Generative Adversarial Network based Synthetic Data Generation System

Presented By:

Md. Aukerul Moin Shuvo
Roll: 1603061
CSE, RUET

Md. Shohanoor Rahman
Roll: 1603112
CSE, RUET

Supervised By:

Barshon Sen
Assistant Professor
CSE, RUET

August 3, 2022

Contents

Introduction

What is Synthetic Data?

- Artificially manufactured rather than generated by real-world events [1].
- **Example**:

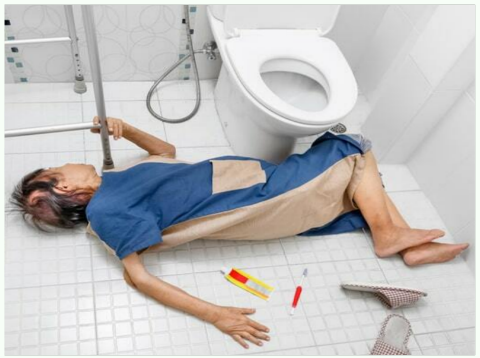
Figure 1: Synthetic People Face [2]

Introduction

Synthetic Data: Use Case

- Used when actual data is rare or expensive.
- **Example**: Rare Objects & Places, Privacy issue, Black Swan Events etc.

Rare Golden Tiger



Heart Attack in Private Place

Falling Stars

Figure 2: Use Case of Synthetic Data

Related Works

Paper Name

Modeling Tabular Data Using Conditional GAN by *Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, Jiayu Zhou* [3].

Contributions

- Model learns complicated distributions.
- Uses mode specific normalization.
- Outperforms Bayesian Networks in terms of distribution learning.
- Conditional data generator.

Overview of Proposed System

Methodology

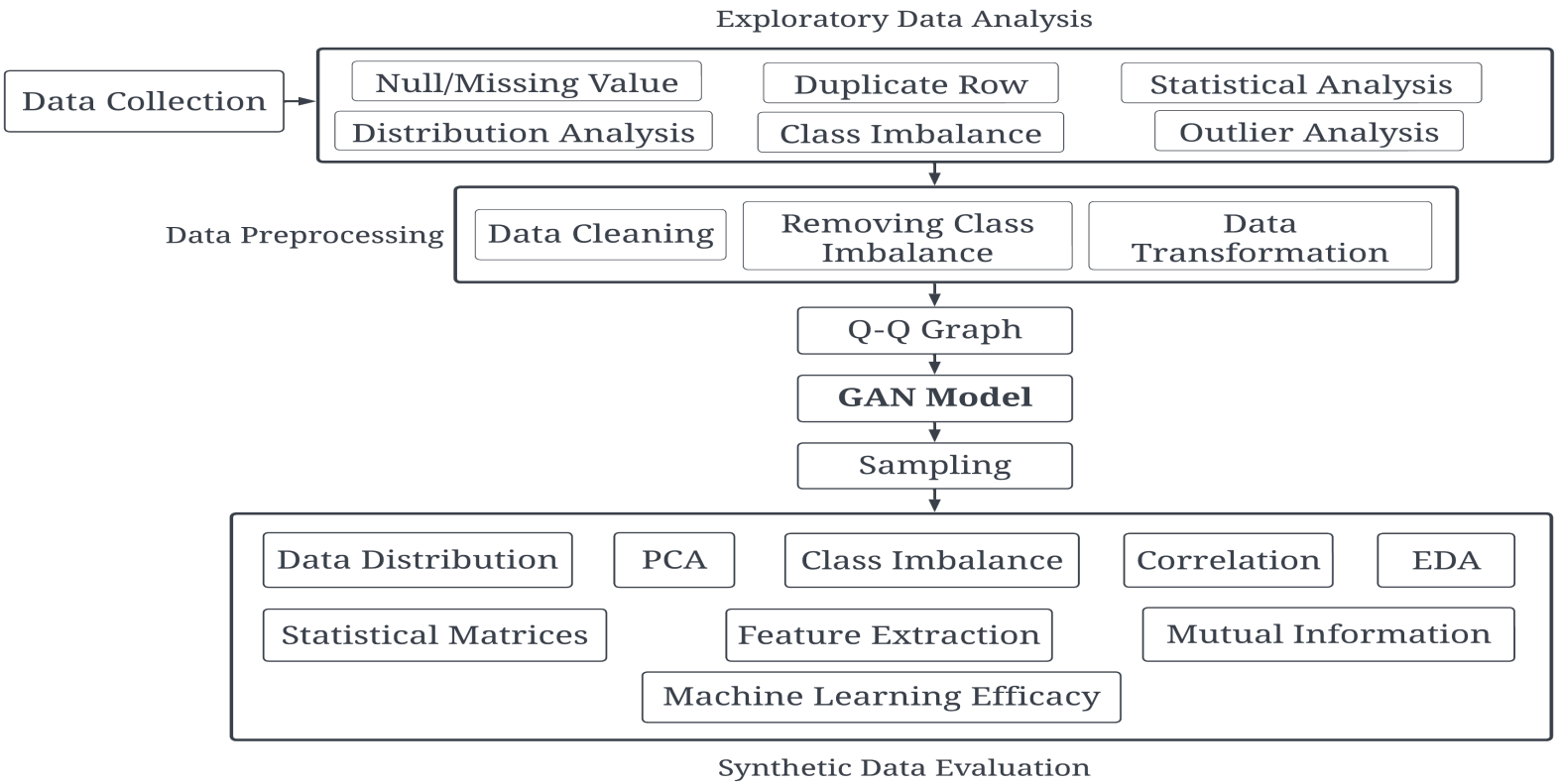


Figure 3: Block Diagram of the Proposed System

Overview of Proposed System (Cont'd)

Exploratory Data Analysis

- **Dataset Information**: Number of entities, features & their types.
- **Missing Value Analysis**: Checking absence of any datapoint.
- **Duplicate Data Observation**: Identifying rows having same value.
- **Data Description**: Data count, Mean, Standarad Deviation, Min, First Quartile, Median, Third Quartile, Max.
- **Data Distribution**: All possibles values of the features and their frequencies.
- **Class Imbalance**: Number of samples per class.

Overview of Proposed System (Cont'd)

Data Preprocessing

- **Data Cleaning**: Removal of Noise & Missing values.
- **Class Imbalance**: Removal of class imbalance using SMOTE technique[4].
- **Data Transformation**:Log Transformation, Power Transformation[5].
- **Quantile-Quantile Plot**: Similar distributional aspects after Data Transformation[6].

Overview of Proposed System (Cont'd)

Feature Transformation and Scaling

- Datasets usually have different columns with different units,ranges and distribution.
- Used to make sure that the model treats both these variable units and ranges equally [5].
- Also to make sure that the data distribution is normal or less skewed [5].

Overview of Proposed System (Cont'd)

GAN Training (Cont'd)

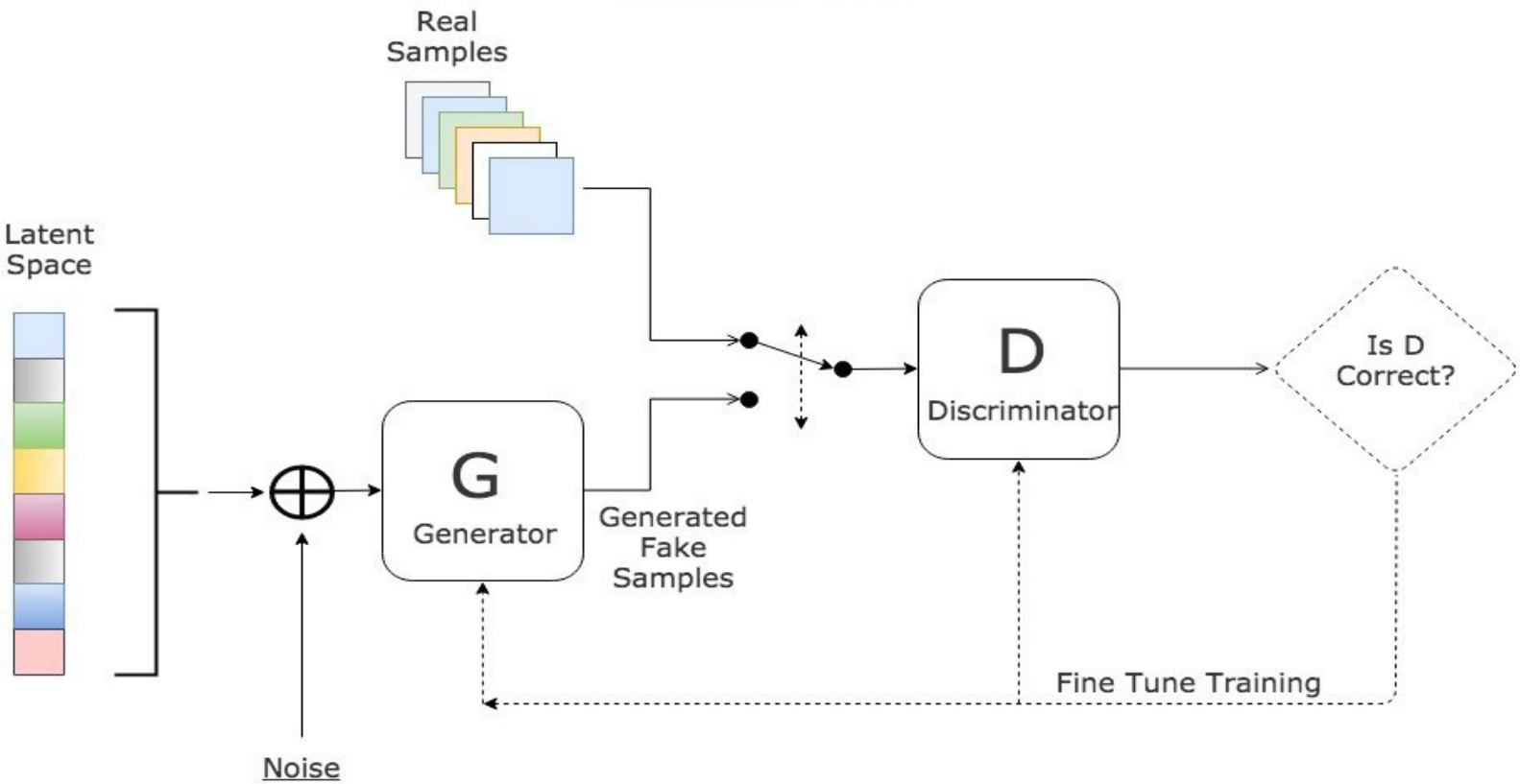


Figure 4: GAN Training Pipeline [7].

Overview of Proposed System (Cont'd)

Overview of CTGAN

- ***Generator*** uses **Batch-normalization** and **ReLU** activation function [3].
- Synthetic row representation is generated using **mixed activation functions**[3].
- Scalar values are generated by **Tanh**[3].
- Mode indicator and discrete values are generated by **Gumbel Softmax**[3].
- ***Discriminator*** uses **Leaky ReLU** function and **Dropout** on each hidden layer[3].

Overview of Proposed System (Cont'd)

CTGAN Model

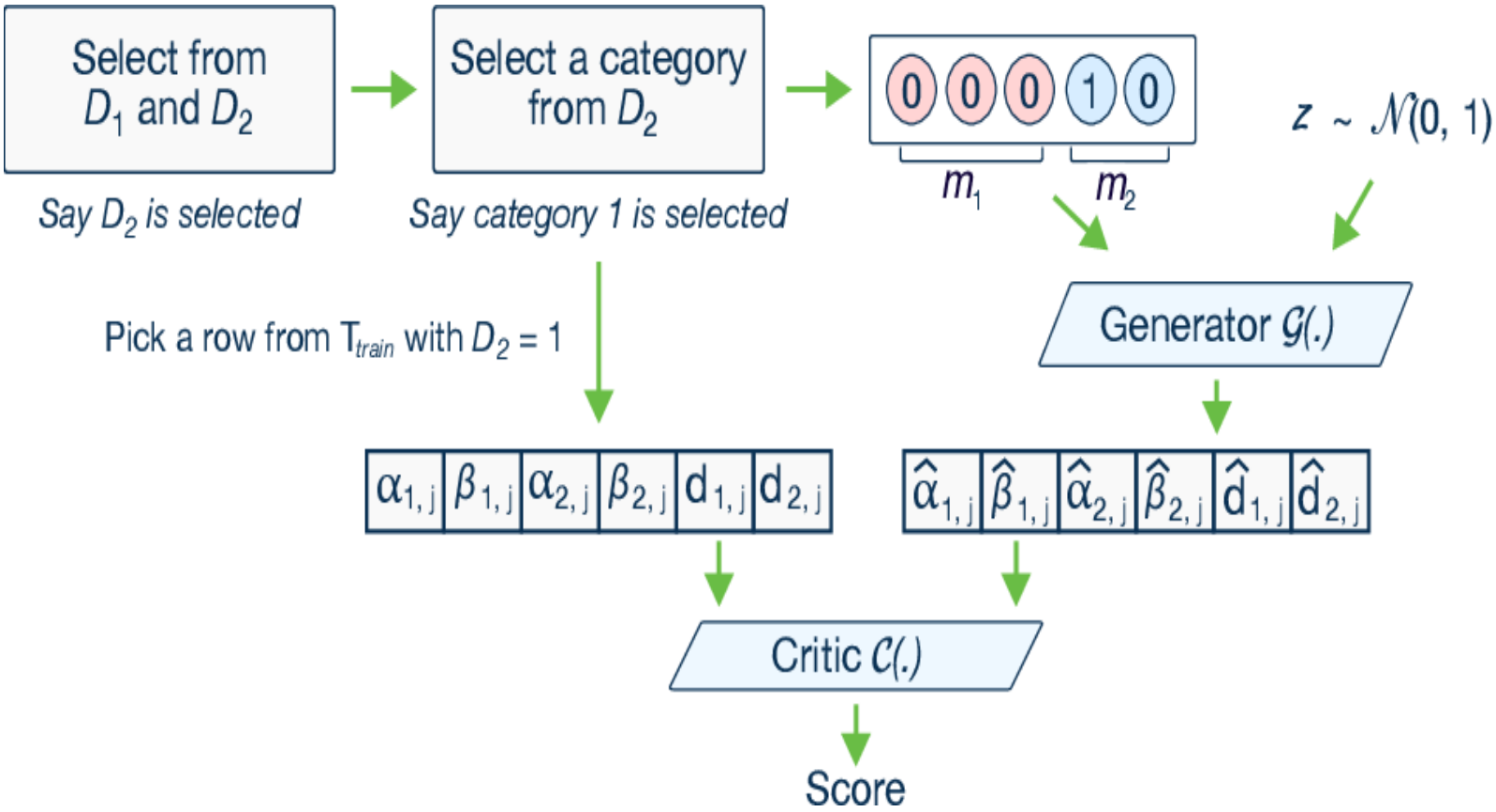


Figure 5: CTGAN Model [3]

Implementation

Exploratory Data Analysis: Data Information

Taiwanese Bankruptcy Prediction Data Set [8]

- Number of Instances: **6819**
- Number of Attributes: **96**

Implementation (Cont'd)

Exploratory Data Analysis: Data Description

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue
count	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000
mean	0.032263	0.505180	0.558625	0.553589	0.607948	0.607929	0.998755	0.797190	0.809084	0.303623
std	0.176710	0.060686	0.065620	0.061595	0.016934	0.016916	0.013010	0.012869	0.013601	0.011163
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.476527	0.535543	0.527277	0.600445	0.600434	0.998969	0.797386	0.809312	0.303466
50%	0.000000	0.502706	0.559802	0.552278	0.605997	0.605976	0.999022	0.797464	0.809375	0.303525
75%	0.000000	0.535563	0.589157	0.584105	0.613914	0.613842	0.999095	0.797579	0.809469	0.303585
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

8 rows × 96 columns

Figure 6: Description of data for a portion of dataset

Implementation (Cont'd)

Exploratory Data Analysis: Data Distribution

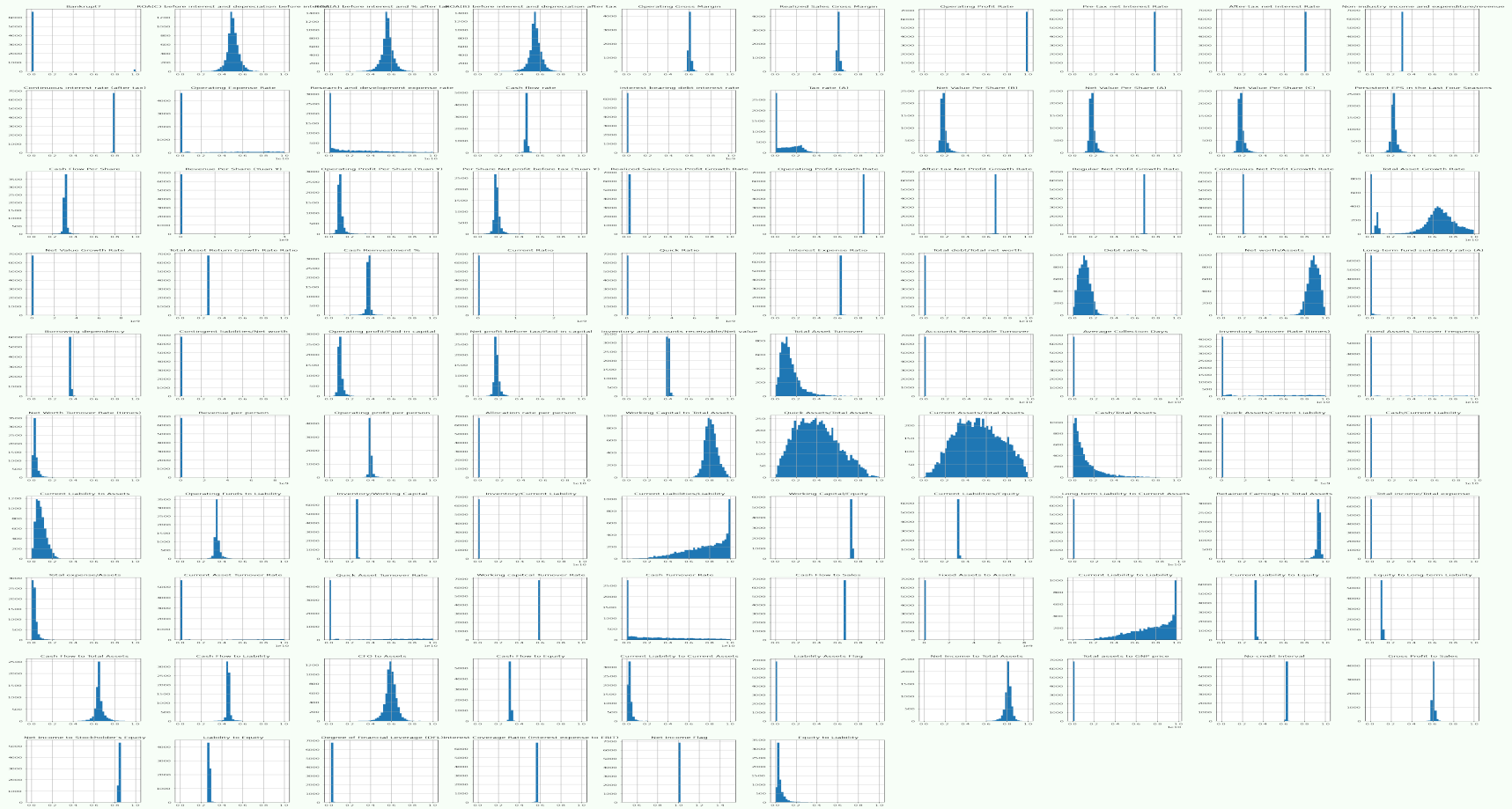


Figure 7: Distribution of Data

Implementation (Cont'd)

Data Preprocessing: Feature Transformation

- Compared **Log Transformation** and **Power Transformation: Yeo-Johnson**.
- **Power Transformation** gives more normal distribution than Natural Log Transformation.

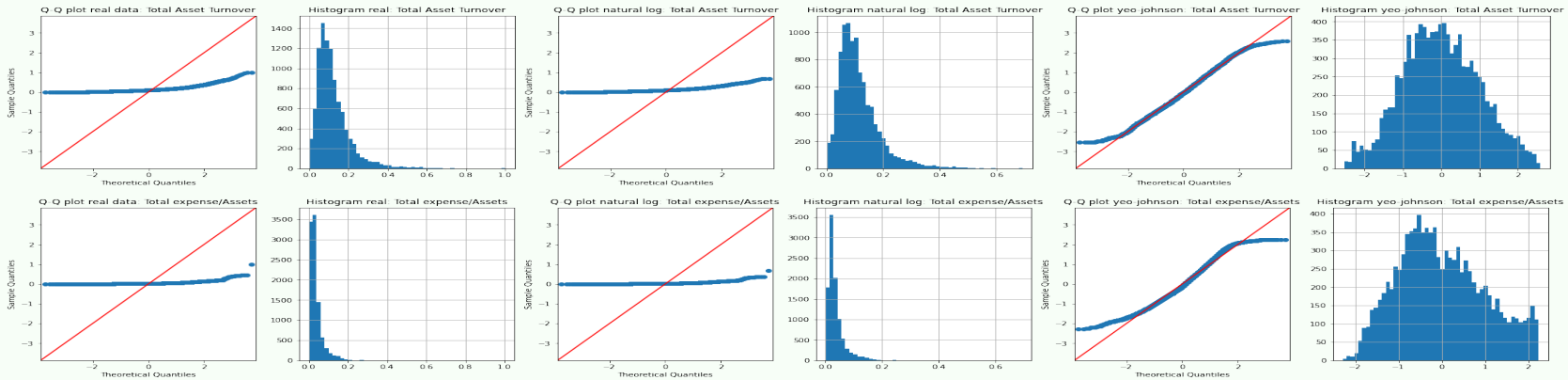


Figure 8: Q-Q graph comparison before(Left) and after Log(Middle) and Power(Right) Transformation

Implementation (Cont'd)

Exploratory Data Analysis: Class Imbalance

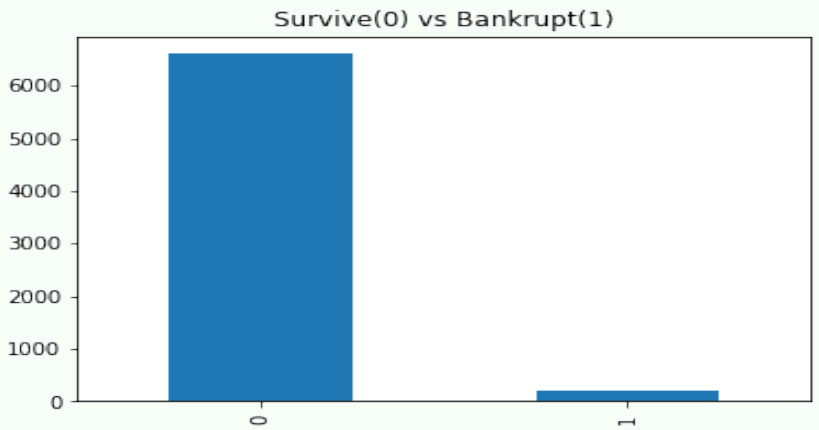


Figure 9: Class Imbalance in Data

- Data set is very imbalanced.
- **96.77%** (6599) data is of Survived company.
- **3.23**% (220) data is of Bankrupted company.

Implementation (Cont'd)

Data Preprocessing: Removal of Class Imbalance

- **SMOTE**: Oversampling Technique[4].

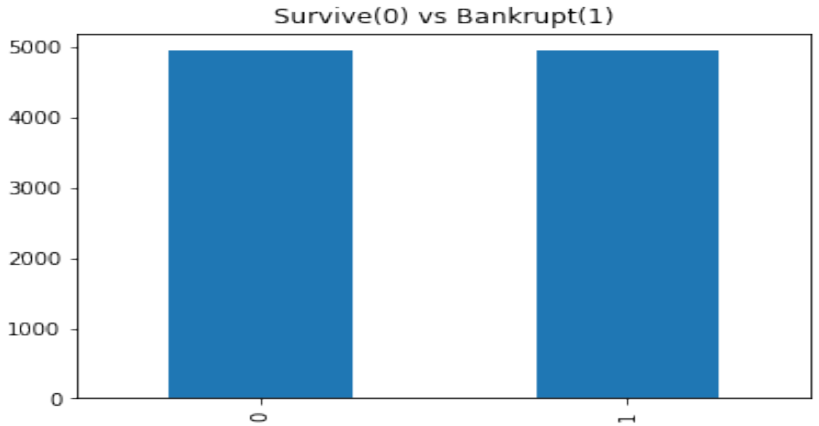


Figure 10: Result of SMOTE Oversampling

Implementation (Cont'd)

CTGAN Model Training

- Batch Size= 500
- Epochs= 10000
- Embedding Dimension= 128
- Generator Dimension= (256, 256)
- Discriminator Dimension= (256, 256)
- Generator Learning Rate = 0.0005
- Discriminator Learning Rate = 0.0005
- Taken from Benchmark implementation [3]

.

Result: Synthetic Data Evaluation

Exploratory Data Analysis: Missing Value Analysis

- No missing value in the synthetic data.

	variables	missing values in percentage
0	Bankrupt?	0.0
1	ROA(C) before interest and depreciation befor...	0.0
70	Total expense/Assets	0.0
69	Total income/Total expense	0.0

Figure 11: Result of Missing Value Analysis in Synthetic Data

- Synthetic data doesn't contain any duplicate row.

Synthetic Data Evaluation (Cont'd)

Exploratory Data Analysis: Class Imbalance

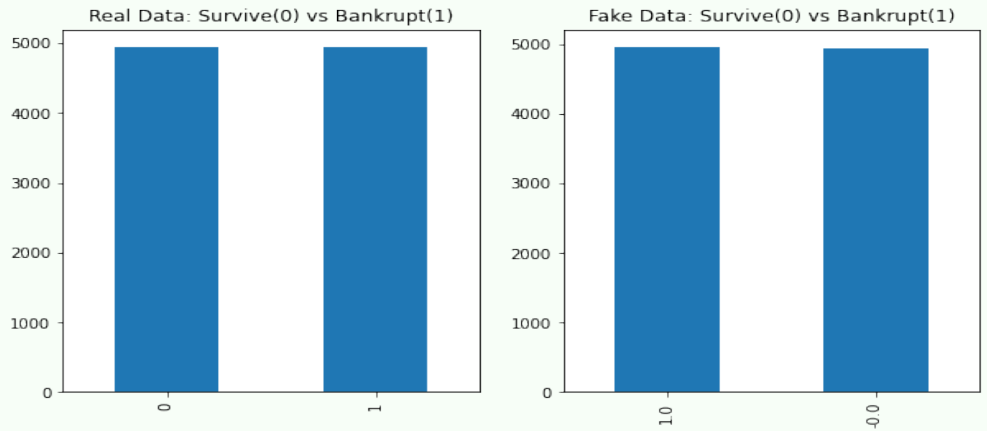
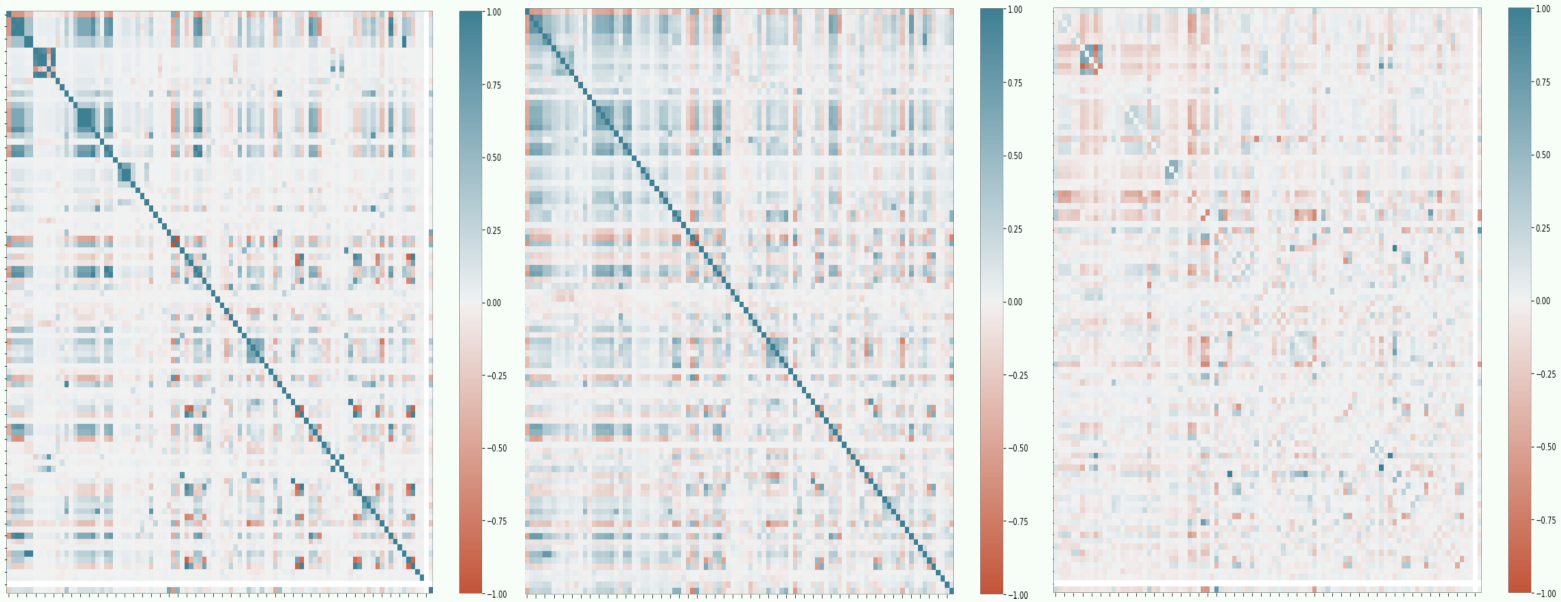


Figure 12: Class Imbalance Analysis in Real Data and Synthetic Data

- No Class Imbalance in synthetic data.
- 50% data is of Survived company.
- 50% data is of Bankrupted company.

Synthetic Data Evaluation (Cont'd)

Exploratory Data Analysis: Correlation Analysis



(a) Real Data (b) Synthetic Data (c) Difference

Figure 13: Correlation Analysis

Synthetic Data Evaluation (Cont'd)

Exploratory Data Analysis: Top 3 Correlation Analysis

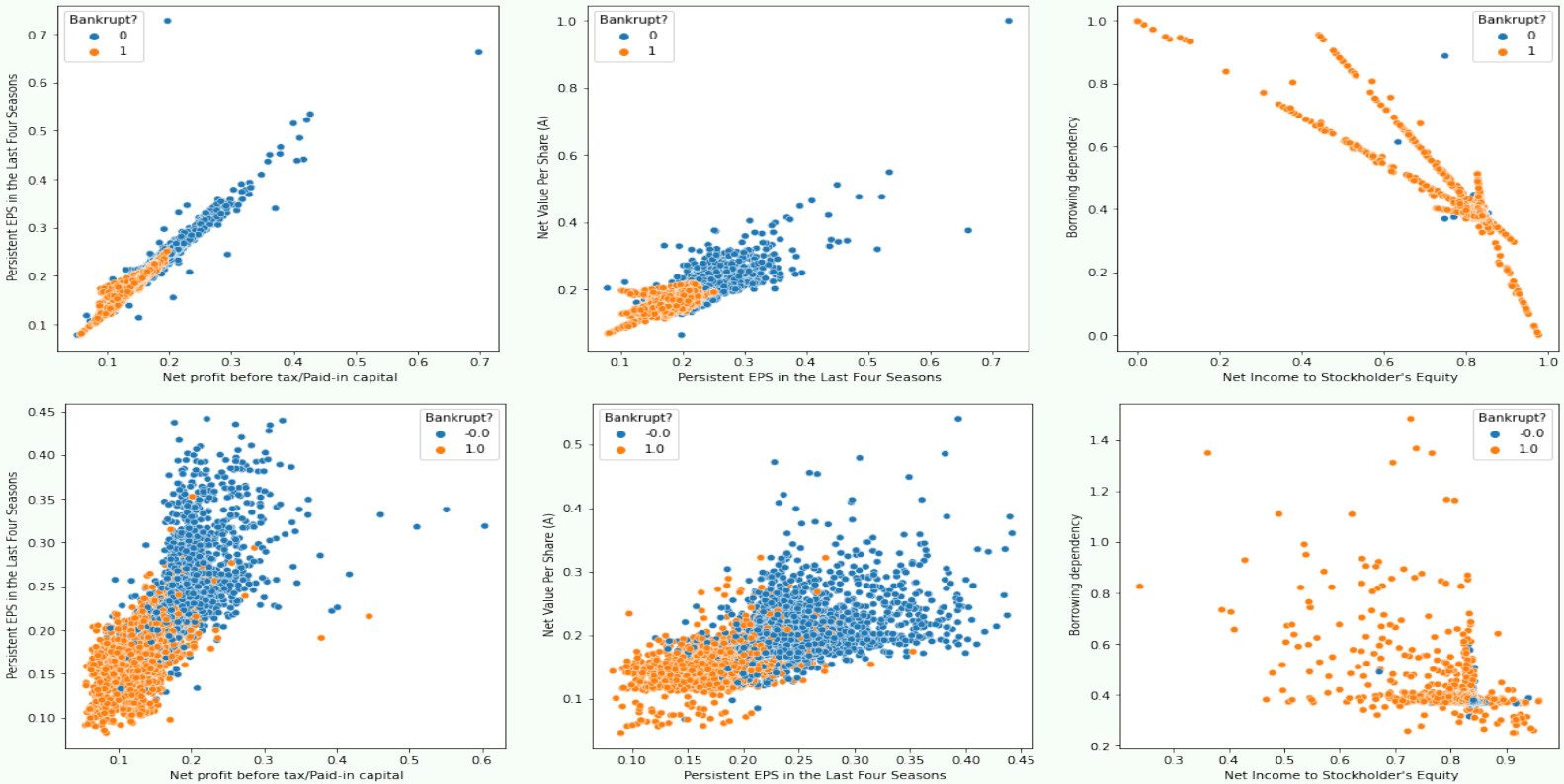


Figure 14: Top 3 Correlation Analysis: Real(Upper) & Synthetic(Lower) Data

Synthetic Data Evaluation (Cont'd)

Data Distribution Comparison: Real and Synthetic

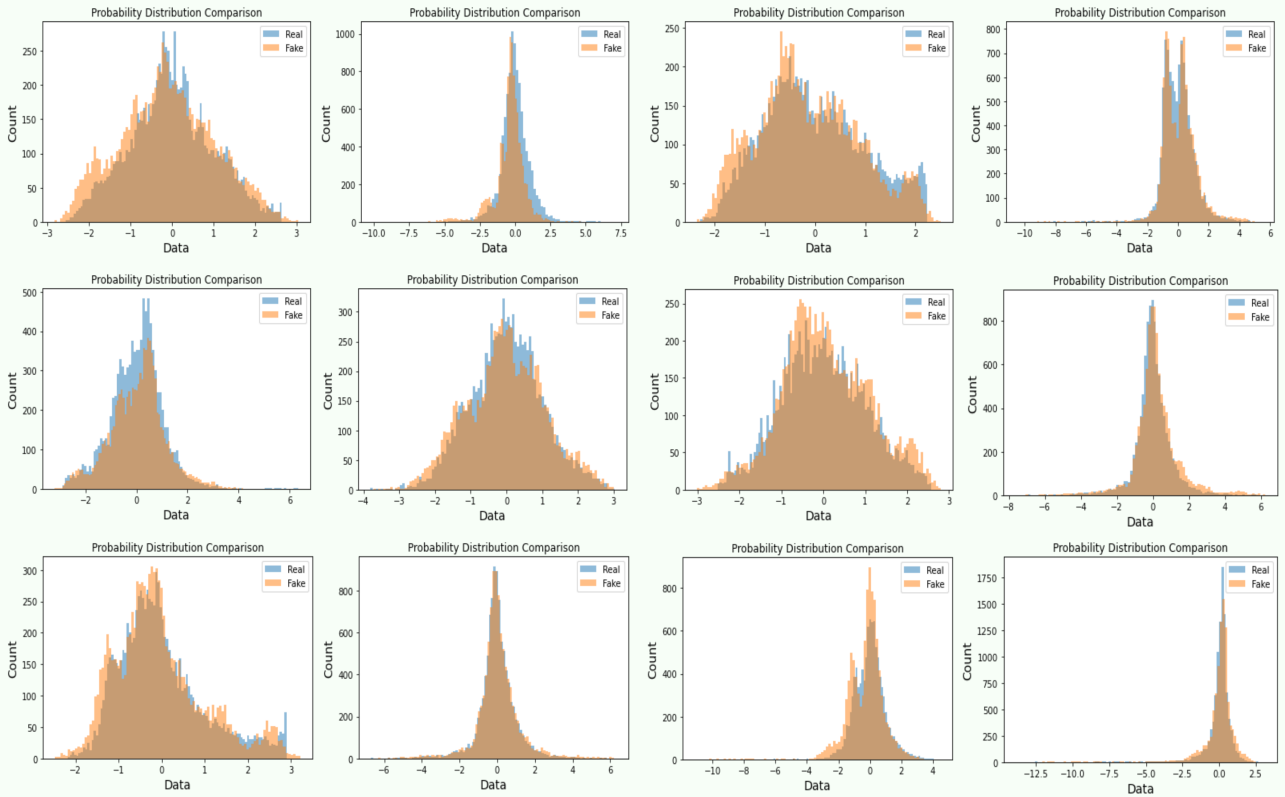


Figure 15: Distribution Comparison of Some Features: Real & Synthetic Data

Synthetic Data Evaluation (Cont'd)

Principal Component Analysis(PCA): Two Most Descriptive Principal Components

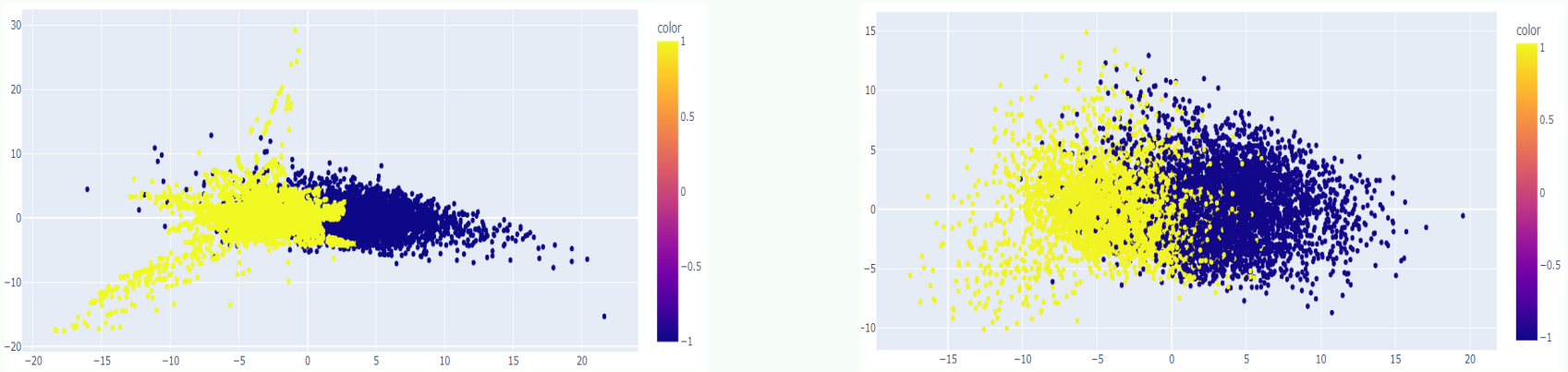
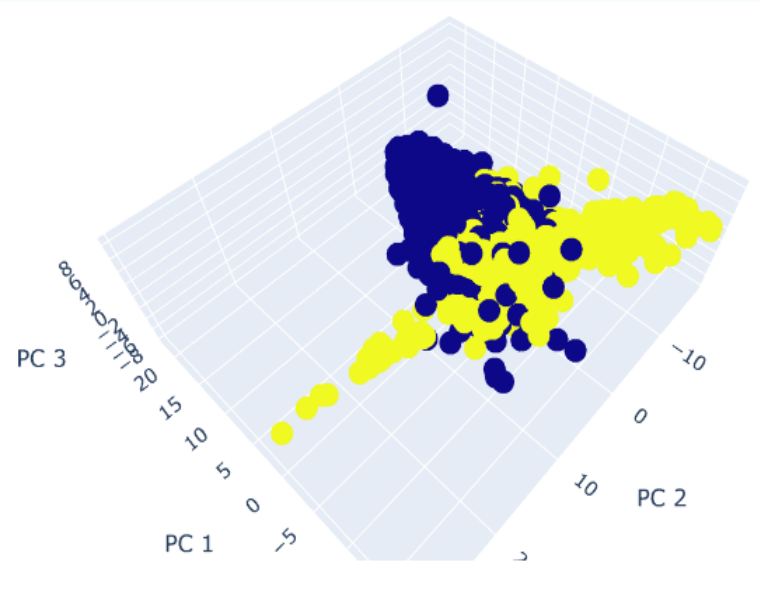


Figure 16: Two Most Descriptive PCA: Real(Left) & Synthetic(Right) Data

Synthetic Data Evaluation (Cont'd)

Principal Component Analysis(PCA): Three Most Descriptive Principal Components



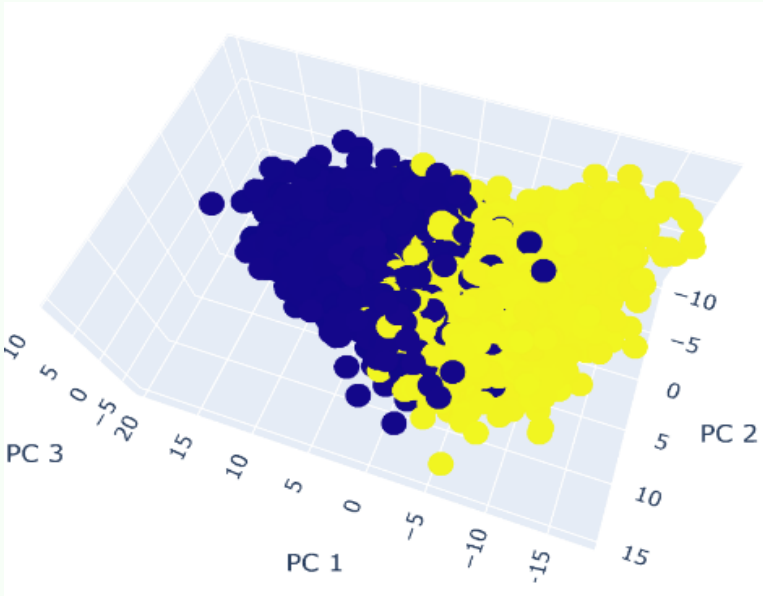


Figure 17: Three Most Descriptive PCA: Real(Left) & Synthetic(Right) Data

Synthetic Data Evaluation (Cont'd)

Statistical Matrices

Metric	Name	Raw Score	Normalized Score	Minimum value	Maximum value	Goal
LogisticDetection	Logistic Regression Detection	1.00	1.00	0.0	1.0	MAXIMIZE
SVCDetection	SVC Detection	1.00	1.00	0.0	1.0	MAXIMIZE
KSTest	Kolmogorov-Smirnov Statistic	8.64e-01	0.86	0.0	1.0	MAXIMIZE
KSTestExtended	Inverted Kolmogorov-Smirnov Statistic	8.63e-01	0.86	0.0	1.0	MAXIMIZE

Figure 18: Different Types of Statistical Tests on Synthetic Data

Synthetic Data Evaluation (Cont'd)

Mutual Information (MI)

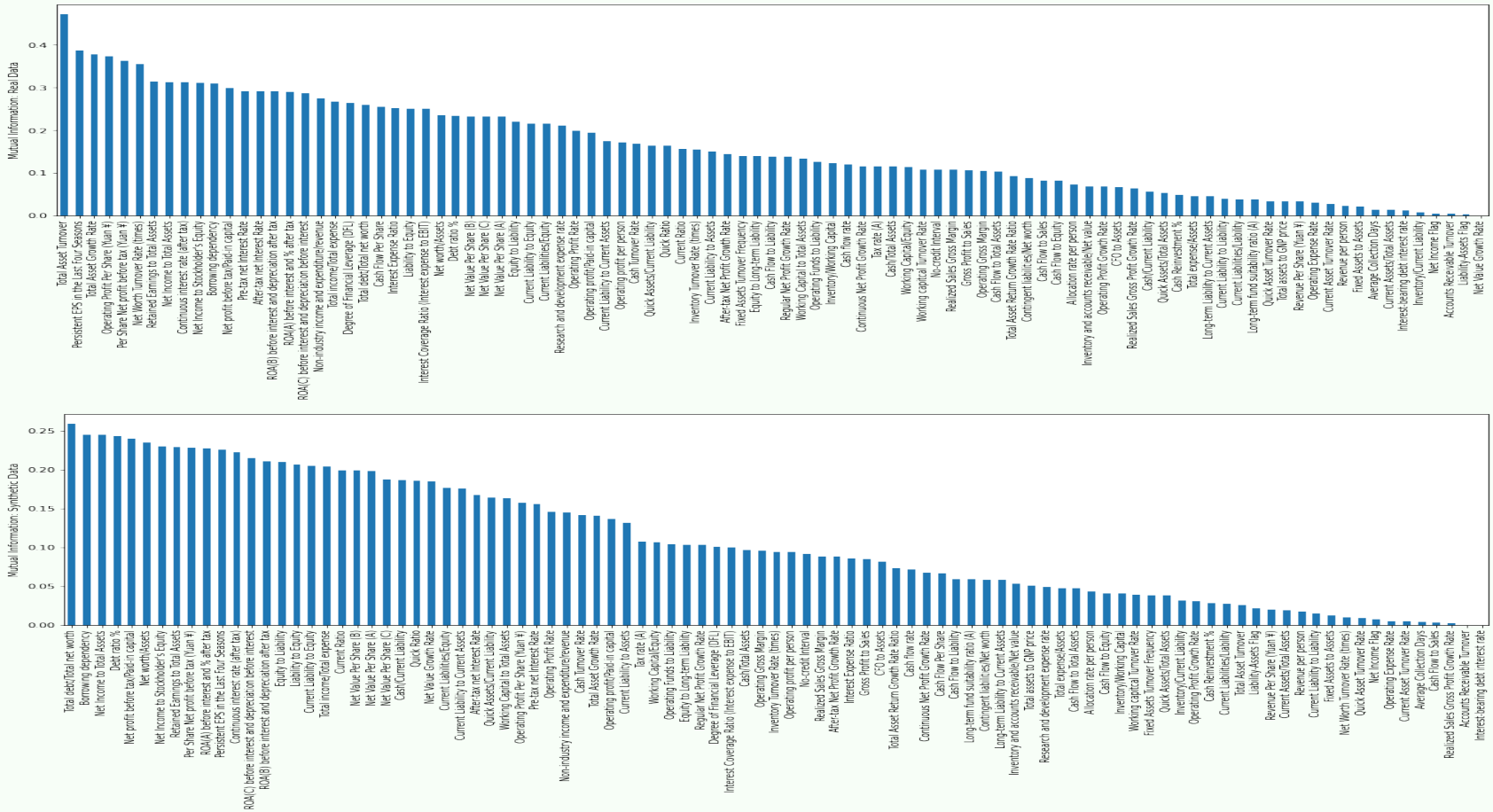


Figure 19: MI of Features: Real(Upper) and Synthetic(Lower) Data

Synthetic Data Evaluation (Cont'd)

Feature Extraction: Recursive Feature Elimination with Cross-validation (RFECV)

- Optimal number of features for **Real Data**: **32**
- Optimal number of features for **Synthetic Data**: **41**

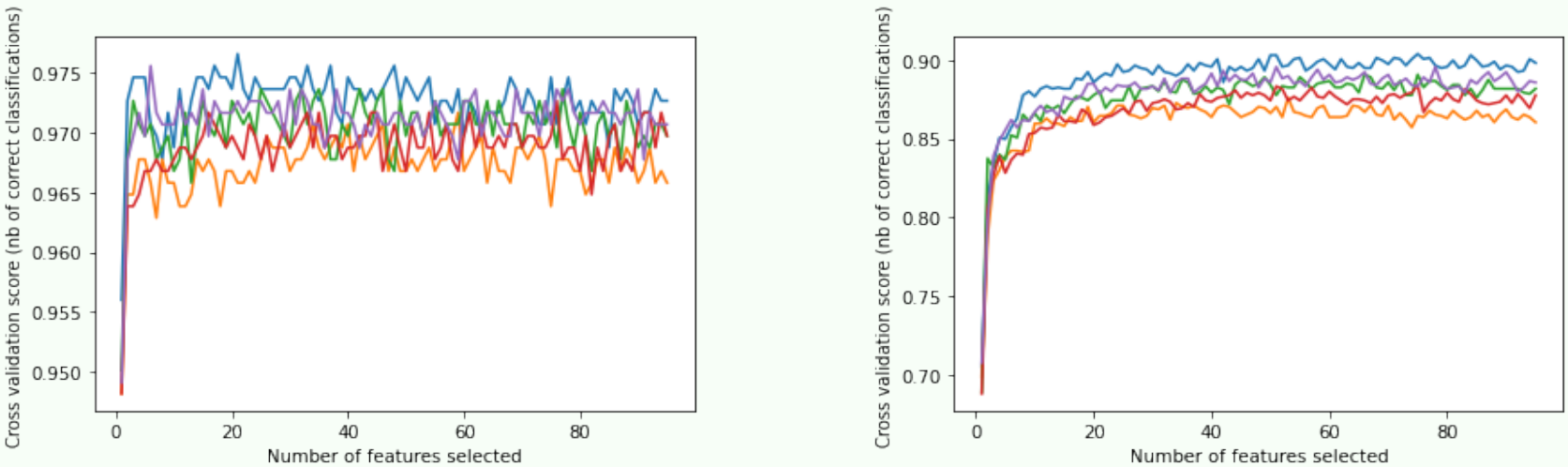


Figure 20: RFECV Feature Elimination: Real(Left) & Synthetic(Right) Data

Synthetic Data Evaluation (Cont'd)

Feature Extraction: Chi-Squared Feature Selection

Real Data		Synthetic Data	
Specs	Score	Specs	Score
Fixed Assets Turnover Frequency	1.73E+12	Cash/Current Liability	1.28E+21
Cash/Current Liability	7.11E+11	Current Asset Turnover Rate	8.10E+12
Total assets to GNP price	2.28E+11	Operating Expense Rate	1.94E+11
Research and development expense rate	1.89E+11	Total Asset Growth Rate	1.88E+11
Total Asset Growth Rate	1.76E+11	Research and development expense rate	1.70E+11
Fixed Assets to Assets	1.32E+11	Cash Turnover Rate	1.69E+11
Net Value Growth Rate	1.09E+11	Fixed Assets to Assets	1.14E+11
Operating Expense Rate	1.08E+11	Net Value Growth Rate	8.30E+10
Cash Turnover Rate	9.00E+10	Fixed Assets Turnover Frequency	3.75E+04
Revenue per person	7.67E+10	Tax rate (A)	2.56E+02

Figure 21: Top 10 Features selected using Chi-Square Feature Selection

Synthetic Data Evaluation (Cont'd)

Feature Extraction: Variance Threshold Feature Selection

- Optimal number of features for **Real Data**: **24**
- Optimal number of features for **Synthetic Data**: **9**

Feature Selection: Real Data	Feature Selection: Synthetic Data
Operating Expense Rate	Operating Expense Rate
Research and development expense rate	Research and development expense rate
Interest-bearing debt interest rate	Total Asset Growth Rate
Revenue Per Share (Yuan ¥)	Inventory Turnover Rate (times)
Total Asset Growth Rate	Fixed Assets Turnover Frequency
Net Value Growth Rate	Cash/Current Liability
Current Ratio	Current Asset Turnover Rate
Quick Ratio	Quick Asset Turnover Rate
Total debt/Total net worth	Cash Turnover Rate
Accounts Receivable	
Turnover	
Average Collection Days	
Inventory Turnover Rate (times)	
Fixed Assets Turnover Frequency	
Revenue per person	
Allocation rate per person	
Quick Assets/Current Liability	
Cash/Current Liability	
Inventory/Current Liability	
Long-term Liability to Current Assets	
Current Asset Turnover Rate	
Quick Asset Turnover Rate	
Cash Turnover Rate	
Fixed Assets to Assets	
Total assets to GNP price	

Figure 22: Top Features selected using Variance Threshold Feature Selection

Synthetic Data Evaluation (Cont'd)

Machine Learning Model Efficacy

Real Data		Synthetic Data	
model	best_score	model	best_score
SVM	0.507173	SVM	0.529846
xgboost	0.982219	xgboost	0.892677
random_forest	0.961103	random_forest	0.855138
logistic_regression	0.59214	logistic_regression	0.540061
naive_bayes_gaussian	0.594059	naive_bayes_gaussian	0.530585
decision_tree	0.951808	decision_tree	0.812553

Figure 23: Machine Learning Model Efficacy

Conclusion & Future Works

Limitations & Future Work

- Various data types (int, decimals, categories, time, text, date) is challenging to generate.
- Some features were not learnt perfectly, needs hyperparameter tuning.

References

[1] S. I. Nikolenko *et al.*, *Synthetic data for deep learning*. Springer, 2021.

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2020.

[3] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *ArXiv*, vol. abs/1907.00503, 2019.

[4] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, 2019.

[5] H. Zhuang, X. Wang, M. Bendersky, and M. Najork, "Feature transformation for neural ranking models," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1649–1652.

[6] S. S. Dhar, B. Chakraborty, and P. Chaudhuri, "Comparison of multivariate distributions using quantile–quantile plots and related tests," *Bernoulli*, vol. 20, no. 3, pp. 1484–1506, 2014.

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *NIPS*, 2014.

[8] D. Liang and C.-F. Tsai, *Taiwanese bankruptcy prediction data set*, 2016. [Online]. Available: `https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction`.

Thanks.

Do you have any question?