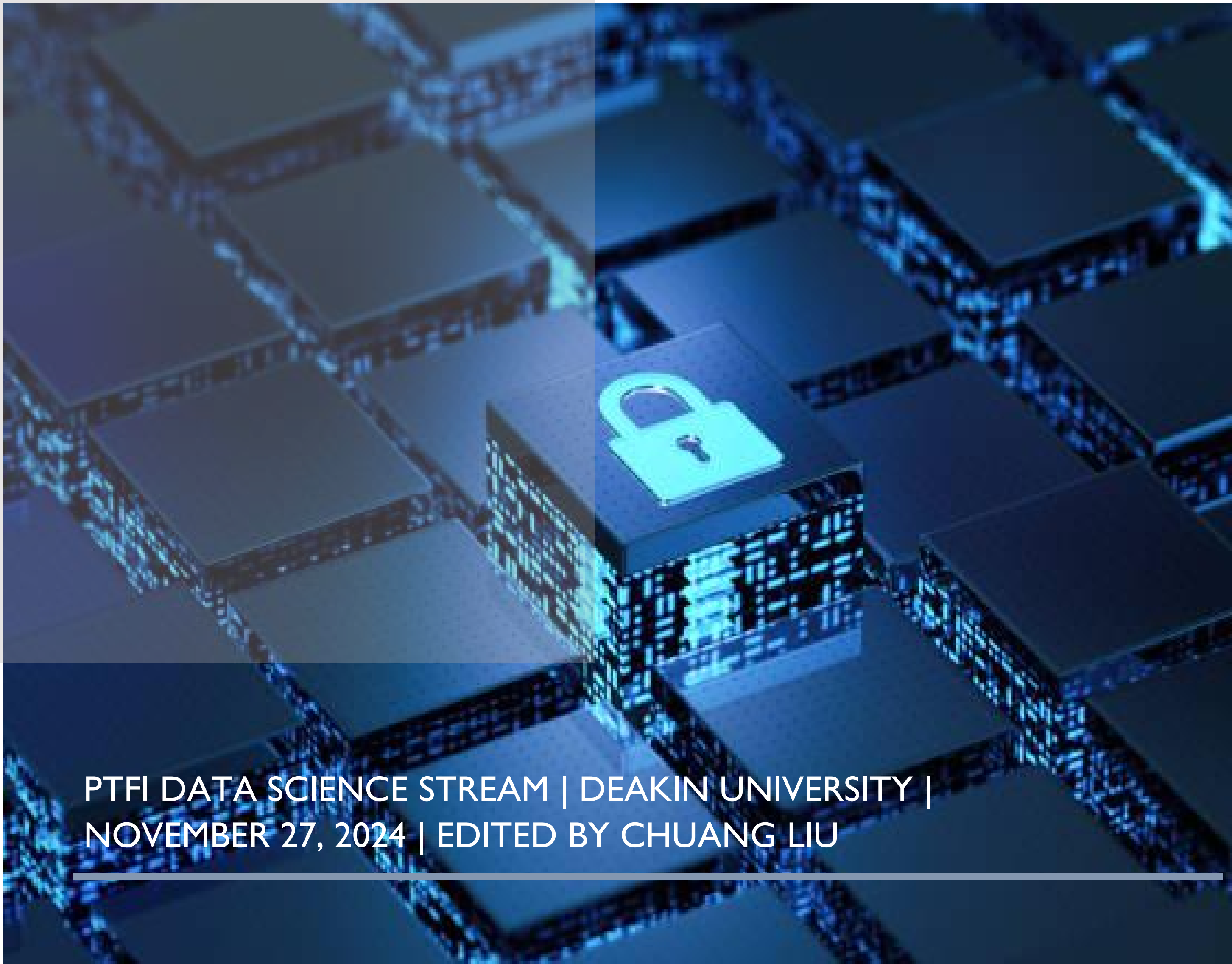
SYNTHETIC DATA ANALYSIS

PRIVACY TECHNOLOGIES FOR FINANCIAL INTELLIGENCE: A DATA BYTES COMPANY PROJECT (2ND EDITION)



PTFI DATA SCIENCE STREAM | DEAKIN UNIVERSITY | NOVEMBER 27, 2024 | EDITED BY CHUANG LIU

1. BACKGROUND

Creating synthetic datasets for financial crime when real datasets are unavailable due to privacy concerns and regulations can be a viable solution. However, it's crucial to note that the effectiveness of synthetic datasets depends on the quality of the generation process. The synthetic data should accurately capture statistical properties and patterns present in real financial data to ensure that models trained on such data are relevant and reliable in real-world applications.

The scope of this analysis was to investigate options for generating synthetic data for applying privacy enhancing technologies. A prior attempt to generate data using the Faker library, mimesis, mockaroo or SDV[1] was successful but there are still shortcomings, such as cannot capture the mathematical properties of mixed data type variables, long tail distributions, skewed multi-mode continuous and single gaussian variables. To solve these four problems, we introduce two new data generation network models CTAB-GAN[2] and CTAB-GAN+[3], which are both open source models on Github.

At the same time, we analyzed four kinds of datasets Approval[4], Default[5], Loan[6] and Adult[7]. They include the study of four financial relationships in order for us to conduct financial data analysis to help further expand the development of PTFI projects.

The report covers as follows:

(i) Some background of datasets (Approval, Default, Loan and Adult);

(ii) some background of models (CTAB-GAN and CTAB-GAN+);

(iii) the application of models (CTAB-GAN and CTAB-GAN+);

(iv) the results achieved and an evaluation of these results.

It is noted that there are many other methods to generate synthetic data, however, due to time constraints only the two methods selected were investigated.

DATASET 1: APPROVAL

In 1987, Quinlan proposed Approval dataset, which title is Australian Credit Approval. This dataset concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This dataset is interesting because there is a good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values. Its number of instances is 690 and its number of attributes is 15 (including target attribute).

What is Approval's goal?

This dataset is used to research Australia credit approval. By analyzing the mathematical properties of 14 features in relation to the target, we can determine whether a customer can pass his/her Australian credit approval.

What is target in Approval?

T1：1,2, class attribute (formerly: +,-).

What are features in Approval?

F1: 0, 1, CATEGORICAL (formerly: a,b).

F2: continuous.

F3: continuous.

F4: 1, 2, 3, CATEGORICAL (formerly: p,g,gg).

F5: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, CATEGORICAL (formerly: ff,d,i,k,j,aa,m,c,w, e, q, r,cc, x).

F6: 1, 2, 3, 4, 5, 6, 7, 8, 9, CATEGORICAL (formerly: ff,dd,j,bb,v,n,o,h,z).

F7: continuous.

F8: 1, 0, CATEGORICAL (formerly: t, f).

F9: 1, 0, CATEGORICAL (formerly: t, f).

F10: continuous.

F11: 1, 0, CATEGORICAL (formerly t, f).

F12: 1, 2, 3, CATEGORICAL (formerly: s, g, p).

F13: continuous.

F14: continuous.

DATASET 2: DEFAULT

In 2009, Yeh proposed Default dataset, which title is Default of Credit Card Clients. This dataset is aimed at the case of customers' default payments in Taiwan, which is used to get predictive accuracy of probability of default. Its number of instances is 30000 and its number of attributes is 24 (including target attribute).

What is Default's goal?

This dataset is used to research default payments in Taiwan. By analyzing the mathematical properties of 23 features in relation to the target, we can determine whether a customer from Taiwan can make default payments.

What is target in Default?

default payment next month： 0, 1, Does this user get default payment next month?.

What are features in Default?

LIMIT_BAL: continuous, amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

SEX: 0, 1, gender (1 = male; 2 = female).

EDUCATION: 1, 2, 3, 4, education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

MARRIAGE: 1, 2, 3, marital status (1 = married; 2 = single; 3 = others).

AGE: continuous, age (year).

PAY_0, PAY_2-PAY_6: continuous, history of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

BILL_AMT1-BILL_AMT6: continuous, amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

PAY_AMT1-PAY_AMT6: continuous, amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

DATASET 3: LOAN

In 2024, Jacob proposed Loan dataset, which title is Bank_Loan_modelling. This dataset is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns to better target marketing to increase the success ratio with a minimal budget. Its number of instances is 5000 and its number of attributes is 13 (including target attribute).

What is Loan's goal?

This dataset is used to research bank loan in Thera Bank. By analyzing the mathematical properties of 12 features in relation to the target, we can determine whether a Thera Bank's customer can handle a bank loan.

What is target in Loan?

Personal Loan：0, 1, Did this customer accept the personal loan offered in the last campaign?.

What are features in Loan?

Age: continuous, customer's age in completed years.

Experience: continuous, years of professional experience.

Income: continuous, annual income of the customer.

ZIPCode: continuous, Home Address ZIP code.

Family: 1, 2, 3, 4, Family size of the customer.

CCAvg: continuous, Avg. spending on credit cards per month, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

Education: 1, 2, 3, Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional.

Mortgage: continuous, value of house mortgage if any.

CreditCard: 0, 1, Does the customer use a credit card issued by UniversalBank?.

Securities Account: 0, 1, Does the customer have a securities account with the bank?

CD Account: 0, 1, Does the customer have a certificate of deposit (CD) account with the bank?.

Online: 0, 1, Does the customer use internet banking facilities?.

DATASET 4: ADULT

In 1996, Becker et al. proposed Adult dataset, which title is Adult. And its duplicate or conflicting instances is 6. Its number of instances of probability for the label '>50K' : 23.93% / 24.78% (without unknowns) and number of instances of probability for the label '<=50K' : 76.07% / 75.22% (without unknowns). Prediction task is to determine whether a person makes over 50K a year. Its number of instances is 48842 and its number of attributes is 14 (including target attribute).

What is Adult’s goal?

This dataset is used to research annual income of an individual from USA. By analyzing the mathematical properties of 13 features in relation to the target, we can determine whether annual income of an individual from USA can exceed $50k/yr.

What is target in Adult?

income：>50K, <=50K, >50K., <=50K.

What are features in Adult?

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

MODEL 1: CTAB-GAN

In ACML 2021, Zhao et al. proposed CTAB-GAN, which is a conditional GAN based tabular data generator. CTAB-GAN advances beyond the prior state-of-the-art methods by modeling mixed variables and provides strong generation capability for imbalanced categorical variables, and continuous variables with complex distributions. To such ends, the core features of CTAB-GAN include as follows:

(i) introduction of the classifier into conditional GAN;

(ii) effective data encoding for mixed variable;

(iii) a novel construction of conditional vectors.

Which type of mathematical properties can be studied by CTAB-GAN?

(i) mixed data type variables;

(ii) long tail distributions;

(iii) skewed multi-mode continuous.

Which type of data was generated?

(i) Approval;

(ii) Default;

(iii) Loan;

(iv) Adult.

Results

The results indicated that evaluation visualisation demonstrates a remarkable similarity between four datasets and the synthetic data generated by CTAB-GAN. CTAB-GAN, a machine learning-based approach, effectively learns intricate patterns from the real data and realistically replicates them in the synthetic dataset. This model is particularly well-suited for our project’s goal of generating high-quality synthetic data closely mirrors the characteristics of the original data records.

Advantages

(i) On Adult, the results of CTAB-GAN are even better than on CTAB-GAN+;

(ii) The running time is less than that of CTAB-GAN+, roughly an order of magnitude difference.

Disadvantages

(i) The results are not as good as CTAB-GAN+ on Approval, Default, and Loan.

MODEL 2: CTAB-GAN+

In Frontiers in Big Data 2024, Zhao et al. proposed CTAB-GAN+, which is a conditional GAN based tabular data generator. CTAB-GAN+ advances beyond SOTA methods by improving performance on regression datasets and allowing control over the quality of synthesized data. The core features ofCTAB-GAN+ include as follows:

(i) introduction of the auxiliary component, i.e., classifier or regressor, into conditional GAN;

(ii) effective data encodings for mixed and simple Guassian variables;

(iii) a novel construction of conditional vectors;

(iv) tailored DP discriminator for tabular GAN.

Which type of mathematical properties can be studied by CTAB-GAN?

(i) mixed data type variables;

(ii) long tail distributions;

(iii) skewed multi-mode continuous;

(iv) single gaussian variables.

Which type of data was generated?

(i) Approval;

(ii) Default;

(iii) Loan;

(iv) Adult.

Results

The results indicated that evaluation visualisation demonstrates a remarkable similarity between four datasets and the synthetic data generated by CTAB-GAN+. CTAB-GAN+, a machine learning-based approach, effectively learns intricate patterns from the real data and realistically replicates them in the synthetic dataset. This model is particularly well-suited for our project's goal of generating high-quality synthetic data closely mirrors the characteristics of the original data records.

Advantages

(i) The results are better than CTAB-GAN on Approval, Default, and Loan.

Disadvantages

(i) On Adult, the results of CTAB-GAN+ are even worse than on CTAB-GAN;

(ii) The running time is more than that of CTAB-GAN, roughly an order of magnitude difference.

RECOMMENDATIONS

Given the information above, both CTAB-GAN and CTAB-GAN+ have their own strengths. If you are in a hurry, you can use tab-gan to generate a good simulation data set. However, if you want to improve the quality and effect more finely, it is recommended to use CTAB-GAN+. Specific situation, specific treatment.

The full notebooks and models for CTAB-GAN and CTAB-GAN+ can be located on editor's GitHub:

https://github.com/liuchuang00/PTFI_Task_1

REFERENCES

[1] PTFI (2024). Synthetic Data Analysis [Online]. Github Repository. Available: Privacy-Technologies-for-Financial-Intelligence/T1-2024-synthetic dataset methods at Data-science · DataBytes-Organisation/Privacy-Technologies-for-Financial-Intelligence.

[2] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "CTAB-GAN: Effective Table Data Synthesizing," in Asian Conference on Machine Learning. PMLR, 2021, pp. 97–112.

[3] Z. Zhao, A. Kunar, R. Birke, H. Van der Scheer, and L. Y. Chen, "CTAB-GAN+: Enhancing Tabular Data Synthesis," Frontiers in big Data, vol. 6, p. 1296508, 2024.

[4] R. Quinlan (1987). Statlog (Australian Credit Approval) [Dataset]. UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval.

[5] I. Yeh (2009). Default of Credit Card Clients [Dataset]. UCI Machine Learning Repository. Available: http://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients.

[6] Sunil Jacob (2024). Bank_Loan_modelling [Dataset]. Kaggle Repository. Available: https://www.kaggle.com/datasets/itsmesunil/bank-loan-modelling/data.

[7] B. Becker and R. Kohavi (1996). Adult [Dataset]. UCI Machine Learning Repository. Available: https://archive.ics.uci.edu/dataset/2/adult.