Privacy Preserving ML: Using CTGAN Algorithm to Generate Synthetic Data and Analyze the Performance Compared to Using Raw Data

Hao Zhang
hzhang62@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Junyan Li
jli56@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Zhuolin Li
zli82@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Abstract

Personal data can be misused in a numbers of way if people cannot have the ability to control and protect it. Company may or unintentionally leak or sell user's personal data to advertiser without user consent[1]. CTgan algorithm which is a powerful tool for protecting user's personal data was introduced in the *Generating Synthetic Tabular Data*[8] paper. However, the proof of correctness in in this article is not sufficient. In this paper, we show CTgan algorithm perform pretty vell on generating synthetic data and also get a fair performance after applying machine learning classification algorithms base on our own experiment.

Keywords: CTGan Algorithm, privacy, synthetic data, machine learning

ACM Reference Format:
Hao Zhang, Junyan Li, and Zhuolin Li. 2018. Privacy Preserving ML: Using CTGAN Algorithm to Generate Synthetic Data and Analyze the Performance Compared to Using Raw Data. In ,. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

In the digital age, data analysis is used in almost every field, and data analysis is based on the collection of private data. Therefore, privacy has gradually become a major and sensitive topic that cannot be taken lightly. Out of the need for personal privacy protection, there are many technologies, such as K-anonymity, t-closeness, l-diversity, and so on. Most of them achieve privacy by making some changes to the data. Corresponding problems also arise, the changes we make when we process the data to a certain extent, it may affect the accuracy of the results. In other words, it is difficult to make this trade-off between privacy and utility. In this project, we are going to use GAN to generate synthetic data, which will address the trade-off properly.

In the *Generating Synthetic Tabular Data*[8] paper, it introduced CTGAN library which have a powerful tool to generate fake tabular data. But the proof of correctness in this paper only used one dataset and the statistical method used in this paper is not sufficient. We believe CTGAN library is a powerful tool. We want to prove this with our own experiment and the concept learned in this course to prove or disprove the performance of this library.

https://doi.org/10.1145/1122445.1122456

2 Related Work

In *Generating Synthetic Tabular Data*[8] paper, generative adversarial network (GAN) is the method being used to generate tabular data. This paper will be our main source to process our research. Gan technique is a powerful method to generate tabular data that contains the similar distribution as the real data. This allows users to share and publish dataset without data loss and maintain data privacy. The algorithm in this paper also has a high adaptability which supports multiple data types which include numerical, categorical, time and text data. It also recognizes different types of distributions(multi-modal, long tail, non-gaussian). With the help of this algorithm. Users will feel more comfortable and convenient to data sharing, experimenting, and analysis on a large scale, without disclosing sensitive information. From the statistics provided by the paper, the tabular data generated by the algorithm have a good similarity with raw data. But the experience only worked on one single dataset, the performance of this algorithm under different data sizes, more various statistical methods are mentioned at the end of this article but never provided. The stability of this algorithm is still pending to prove or disprove.

Generating Synthetic Tabular Data[8] paper comes from a series of synthetic data paper by Lulu Tan. This is the second article that concentrate on generating synthetic data. In this paper, author works on a data set to generate fake data with CTGAN library. Then author used the TableEvaluator library to produce statistical plots for his result which includes distribution of columns, Correlation Matrix between Real and Synthetic Data and Absolute Log Mean and STDs of Real and Synthetic Data. Author have not discuss about the results

from these plots and in the conclusion section it mentioned "In the next article (last of the series), we will discuss some limitations, challenges and the future of synthetic data." [8] But from our search we can not find the third article which should concentrate on discussing limitation of the CTGAN library and proof of correctness and performance. We assume author will not upload the final piece of this series paper. Thus, we decide to proof it with our own experiment and statistical method.

3 General Plan of Work

3.1 Workload Distribution and Finished Date

Here are the roles and responsibilities of each team member and Finished Date:

- 9/17/2021 Finish project proposal(Hao Zhang, Junyan Li, Zhuolin Li)
- 9/25/2021 Find the appropriate dataset (Zhuolin Li)
- 10/09/2021 Finish code review and algorithm analysis (Hao Zhang)
- 10/13/2021 Progress discussion with instructor(Hao Zhang, Junyan Li, Zhuolin Li)
- 10/23/2021 Finish code development(Hao Zhang, Junyan Li, Zhuolin Li)
- 10/30 Find appropriate statistical methods to analyse the data(Junyan Li)
- 11/16 Project final presentation(Hao Zhang, Junyan Li, Zhuolin Li)
- 11/20 Finish final report(Hao Zhang, Junyan Li, Zhuolin Li)

3.2 Meeting Schedules

We decide to have meetings every Friday to report weekly progress and update final report:

1 - Sept 17th: all attended
2 - Sept 24th: all attended
3 - Oct 1st: all attended
4 - Oct 8th: all attended
5 - Oct 15th: all attended
6 - Oct 29th: all attended
7 - Nov 5th: all attended
8 - Nov 12th: all attended
9 - Nov 19th: all attended

4 Approach

Our proposed approach is to use GAN machine learning algorithms to generate synthetic data and analyze the performance compared to using the raw data. Our first task will be finding appropriate datasets to test the algorithm. Second, we will dive into the actual code of CTGAN library[3] to analyze the correctness of the algorithm. Our third task will be generating the synthetic data from our datasets. Lastly, we will look for appropriate statistical methods to analyse the mock data and write a report about the correctness, performance and stability of this algorithm.

4.1 Synthetic Data

Synthetic data is data that is artificially derived from real data, and can capture the similar structure and similar statistical distribution as the raw data, making it indistinguishable from real data. There are some methods to protect data privacy (such as the k-anonymity, l-diversity, t-closeness). However, it might involve the omission of data records to some extent. This leads to an overall loss of information and the utility. In this case, synthetic data is a good alternative solution for data anonymization. The advantage of synthetic data is that it can protect data privacy. Real data contains sensitive and private user information that cannot be shared freely and is subject to legal restrictions. Synthetic data sets can be published, shared, and analyzed more publicly without revealing actual personal information.

4.2 GAN Algorithm

Generative adversarial network(GAN) is a deep-learning based generative model. It consists of two parts: (a)A generator that focuses on generating new fake data from the original dataset. (b)A discriminator that classifies these data as real data and fake data. The generator and discriminator compete against each other based on game theory in order to train the generator to become more powerful in generating new fake but plausible data and discriminator become more powerful in distinguishing whether the data from the generator is fake or not. In order to get high quality tabular data, we plan to use the conditional generative adversarial networks(CTGAN). This model is developed by Xu et al. of MIT[9].

4.3 Machine Learning

We divide the two datasets into training set and test set respectively. The training set of the two datasets accounts for 75% of each total sample respectively, and the test set accounts for 25% of each total sample.

- `Logistic Regression`: Logistic regression is a classical classifier of supervised learning, which is often used in data mining, diseases diagnosis and economic prediction. The output of logistic regression can predict the probability of a class. The default threshold of logistic regression is 0.5.
- `Naïve Bayes`:Naïve Bayes classification usually adopts the strategy of content-based filtering technique[10]. This method analyses words, the occurrence, distributions of words and phrases in the content of emails and then use generated rules to filter all the incoming spam emails. It can be further illustrated as an approach which is based on a statistical machine learning process which has the properties of an independence

which is strong and equally can handle a large number of datasets. In the concept of Naïve Bayes, the distribution of a probability is usually assessed from the rate of distribution of the dataset.

- `Decision Tree`:Decision Tree is a classic supervised machine learning algorithm[5]. And decision tree algorithm could be both used for regression and classification problem. Decision tree could build a model which trained by simple decision rules from training data that could be used to predict the class or label of the target value.

5 Dataset

All of the dataset come from Kaggle, and all of our datasets are real world datasets. These datasets have different attributes and distributions. The following is an introduction to the two datasets.

5.1 Adult Income Dataset

This dataset is a real-world dataset extracted from the 1994 Census database by Barry Becker[6]. It consists of quasi-identifier data(age and education) and sensitive data(income). And the prediction task is to determine whether a specific person's salary is above 50K per year.

5.2 Bank Term Deposit Dataset

This dataset is a popular dataset from Kaggle for machine learning classification problem[4]. There are 4521 records in this dataset. It also consists of quasi-identifier data(age and job) and sensitive data(balance and loan). And the prediction task is to determine whether customer will subscribe for a term deposit in a bank institution.

6 Software and Tools

- `Coding Environment`: Google Colab.
- `Coding Language`: Python 3
- `Libraries`: Pandas, Numpy, Sklearn, CTGAN,etc.

7 Result

As figure 1, and figure 2 shown, The synthetic data and the raw data are very similar, but synthetic data does protect the user's privacy. The comparison of all features' distribution and cumulative sums graphs are listed in Figure 3. These graphs provide better visualization for the difference between raw data and synthetic data. We could see that the distribution for some feathers are similar. For the experiment purpose, we generated graph for all features.

8 Evaluation

We will use "TableEvaluator"[7] (a python library that evaluates the similarity between a synthesized dataset and the raw data.

8.1 2 sample z test

We did the 2 sample z test on the mean[2] from the original data and synthetic data with Bank Term Deposit data set. We choose the Duration column as our experiment subject because it is a numerical column which allow us to perform analysis method.

Before doing the test we checked the assumption of 2 sample Z test:

1. The two data set have large sample size. From definition the sample size > 30 could consider as large. we are using the sample size of 4521 which satisfy the requirement.

2. Two data set have normal bell shape distribution.

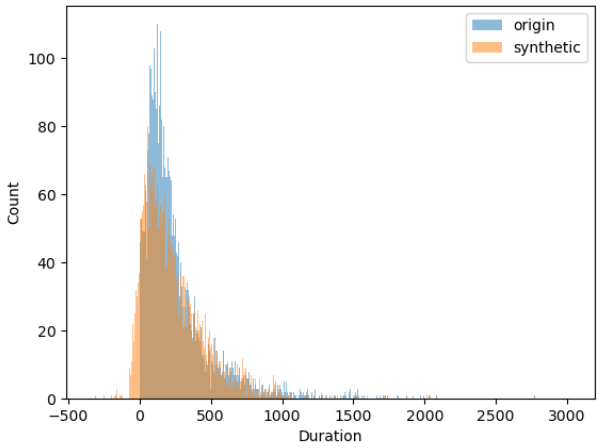

Figure 5. Distribution plot

As figure 5 shown the distribution is bell shape but a little left skewed.

3.the variance of two data set is known. Yes we can calculate the variance for these two data set.

We set our null hypothesis as: u1 - u2 = 0 which means the difference of the mean is 0 between two data set.

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Figure 6. 2 sample Z test equation

We use the 2 sample Z test formula from Figure 6 to calculate the Z value for our test. From calculation the Z score is 4.77

Raw data(Bank)

	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
1	services	married	secondary	no	4789.0	yes	yes	cellular	11	may	220	1	339	4	failure
2	management	single	tertiary	no	135.0	yes	no	cellular	16	apr	185	1	330	1	failure
3	management	married	tertiary	no	1476.0	yes	yes	unknown	3	jun	199	4	-1	0	unknown

Synthetic data(Bank)

	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome
0	technician	divorced	secondary	no	495.977319	yes	no	cellular	1	may	141	2	-1	4	failure
1	management	single	tertiary	no	4444.817445	no	no	cellular	29	aug	86	2	0	0	unknown
2	admin.	divorced	tertiary	no	9.069700	yes	yes	telephone	11	mar	6	2	406	0	unknown

Figure 1. Raw data and synthetic data sample for bank dataset

Raw data(adult)

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States

Synthetic data(adult)

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country
0	28	Private	151886	HS-grad	8	Married-civ-spouse	Craft-repair	Husband	White	Female	10	4	40	United-States
1	51	Private	100938	HS-grad	14	Married-civ-spouse	Exec-managerial	Not-in-family	Black	Male	-24	1	40	United-States
2	33	Without-pay	366507	11th	9	Separated	Protective-serv	Own-child	Black	Male	-39	-1	40	United-States

Figure 2. Raw data and synthetic data sample for adult dataset

z	.00	.01	.02	.03	.04	.05
−3.4	.0003	.0003	.0003	.0003	.0003	.0003
−3.3	.0005	.0005	.0005	.0004	.0004	.0004
−3.2	.0007	.0007	.0006	.0006	.0006	.0006
−3.1	.0010	.0009	.0009	.0009	.0008	.0008
−3.0	.0013	.0013	.0013	.0012	.0012	.0011
−2.9	.0019	.0018	.0018	.0017	.0016	.0016
−2.8	.0026	.0025	.0024	.0023	.0023	.0022
−2.7	.0035	.0034	.0033	.0032	.0031	.0030
−2.6	.0047	.0045	.0044	.0043	.0041	.0040
−2.5	.0062	.0060	.0059	.0057	.0055	.0054
−2.4	.0082	.0080	.0078	.0075	.0073	.0071
−2.3	.0107	.0104	.0102	.0099	.0096	.0094
−2.2	.0139	.0136	.0132	.0129	.0125	.0122
−2.1	.0179	.0174	.0170	.0166	.0162	.0158
−2.0	.0228	.0222	.0217	.0212	.0207	.0202
−1.9	.0287	.0281	.0274	.0268	.0262	.0256
−1.8	.0359	.0351	.0344	.0336	.0329	.0322
−1.7	.0446	.0436	.0427	.0418	.0409	.0401
−1.6	.0548	.0537	.0526	.0516	.0505	.0495

Figure 7. Z table

By checking the Z table from figure 7 with = 5% we get 1.65. Because our z value 4.77 > 1.65 we will reject our null hypothesis which means u1 - u1 != 0. This test shows the two data set have different means.

This shows when we need to use mean of the synthetic data it will not have the same mean as the origin data. This result maybe caused by the left skewed shape of our distribution. Our data set does not have a perfect normal distributed numerical value column. If we can come up with a normal distributed data set we can get a better result from the z test.

8.2 F test

We also did the F-test[2] to compare the variance of two data set with the Duration column. We have our null hypothesis as v1 = v2 which means the two data set have the same variance. From the F test equation F = $s^2 1/s^2 2 We got our F value = 1.23. with the degree of freedom df1 = 4521, df2 = 4521.$

By checking the Z table with = 5% we get 1.35. Because our F value 1.23 < 1.35 we fail to reject our null hypothesis which means v1 = v1.

This result show the synthetic data have the same variance as the origin data. This test also prove that if we are doing calculation with the variance and standard deviation CTGan's synthetic data will provide the same value as the original data. The synthetic data will not influence the utility of the data regarding to variance.

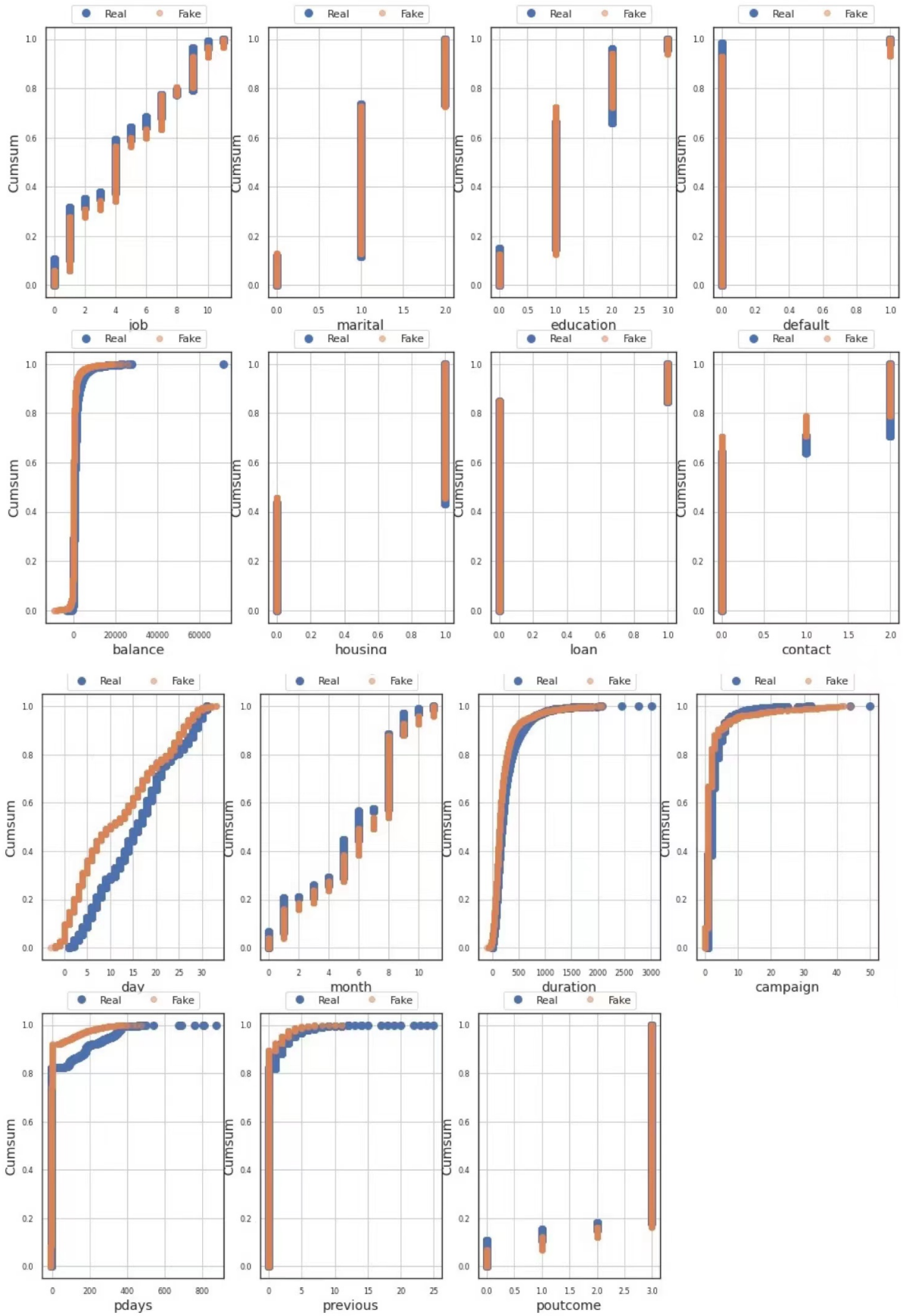

Figure 3. Cumulative Sums and Distribution per Feature for Bank Deposit Dataset

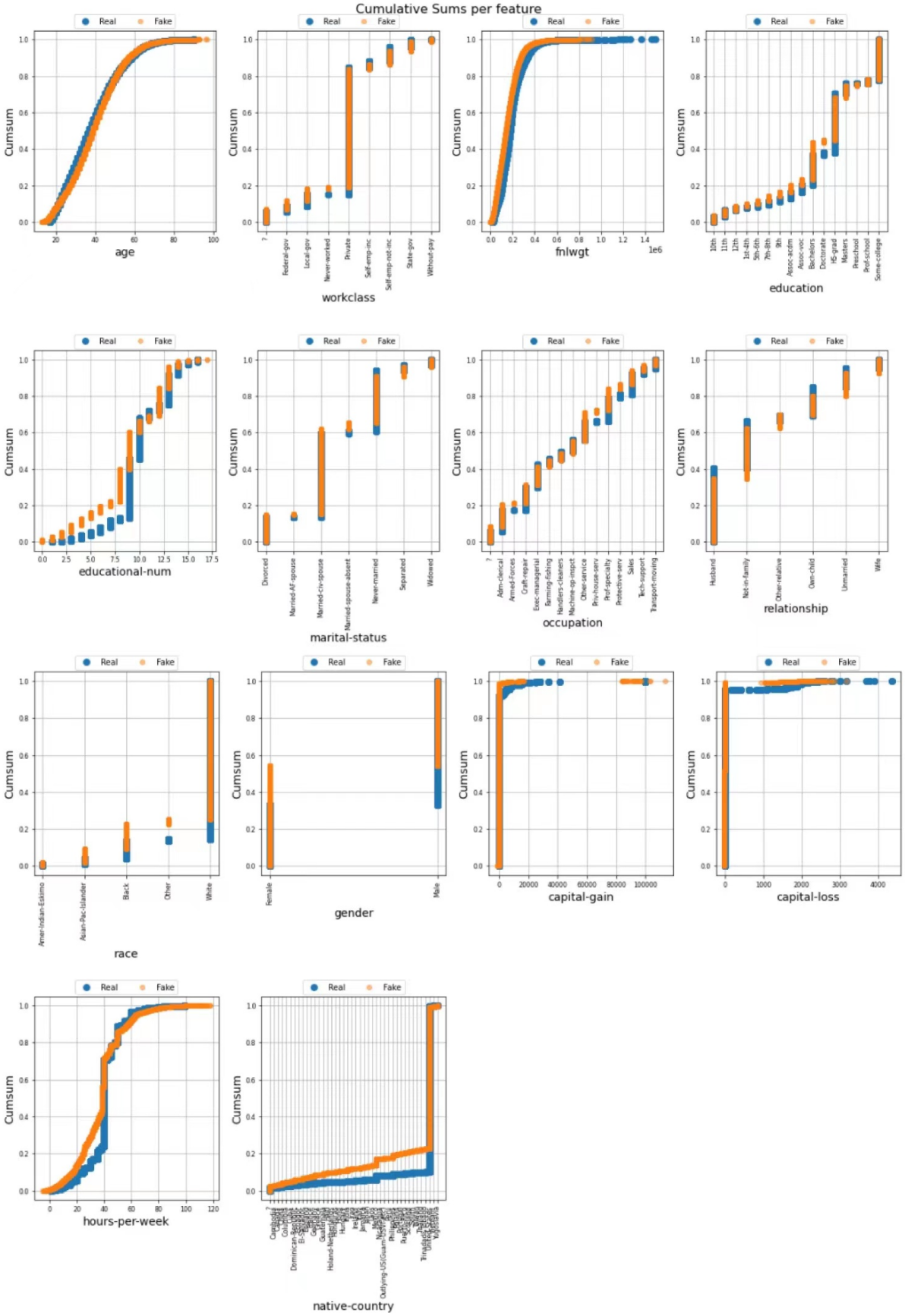

Figure 4. Cumulative Sums and Distribution per Feature for Adult Dataset

8.3 Result from TableEvaluator library

By using the TableEvaluator library. We generated the cumulative Sums and distribution for our data sets in Figure 3 and Figure 4. These plots shows how the synthetic data different from the original data from every data columns of our original data set. The CTGAN library have strong adaptability to generate fake data on different category of data columns. There are similar performance on numerical, categorical, time and text data. From the distribution of data, we notice CTGAN library mainly aims on learning the distribution of the data columns the fake data always match the pattern of the original data. But if you trying to match a single record to original data, you will not get the same information as the original record. In this case the privacy of data is preserved. But at the same time. But at the same time the utility of the fake data is also harmed because CTGAN library learns the distribution of data not accurately match the original data. If user trying to use the fake data to perform number-sensitive calculations. The result will not likely to be accurate.

		Actual	
		True	**False**
Predicted	**True**	931(894)	85(122)
	False	61(93)	53(17)

Table 1. Confusion Matrix on Raw and (Synthetic Data) for Bank Data set for Decision Tree

8.4 Machine Learning Analysis

In table 1, 2, and 3, confusion matrix analysis for the raw data and synthetic dataset were performed. Decision tree, logistic regression algorithm and Naive Bayes algorithm which are the popular ML algorithm for classification are applied into our datasets. First, for bank dataset, the accuracy rate of decision tree for raw data is 87.08%, and the accuracy rate for synthetic data is 80.62%. And which is not a big difference compare these two different datasets. The accuracy for logistic regression for raw data is 89.82%. And synthetic data is 87.17%. The accuracy of Naive Bayes algorithm for raw dataset is around 84%, the accuracy for synthetic data is is even higher(89.56%). But we could see the value of false negative in the confusion matrix is 0, which means CTGan algorithm cannot classify the minority data correctly.

Second, for adult income dataset, the accuracy rate of decision tree for raw data is 81.93%, and the accuracy rate for synthetic data is 62.90%. The accuracy for logistic regression for raw data is 79.08%. And synthetic data is 71.80%. Finally, the accuracy rate of Naive Bayes algorithm for raw dataset is 79.68%, the accuracy rate for synthetic data is 73.18%. The different between raw data and synthetic data of decision tree is greater then the other two algorithms. The reason is that a small change in the data can cause a large change in the structure of decision tree, so this will cause instability.

		Actual	
		True	**False**
Predicted	**True**	996(979)	20(37)
	False	95(108)	19(6)

Table 2. Confusion Matrix on Raw and (Synthetic Data) Bank Data set for Logistic Regression

Predicted \ Actual	True	False
True	911(1012)	105(4)
False	74(114)	40(0)

Table 3. Confusion Matrix on Raw and (Synthetic Data) Bank Data set for Naive Bayes

9 Conclusion

CTGAN library was introduced in the paper Generating Synthetic Tabular Data[8]. And it is a powerful tool to generate synthetic tabular data. We performed performance test on CTGAN library with 2 extra data sets. We performed evaluation experiment with TableEvaluator library, statistical tests and Machine learning Analysis. Our experiment and evaluation show that CTGan performs pretty well on generating synthetic data, and we get a fair performance after running machine learning algorithms with synthetic data. With the trade off of utility user's privacy is preserved and the harm to utility is manageable.

10 Limitation

However, the limitation is that the stability of the Machine learning performance vary between different data set. When data set is high unbalanced, we find that synthetic data does not work very well for classifying the minority data set.

11 Future Work

For future work, we could test more data set and compare the result with raw data.

We could also trying to implement algorithm to adjust the strength of privacy level. Since the library have a good time complexity on dealing with large data. it is easy to implement the algorithm that generate multiple copies of fake data and take average of them to get the data which is more similar to original data. With this new function Users will be able to adjust the privacy strength level and control the trade off between privacy and utility.

References

[1] Cloudflare. [n. d.]. What Is Data Privacy? | Privacy Definition. https://www.cloudflare.com/zh-cn/learning/privacy/what-is-data-privacy/.

[2] J.L. Devore and N.R. Farnum. 1999. *Applied Statistics for Engineers and Scientists*. Duxbury Press. https://books.google.com/books?id=450ZAQAAIAAJ

[3] CTGAN An Open Source Project from the Data to AI Lab. [n. d.]. In *MIT*. https://github.com/sdv-dev/CTGAN

[4] Kaggle. 2021. Bank Term Deposit. https://www.kaggle.com/faviovaz/bank-term-deposit.

[5] OmniSci. [n. d.]. What Is Decision Tree Analysis? Definition. https://www.omnisci.com/technical-glossary/decision-tree-analysis

[6] UCI Machine Learning Repository. [n. d.]. Adult Data Set. archive.ics.uci.edu/ml/datasets/adult

[7] Table-Evaluator. [n. d.]. PyPI. pypi.org/project/table-evaluator/

[8] L. T. Tan. [n. d.]. Generating Synthetic Tabular Data. In *Project Alesia*. https://projectalesia.com/posts/generating-synthetic-tabular-data/

[9] Skoularidou M. Cuesta-Infante A. Veeramachaneni K Xu, L. 2019. Modeling tabular data using conditional gan. In *arXiv*.

[10] Yan Zheng Wang. Xu, Shuo Li. 2017. Multinomial Naïve Bayes Classifier to Text Classification. 347-352.10.1007/978-981-10-5041-1_57.