data for selecting the privacy guaranteed data within the given privacy budget. The framework takes any dataset as input and, generates another dataset which is structurally and statistically very similar original dataset. The newly generated dataset has much stronger privacy guarantee on the selected sensitive and non-sensitive datatypes. We have also demonstrated analytical models developed using the privacy safe data from the framework as substitute to the models developed on the original datasets. We have demonstrated the framework and analytical model with sample data sets to present the similarity of original and differential privacy safe datasets.

Keywords: Differential Privacy (DP), Epsilon, Privacy Budget, Analytics.

1. Introduction

1.1 Data Privacy

The volume of available data is increasing inexorably, robust, and forward-looking technology is much needed to protect this data from adversaries, who may be equipped with sophisticated attack methodologies and superior computing powers. Securing only the sensitive data fields is no longer adequate for ensuring data privacy. With enhanced computing power, even the non-sensitive data fields can be used to obtain the background knowledge and make educated guesses about the sensitive or private information. On the other hand, there has been a growing adoption of analytics and data science for data driven decision making. While traditional privacy methods, like masking or anonymizing the sensitive and non-sensitive data fields have been useful, they also obstruct the data owner's ability to get valuable insights from analytical models.

Therefore, there is a demand for extracting useful insights from data and analytical models while still ensuring data privacy. This dual requirement of ensuring privacy and utility of data simultaneously has been presenting challenges at the existing data privacy techniques. The objective of privacy preserving data analysis is to release necessary information without compromising privacy of individuals. Differential Privacy (DP), which was introduced by Dwork is an emerging data privacy technique which provides utility data with guaranteed privacy [1]. DP aims to ensure that the output of the algorithm does not significantly depend on any individual's data and guarantee that an adversary will not be able to confidently infer about any information for any individual in the database. The DP is getting much importance with an increasing need of privacy safe datasets to be utility focused. The data analysis performed on the DP enabled datasets provide useful information just like the real dataset. Also, the analytical models developed on these datasets can perform at levels almost similar to the models developed on real data. These privacy safe datasets help in sharing data within or outside business units for analytical model development and these models will have direct applicability as real data models.

The important fact about DP is that it creates privacy enabled data for complex statistical calculations, Machine Learning (ML) algorithms, Artificial Intelligence (AI) algorithms and can be applied to several different functions where the privacy dataset will have similar statistical properties of the original dataset.

1.2 Differential Privacy

DP is not a method, but rather a property that an analysis (or algorithm) may have. Furthermore, DP is not a property of merely a particular output of an analysis, but of the information relationship the analysis creates between its input and its output. In contrast with other privacy approaches which use syntactic property, the DP treats semantic property of the relationship between analysis input and output distributions by restricting the observer to learn about the single contributor [2]. The DP guarantee is a worst-case guarantee as it does not depend on specific input or output, nor does it depend on the specifics of an attempted attack. Differential privacy has a parameter epsilon (ε) which represents the strength of the privacy guarantee. The parameter 'ε' is a non-negative numerical value and can be used as a means of quantifying privacy risks. The differential privacy framework provides tools for reasoning about how 'ε' changes because of multiple uses of differentially private analyses.

The formal definition of differential privacy, which intuitively will guarantee that a randomized algorithm behaves similarly on similar input databases.

Definition 1.1: (Differential Privacy). A randomized algorithm M with domain $N^{|X|}$ is (ε, δ)-differentially private if for all $S \subseteq Range(M)$ and for all $x,y \in N^{|X|}$

$$\text{such that } \left|\left|x - y\right|\right|_1 \leq 1:$$

$$\Pr[M(x) \in S] \leq \exp(\varepsilon) \Pr[M(y) \in S] + \delta$$

If $\delta = 0$, we say that M is ε-differentially private

The DP guarantee is achieved by adding uncertainty to the outcome of an analysis, often in the form of noise perturbation. The perturbation magnitude affects the level of privacy protection, in terms of how similar the outcome distributions are in the presence and absence of an individual's data. As a rule of thumb, achieving a higher level of privacy (that is more similar outcomes in the presence or absence of any single individual's data) requires adding a higher level of noise and hence results in lesser accuracy. The uncertainty needed for DP is less than other types of uncertainties like sampling errors and collection errors. DP techniques protect user privacy while allowing meaningful analysis over the dataset. By adding noise to individual data points, it protects the user's privacy, but on aggregating these data points, the noise is averaged out, obtaining a result closer to the original one.

There are various differential techniques like Laplace, Bounded Laplace, Randomized response technique and Exponential methods. These differential techniques can be controlled on the 'ε' and sensitivity ('s') parameter to achieve right tradeoff of privacy and utility of the data [3]. Even though there are different methods available for DP enabled data for synthetic data and for differential privacy data, there is a need for having a single framework where a data analyst can iteratively execute the query to get the privacy enabled data for different data types within the privacy budget [4]. To achieve this, the data analyst must use multiple queries, selecting different techniques and different metrics for measuring the utility and privacy trade off to arrive on the privacy enabled data.

If the organization or data analyst get different types of data, it becomes difficult to decide the technique, parameter, and metrics that must be used to measure the privacy enabled data and the costs may overrun the privacy budget. The proposed augmented framework consists of different DP techniques, easy parameter selection and metrics to create the right data for a set of different data types in iterative and interactive process within the privacy budget. The augmented framework provides Laplace, Laplace bounded, Randomized response and Exponential techniques which can be used by the data analyst based on data properties. Any new differential privacy techniques or traditional techniques can also be easily integrated into the framework. The augmented framework and techniques are discussed in the next section.

1.3 Laplace Mechanism

The technique is the go-to function of differential privacy having widespread applications on numerical data. The strength of the technique lies in its computational and mathematical simplicity. The technique simply computes function (f) and perturbs each data point with noise drawn from the Laplace distribution. The scale of the noise will be controlled by using the privacy parameter 'ε'.

$$Lap(x|b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$$

The variance of this distribution is $\sigma^2 = 2b^2$.

Definition 1.2 (The Laplace Mechanism). Given any function $f: N^{|X|} \rightarrow R^k$, the Laplace mechanism is defined as:

$$ML(x,f(\bullet),\varepsilon) = f(x) + (Y1,...,Yk)$$
$$\text{where } Y_i \text{ are i.i.d.random variables drawn from Lap } \left(\frac{\Delta f}{\varepsilon}\right)$$

Where i.i.d. stands for independent and identical distribution.[8]

1.4 Bounded Laplace

Laplace distribution is infinite, and it is common for the output of the Laplace mechanism to fall outside the range of the query. The bounded approach is to support the response mechanism, and then sample directly from the output domain (that is by inverse transform sampling). This can also be achieved through rejection sampling, by continually redrawing from the unbounded distribution until the output falls within the domain. This whole process is called bounding, as pure outputs of the mechanism are bounded by design [5].

Definition 1.3 (Bounded Laplace Mechanism). Given b > 0 and D ⊂ R, the bounded Laplace mechanism W_q: Ω → D, for each q ∈ D, is given by its probability density function fW_q):

$$f_{W_q}(x) = \begin{cases} 0, & if\ x \notin D \\ \frac{1}{C_q}\frac{1}{2b}e^{-\frac{|x-q|}{b}}, & if\ x \in D \end{cases}$$

$$where\ C_q = \int_D \frac{1}{2b}e^{-\frac{|x-q|}{b}}dx\ \ is\ a\ normalization\ constant$$

Random Response: Definition 1.4: The standard notation is Random response is (Ω,*F*,*P*) denotes a probability space. X_i **:** Ω → {0,1} is then a random variable for each *i* **:** [*n*], dependent on the truthful value x_i. Reference [6].
We define the randomized response mechanism by

$$P(X_i = k \mid x_i = j) = p_{jk}$$

Consider an activity X. Simulate the probability distribution space of X using the epsilon as bias. From the distribution space depending upon the truthful value the answer will be truthful or complementary of true response.

1.5 Exponential Technique

This technique can be considered a natural building block for answering queries with arbitrary utilities (may be within the arbitrary non-numeric range) while still preserving differential privacy. This employs a utility function which maps the database/output pairs to their utility scores and when invoked, it tries to provide the best pair with highest utility score as the output depending on the epsilon. Formal definition of exponential mechanism is as follows:
Utility function u: N |X| × R → R, which maps database/output pairs to utility scores

Definition 1.5 (The Exponential Mechanism): The exponential mechanism M_E (*x*,*u*,*R*)selects and outputs an element r ∈ R with probability proportional to exp $\left(\frac{\varepsilon u(x,r)}{2\Delta u}\right)$.

the privacy loss is approximately:

$$\ln\left(\frac{\exp\left(\frac{\varepsilon u(x,r)}{\Delta u}\right)}{\exp\left(\frac{\varepsilon u(y,r)}{\Delta u}\right)}\right) = \varepsilon[u(x,r) - u(y,r)]/\Delta u) \leq \varepsilon.$$

1.6 Sensitivity Parameter (s)

It is a function that reflects the amount of function's output that will change when its input changes. The amount of noise necessary to ensure DP for a given query depends on the sensitivity of the query. The framework supports sensitivity parameter as '1' and 'min-max' sensitivity. DP local means noise is added to individual data before it is centralized in a database. DP global means noise is added to raw data after it is collected from several individuals. In this framework, we have applied global DP where noise is added to original data items when querying the database.

1.7 Privacy Budget

The parameter is denoted as epsilon (ε). 'ε' controls how much noise or randomness is added to the raw dataset. The differential privacy algorithms are based on the parameter 'ε', which controls the trade-off between privacy and utility of data. A high value of 'ε' means more accurate but less private data. It is the maximum distance between a query on database (x) and the same query on database (y). It is also called metric of privacy loss at a differential change in data (that is adding or removing 1 entry). Privacy budget is a cumulative sum of epsilon (ε) for each database query. Privacy budget within control means the number of queries executed on database are minimum so that there will be no exposure of the data.

2. The Augmented Framework

Currently, there are several ways to implement DP with various methods and parameters. The proposed framework will follow augmented model architecture where a data analyst can use the system iteratively and interactively to create privacy data for a given query by using different DP techniques (Laplace, Bounded Laplace, Randomized Response and Exponential) with privacy and utility guarantee. We have various DP techniques like Laplace, Bounded Laplace for numerical data, Random response for binary and Exponential technique for binary/categorical data which are essential in protecting the privacy of sensitive and non-sensitive data. The framework also consists of some traditional masking techniques particularly for text data which is not part of the discussion.

The proposed framework aims to facilitate any given user scenario, based on the DP techniques, sensitivity('s'), 'ε' value and data type of variables, to provide privacy safe data with the defined metrics and measure the utility and privacy tradeoff. The architecture allows implementing iterative and interactive selection of DP safe data within utility and privacy tradeoff based on the metrics. The framework consists different privacy metrics like privacy match, privacy DigiMatch and privacy error to demonstrate the strength of the privacy. Different utility metrics like mean, standard deviation, correlation charts, histograms, and kernel density graphs (KDE) have been used for illustrating data similarity.

In this framework, the data analyst submits queries iteratively in an interactive way, based on the observed metrics of the previous queries, and considering the privacy and utility of the data submits in the next iteration. This framework addresses issues of correct technique selection, parameters, and metrics for

utility and privacy trade off criteria of the privacy enabled data and controls the privacy budget.

Algorithm: Iterative query

1. Start
2. Profile the data to know the data types
3. Input Privacy budget, Initial 'ε', sensitivity (s)=1
4. Select DP techniques for the data types and build the query
5. Execute the Query on the data Q(Di)
6. Measure privacy, Utility Metrics
7. If Privacy and Utility Metrics within Tolerance level then Go to Step 11
8. Else Privacy budget ← Privacy budget – epsilon ('ε')
9. If Privacy budget is exhausted, then Go to Step 12
10. Else change the epsilon ('ε') and go to Step 4
11. Stop, consider the data for Privacy
12. Stop

The Augmented framework Architecture:

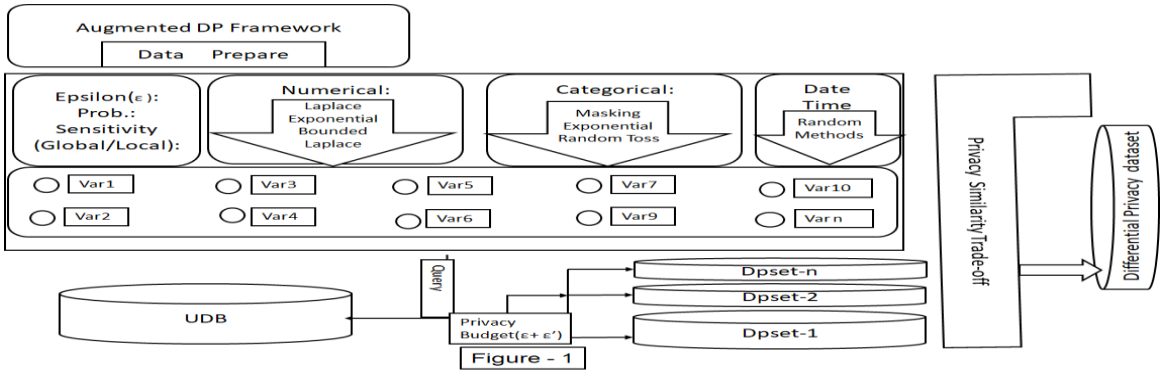

Figure - 1

2.1 Privacy & Utility Metrics

The framework consists of different types of metrics for measuring privacy and utility tradeoff of the data. Three types of privacy measures are defined to enable privacy guarantee for the data. The techniques offer more diligent understanding of the privacy enablement of the data. The techniques are Privacy Match, Privacy DigiMatch and Privacy Error. The framework supports utility metrics like average, and standard deviation. It also supports visualization graphs like Histogram, KDE, Violin plot, Box plot and Pie charts to check the statistical properties of original data and privacy safe data.

2.2 Privacy Match

It is the privacy percentage based on number of matches, the metric count number of matches of original column and DP columns. If 'm' is the number of matches and column count is 'n' then privacy percentage is (1-(m/n))*100. If the percentage is higher, then it means that there is more variation in the private column data.

2.3 Privacy DigiMatch

Privacy percentage based on number of digits mismatch. It will consider the pattern of the digits and find the percentage of difference. The metrics consider original value and DP column value, whichever digits are small it compares with other number each digit position. It takes 'k' as number of digits does not match and calculate percentage as (k*100/m) for all the values for all the 'm' digits. The overall privacy percentage is average percentage of numbers of rows. If the privacy percentage is high, then it means the digit patterns of the columns variate largely.

2.4 Privacy Error

It is the privacy percentage based on distance difference between original column and DP column. The data elements of the columns are subtracted and divided by the original element. This indicates how much the private data elements differ from the original data elements. If the privacy percentage is high, then it means that difference exists between original and DP column.

3. Results & Discussion

The proposed augmented framework was applied on two dissimilar data sets with different selection of differential privacy techniques and parameters. Table 1.1 shows original dataset of some insurance business data and table-1.2 is the DP safe data queried from the framework on the original data. The analytical model logistic regression was applied on the dataset1 for original and DP safe data. The model derived from the original data and privacy safe data had similar performance. The Bounded Laplace technique was applied on numerical datatype variables 'Annual_Premium', 'Age' and 'Sum_Assured' and Exponential technique was applied for categorical data variables 'purchase_reason' and 'employment'. Table-1.3 shows the privacy metrics of the data of different queries for privacy parameter 'ε'=0.75, 'ε'=0.25, 'ε'=0.05 and 'ε'=0.005. The metrics for 'ε'=0.005 show a strong agreement between privacy and utility tradeoff of the original and DP safe data within privacy budget of 1.5. Fig.2, fig.3 and fig. 4 provide the density curves for 'Annual_Premium', 'Sum_Assured and violin curve for 'Age' variables which demonstrates that a strong statistical similarity exists between the datasets. Fig. 6 which shows correlation charts of original and privacy datasets, also demonstrates statistical similarity between datasets. The logistic regression model is trained on the DP data applied on original data to test the DP model accuracy. Table 1.4 shows the accuracy of the logistic regression on original data model, privacy safe data model and privacy data model tested on the original data. The accuracy of the original data model and privacy data models was very similar. The performance of the privacy data model which was

tested on original data is of same accuracy of the original data model. The gain charts of the models in Fig. 5 demonstrate the same on the accuracy of the models. This demonstrated that the model built on DP safe data using this framework can be successfully used as a model on the original data. Dataset 2 is from banking domain to test how privacy safe clustering data model will perform compared with the model on original data. The query constituted of hybrid techniques like Laplace, Exponential and traditional masking. Traditional masking was applied on variable 'city-name' because the DP techniques cannot hide the text data and masking is very important for non-disclosure of text variables. Tables 2.1 and 2.2 provide the original dataset and respective DP safe dataset. The queries using different parameters of 'ε'=0.25, 'ε'=0.05 with sensitivity 's'=1 applied on the dataset. Fig. 7 shows that the similarity of variable 'Income' for 'ε'=0.05 is approximately same as the original column data. The clustering model was applied on the original and privacy safe data for three cluster solutions. Fig. 8 demonstrates that models based on original data and privacy safe data provided similar cluster solution and table 2.3 shows that even cluster percentages are more or less same for both the datasets. This demonstrates that proposed augmented hybrid framework enabled for easy selection of DP data within the privacy budget and the models developed on privacy safe data can substitute the model on the original data. In both the cases of datasets the augmented framework demonstrated that the optimum privacy and utility tradeoff DP safe data can be queried within the privacy budget. The analytical model developed on privacy safe data can be used as proxy for the original model. The augmented framework will help in providing privacy guaranteed data for model development as an alternative to original data models within the privacy budgets.

4. Conclusion

Differential Privacy may become the de facto standard to ensure privacy, as it is fast becoming the most trending research topic in privacy-enhancing computation and used in a wide range of analytical applications.

The Augmented framework presented in this paper consists of three parts:

· The first part focuses on the Differential Privacy techniques - Laplace, Bounded Laplace, Random Response, and Exponential Technique.

· The second part describes usage of parameters viz. privacy parameter 'ε', sensitivity ('s') and privacy budget.

· The third part discusses the iterative and interactive processing of queries and the metrics for optimum privacy & utility tradeoff enabling the selection of differential privacy data.

The experiment, results, and the discussion thereafter of using the Augmented framework demonstrates that it is easy to select and test queries to get the right privacy safe data using the suggested privacy metrics within the budget. The comparison of analytical models developed using DP safe data and the original datasets affirms that the Augmented framework helps the data analyst to a) select the right privacy and utility tradeoff data and, b) the DP data models that will be the right replacement for the original models. The proposed Augmented framework can be seamlessly configured for any new DP techniques and new parameters. In the last two years, new techniques are emerging in numerical, text based Differential Privacy. Research is underway in Differential Privacy inclusive of AI models that may lead to better privacy safe data analytics.

sl_no	Sales_Channel	Gender	Product_Group	Product_Type	Decision	Purchase_Reason	Age_at_POS	Manual_Decision	Employment	Sum_Assured	Annual_Premium	Evidence
1	BP	F	Life	JL	AST	MP	55	Y	Employed	65038	600	Y
2	IFA	F	Life	SL	AST	BP	37	Y	Employed	300000	442.8	Y
3	IFA	M	Life	SL	AST	FP	42	N	Employed	200000	210	N
4	IFA	M	Life	JL	AST	FP	32	Y	Employed	130000	180.36	Y
5	BP	M	Life	SL	AST	MP	43	N	Employed	157500	217.92	N
6	IFA	F	CIC	JL	AST	MP	36	N	Employed	80000	360	Y
7	BP	M	Life	JL	AST	MP	31	N	Employed	127995	196.44	N
8	IFA	M	CIC	SL	AST	MP	24	N	Employed	90000	269.16	N
9	IFA	M	Life	SL	AST	FP	39	N	Employed	45000	56.16	N
10	BP	M	Life	SL	AST	MP	47	N	Employed	124365	703.08	N

Table 1.1 Original Dataset1

sl_no_d	Sales_Channel_d	Gender_d	Product_Group_d	Product_Type_d	Decision_d	Purchase_Reason_d	Age_at_POS_d	Manual_Decision_d	Employment_d	Sum_Assured_d	Annual_Premium_d	Evidence_d
1	BP	F	CIC	JL	AST	MP	41	Y	Unemployed	65938	1047	Y
2	IFA	F	Life	SL	ANST	BP	52	Y	Employed	300034	479	Y
3	IFA	M	Life	SL	ANST	FP	56	N	Unemployed	199540	42	N
4	BP	F	CIC	SL	AST	FP	29	Y	Employed	130485	128	Y
5	IFA	M	Life	SL	AST	FP	49	Y	Contractworker	157376	208	N
6	IFA	F	CIC	JL	ANST	MP	49	N	Employed	80654	548	N
7	BP	F	Life	JL	AST	MP	27	N	Unemployed	127910	435	N
8	IFA	M	CIC	SL	ANST	BP	74	N	Employed	90117	47	Y
9	BP	M	Life	SL	AST	FP	64	N	Unemployed	44974	93	N
10	IFA	M	Life	JL	AST	MP	60	N	Retired	124338	462	N

Table1.2 Differential Privacy data

	Data type- Numerical -Laplace												Categorical _Exponential	
	Var. Annual_Premium original data mean =373				Var. Age original data mean = 43				Var. Sum_Assured Original data mean				purchase_reason	employment
epsilon	pure_match (%)	privacy-digit (%)	privacy_error (%)	mean	pure_match (%)	privacy-digit (%)	privacy_error (%)	mean	pure_match (%)	privacy-digit (%)	privacy_error (%)	mean	mismatch%	
0.075	99.98	99	6.58	373	95.48	95	27.55	44	96.35	95	0.01	140770	67.8	82.87
0.05	100	99	9.59	374	96.99	96	36.01	45	97.45	96	0.02	140770	67.12	85.61
0.025	100	99	18.74	377	97.86	97	46.87	45	98.86	97	0.05	140769	66.4	81.5
0.005	100	99	83.4	450	98.43	98	53.09	46	99.76	98	0.25	140769	68.49	85.23

Table 1.3

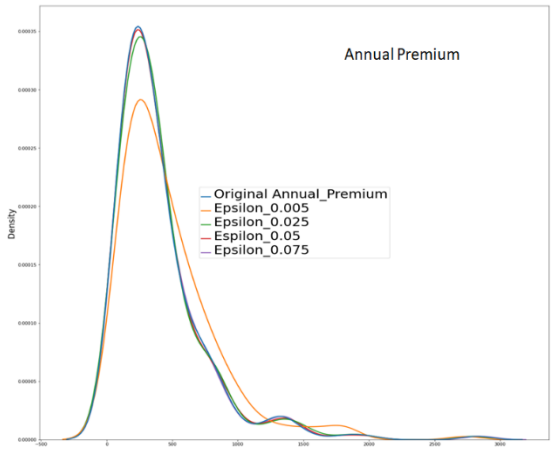

Figure-2

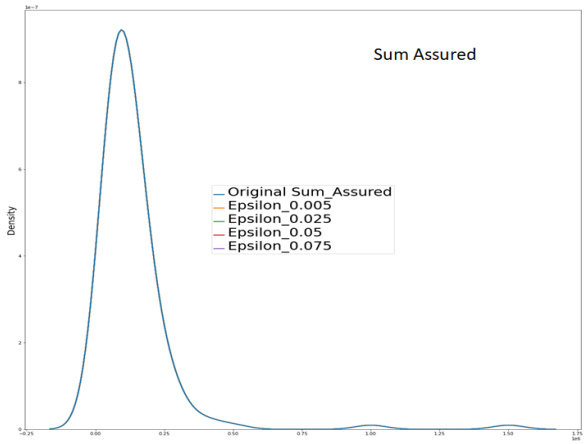

Figure-3

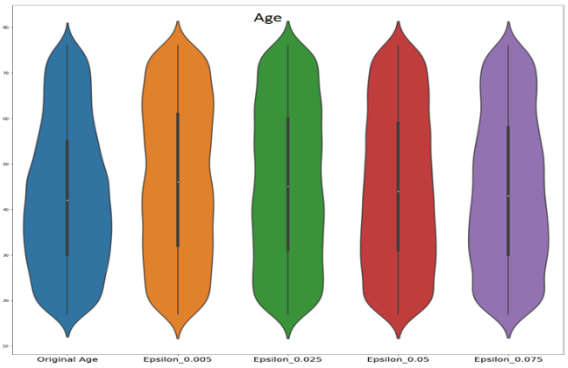

Figure -4

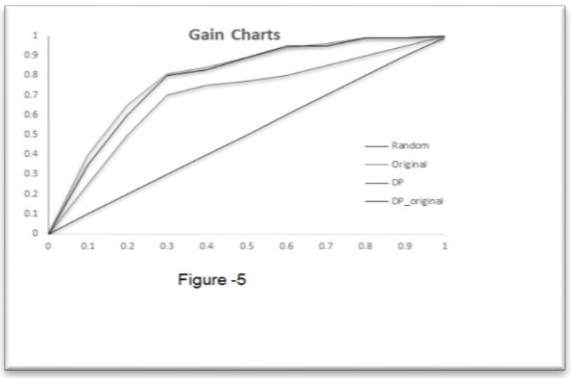

Figure -5

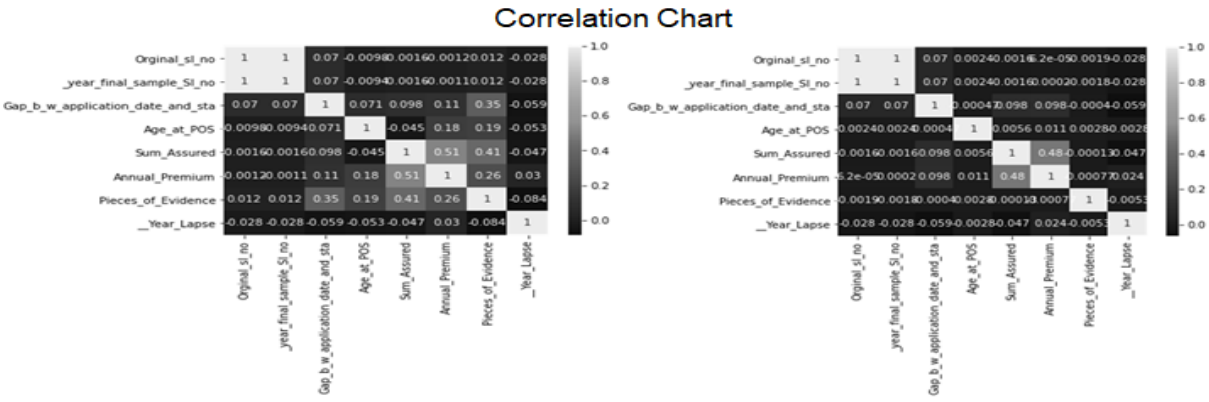

Figure -6

	Model metrics & accuracy of test data1 of table2			
		Precision	Recall	Accuracy
Original Data	0	0.80	0.83	**0.78**
	1	0.77	0.73	
DP Data model	0	0.72	0.72	0.69
	1	0.66	0.65	
DP model on original data	0	0.81	0.82	**0.78**
	1	0.76	0.75	

Table1.4

Sl.No	Loan_ID	Income	Lterm	City_name	lati	long
0	LP001002	5842	155	Delhi	28.66	77.23
1	LP001003	4583	266	Mahārāshtra	18.9667	72.8333
2	LP001005	3019	98	West Bengal	22.5411	88.3378
3	LP001006	2604	152	Karnātaka	12.9699	77.598
4	LP001008	5982	179	Tamil Nādu	13.0825	80.275
5	LP001011	5364	334	Telangana	17.3667	78.4667
6	LP001013	2311	86	Mahārāshtra	18.5196	73.8553
7	LP001014	3002	146	Gujarāt	23.03	72.58
8	LP001018	4000	200	Gujarāt	21.17	72.83
9	LP001020	12853	111	Uttar Prades	26.847	80.947
10	LP001024	3186	29	Rājasthān	26.9167	75.8667

Table 2.1 Banking Loan dataset2

Original data					**Differential Privacy data**			
Age	**L_Mon**	**A_Bal**	**L_Amt**		**Age_DP**	**L_Mon_DP**	**A_Bal_DP**	**L_Amt_DP**
52	40	607.081	65020		32	330	691.466	65247
31	20	415.038	300005		25	83	550.603	301232
74	25	181.76	199999		23	173	461.594	199904
32	11	128.355	130013		62	24	306.318	129742
62	41	195.669	157528		22	148	171.131	157436
47	20	398.273	80005		43	182	112.395	80131
40	18	145.129	127988		52	308	90.358	128246
23	9	264.45	90034		65	54	182.133	90043
24	7	44.791	45018		74	96	175.041	45121
75	12	718.057	124380		41	86	1003.924	125029
48	10	79.084	37992		42	130	175.492	37978
67	24	119.634	57255		58	145	176.766	56733
50	56	754.611	187986		30	95	346.145	187919
L_Mon -> Tenure of loan pending					**L_Amt** -> Loan Amount			

Figure 2.2: DP Privacy Dataset2

Sl	Loan_ID	DP-Income	DP-Lterm	ty_name_	DP-lati	DP-long
0	LP001002	5859	360	SCCYVXXA	25.153	76.726
1	LP001003	4586	180	UHPZVUGI	12.154	92.909
2	LP001005	2986	240	QXFTWQU	23.213	93.236
3	LP001006	2575	60	OTDQKDO	13.556	93.451
4	LP001008	6014	360	HPXPYZKIF	8.294	90.598
5	LP001011	5402	360	GHRDGYG	12.868	81.22
6	LP001013	2340	360	FOSTTGDI	18.773	73.278
7	LP001014	3032	12	OIRRQWQ	22.656	77.521
8	LP001018	3996	360	INAPNBER	29.219	82.567
9	LP001020	12856	360	JIIYLGVEI	27.831	77.359
10	LP001024	3201	240	FUJQJVJJJ	17.869	93.689

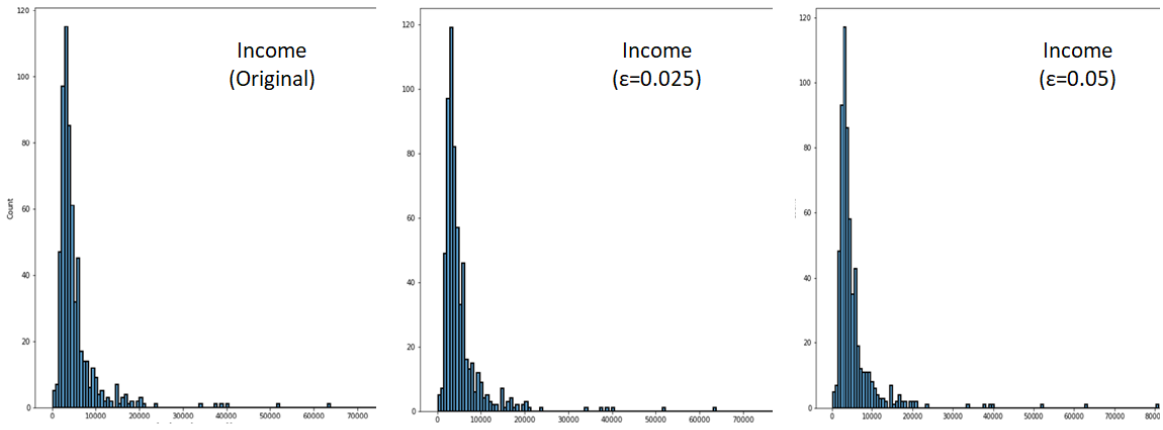

Figure-7

F

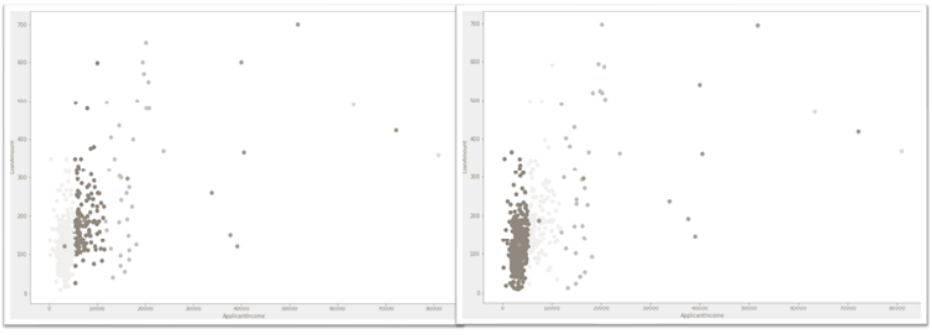

Figure-8

Cluster Number	Original Data Count	DP Data Count
0	430 (70%)	435 (71%)
1	144 (23%)	135 (2%)
2	40 (6%)	44 (7%)

Table 2.3: Cluster Results

References

1. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211-407.
2. Dr. Hector Page, Privitar Charlie Cabot, Privitar Professor Kobbi Nissim, Differential privacy: an introduction for statistical agencies, , Georgetown University, December2018.
3. Alvim, M., Chatzikokolakis, K., Palamidessi, C., & Pazii, A. (2018, July). Local differential privacy on metric spaces: optimizing the trade-off with utility. In 2018 IEEE 31st Computer Security Foundations Symposium (CSF) (pp. 262-267). IEEE.
4. Ping, H., Stoyanovich, J., & Howe, B. (2017, June). Datasynthesizer: Privacy-preserving synthetic datasets. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management (pp. 1-5).
5. Holohan, N., Antonatos, S., Braghin, S., & Mac Aonghusa, P. (2018). The bounded laplace mechanism in differential privacy. arXiv preprint arXiv:1808.10410.
6. Holohan, N., Leith, D. J., & Mason, O. (2017). Optimal differentially private mechanisms for randomised response. IEEE Transactions on Information Forensics and Security, 12(11), 2726-2735.
7. Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., & Talwar, K. (2007, June). Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 273-282).
8. Hay, M., Rastogi, V., Miklau, G., & Suciu, D. (2009). Boosting the accuracy of differentially-private histograms through consistency. arXiv preprint arXiv:0904.0942.
9. Bhaskara, A., Dadush, D., Krishnaswamy, R., & Talwar, K. (2012, May). Unconditional differentially private mechanisms for linear queries. In Proceedings of the forty-fourth annual ACM symposium on Theory of computing (pp. 1269-1284).
10. Zhang, J., Zhang, Z., Xiao, X., Yang, Y., & Winslett, M. (2012). Functional mechanism: Regression analysis under differential privacy. arXiv preprint arXiv:1208.0219.
11. Nissim, K., Orlandi, C., & Smorodinsky, R. (2012, June). Privacy-aware mechanism design. In Proceedings of the 13th ACM conference on electronic commerce (pp. 774-789).
12. Vaidya, J., Shafiq, B., Basu, A., & Hong, Y. (2013, November). Differentially private naive bayes classification. In 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) (Vol. 1, pp. 571-576). IEEE.
13. Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2013, October). Local privacy and statistical minimax rates. In 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (pp. 429-438). IEEE.
14. Holohan, N., Leith, D. J., & Mason, O. (2015). Differential privacy in metric spaces: Numerical, categorical and functional data under the one roof. Information Sciences, 305, 256-268.
15. Liu, F. (2016). Statistical properties of sanitized results from differentially private laplace mechanism with univariate bounding constraints. arXiv preprint arXiv:1607.08554.
16. Smith, A., Thakurta, A., & Upadhyay, J. (2017, May). Is interaction necessary for distributed private learning?. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 58-77). IEEE.
17. Liu, B., Zhou, W., Zhu, T., Gao, L., & Xiang, Y. (2018). Location privacy and its applications: A systematic study. IEEE access, 6, 17606-17624.
18. Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., & Wang, T. (2018, May). Privacy at scale: Local differential privacy in practice. In Proceedings of the 2018 International Conference on Management of Data (pp. 1655-1658).
19. Kalantari, K., Sankar, L., & Sarwate, A. D. (2018). Robust privacy-utility tradeoffs under differential privacy and hamming distortion. IEEE Transactions on Information Forensics and Security, 13(11), 2816-2830.
20. Wagner, I., & Eckhoff, D. (2018). Technical privacy metrics: a systematic survey. ACM Computing Surveys (CSUR), 51(3), 1-38.
21. Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. Journal of the American Statistical Association, 113(521), 182-201.
22. Yang, X., Gao, L., Zheng, J., & Wei, W. (2020). Location privacy preservation mechanism for location-based service with incomplete location data. IEEE Access, 8, 95843-95854.