National College of Ireland

Promoting Financial Inclusion: Evaluating CTGAN-Based Synthetic Alternative Data for the Credit-Invisible.

MSc Research Project
Data Analytics

Moyosoreoluwa Monsurat Ayomide Odufuwa
Student ID: x23338491

School of Computing
National College of Ireland

Supervisor: Abid Yaqoob

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Moyosoreoluwa Monsurat Ayomide Odufuwa
Student ID:	x23338491
Programme:	Data Analytics
Year:	2024/2025
Module:	MSc Research Project
Supervisor:	Abid Yaqoob
Submission Due Date:	11/08/2025
Project Title:	Promoting Financial Inclusion: Evaluating CTGAN-Based Synthetic Alternative Data for the Credit-Invisible.
Word Count:	4414
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	13th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Promoting Financial Inclusion: Evaluating CTGAN-Based Synthetic Alternative Data for the Credit-Invisible.

Moyosoreoluwa Monsurat Ayomide Odufuwa
x23338491

Abstract

The availability of high-quality credit-scoring data is essential for the development and validation of predictive models, yet privacy and regulatory constraints often limit access to such datasets. This study investigates the use of Conditional Tabular Generative Adversarial Networks (CTGAN) to generate synthetic credit-scoring data as a potential substitute for original datasets, particularly in research on alternative credit scoring for credit-invisible populations. The proposed approach involves minimal preprocessing of the original dataset to maintain fidelity, followed by CTGAN training to capture complex feature distributions and dependencies. The resulting synthetic dataset is evaluated for statistical similarity, predictive utility, and fairness. Results show high Pearson (0.8846) and Spearman (0.8335) correlation similarities, with tree-based classifiers trained on synthetic data achieving competitive ROC-AUC scores compared to those trained on real data. However, fairness metrics reveal that CTGAN reproduces much of the bias present in the original dataset. The findings suggest that while CTGAN-generated synthetic data holds strong promise as a privacy-preserving resource for academic research and early-stage prototyping, it cannot fully replace original datasets in production-grade applications without targeted efforts to address residual distributional gaps and embedded bias.

1 Introduction

Access to credit is a cornerstone of financial inclusion, yet a significant proportion of the global population remains excluded due to the absence of sufficient credit history - a phenomenon commonly referred to as credit invisibility. CredoLab[1] estimates almost 3.5 billion individuals globally are unbanked, with many lacking the credit records required by traditional scoring systems. Conventional credit scoring methods, such as those employed by FICO [2] and similar agencies, rely heavily on historical repayment behaviours and established financial activity. While effective for individuals with mature credit histories, these models systematically disadvantage thin-file borrowers - individuals with limited

[1] https://www.credolab.com/ebook/analysing-credit-invisibles-know-who-they-are-and-how-to-help-them#:~:text=Get%20to%20know%20who%20credit,solutions%20to%20financial%20including%20them.&text=Almost%203.5%20billion%20people%20in,most%20of%20them%20are%20unbanked

[2] https://www.myfico.com/credit-education/what-is-a-fico-score

or no recorded credit data - including young adults, recent immigrants, and those from certain sectors of the population[3].

The growing interest in alternative data - information outside the traditional credit bureau scope, such as transactional and behavioral data including mobile and social footprint (Agarwal *et al.*, 2023), email usage and psychometric data (Djeundje *et al.*, 2021) - has shown promise in addressing this disparity. However, advancing research in alternative credit scoring faces a critical barrier: access to real-world alternative data. Such datasets are often proprietary[4], privacy-sensitive, and subject to strict data protection regulations, including the General Data Protection Regulation (GDPR) in the EU and the Fair Credit Reporting Act (FCRA) in the US. These restrictions limit the ability of researchers and practitioners to develop, test, and validate innovative credit risk models on diverse populations.

Synthetic data - data derived from (original) data - and its generation has emerged as a potential solution to this challenge. By simulating datasets that statistically resemble real data without exposing individual records, synthetic data offers a path to both privacy preservation and research scalability(Jordon *et al.*, 2018). Among the various techniques, Conditional Tabular Generative Adversarial Networks (CTGAN) stand out as a state-of-the-art approach designed specifically for tabular datasets with mixed data types and imbalanced distributions (Xu *et al.*, 2019). CTGAN leverages adversarial training and conditional generation to capture complex relationships between features while allowing controlled sampling of underrepresented categories - a crucial capability in credit scoring contexts where datasets are heavily imbalanced.

This study seeks to address the following research question:

Is CTGAN-generated synthetic credit data able to match the predictive performance, fairness characteristics, and data fidelity of real-world alternative credit datasets, and thus potentially replace them in alternative data credit scoring research for credit-invisible individuals? The research objectives are therefore to:

- Generate synthetic credit data using CTGAN, trained on a real-world alternative credit dataset.
- Build and compare machine learning models - Logistic Regression, Random Forest, and XGBoost - on three scenarios: original data, combined (original and synthetic) data, and synthetic-only data.
- Evaluate the models on discrimination ability (ROC-AUC and related metrics), fairness across protected attributes, and calibration.
- Assess the synthetic data's fidelity, diversity, and utility using statistical tests (Kolmogorov–Smirnov, correlation similarity), feature importance stability, and Train-on-Synthetic-Test-on-Real (TSTR).

The contribution of this work lies in providing a practical, multi-dimensional assessment of CTGAN's ability to generate research-grade synthetic credit data, with an emphasis on high-stakes applications such as alternative credit scoring for underserved

[3]https://www.businessinsider.com/personal-finance/credit-score/what-is-credit-invisibility

[4]https://riskseal.io/blog/top-alternative-data-providers-that-serve-the-credit-industry

populations. By integrating model utility, fairness assessment, and data quality evaluation, this study addresses a gap in the literature, where prior work has often focused on either model accuracy or data augmentation, but not the viability of synthetic data as a substitute for real data in credit risk research.

The remainder of this report is structured as follows: Literature review which examines relevant literature on credit invisibility, alternative data, GAN-based synthetic data generation, and evaluation metrics for synthetic data quality. Research methodology which maps the project steps to the CRISP-DM framework. Design specification and implementation presents the CTGAN model and the model training pipelines. Evaluation which highlights the results, comparing predictive performance, fairness outcomes, and data fidelity across datasets and finally the report ends with the conclusion on key findings, implications for research and practice, and potential directions for future work.

2 Related Work

2.1 Credit Invisibility and Alternative Data in Credit Scoring

Hlongwane *et al.* (2024), having used the Kaggle Home Credit dataset —a real dataset released by a lending institution (Home Credit) with non-identifying customer data—, found that incorporating alternative data such as social circle default rates, regional economic ratings, and telco information improved the model's AUC ROC score to 0.79360, outperforming models built solely on bureau data. Notably, around 30% of their sample had no credit bureau records, yet could still be effectively scored using these features. Nonetheless, such datasets are rarely made public due to privacy, competitive, and regulatory constraints, and without Home Credit's voluntary release, researchers would lack access to this valuable resource. This underlines the potential of synthetic data generation, such as CTGAN, to create realistic yet privacy-preserving datasets, enabling broader research into alternative credit scoring for the credit invisible.

Similarly, Ngwenya (2024) addressed the challenge of credit invisibility in Africa, where large portions of the population remain excluded from financial services due to the absence of formal credit histories. The study proposes leveraging alternative data from Mobile Network Operators (MNOs)—including call frequency, data usage, and mobile money transactions—to build a logistic regression–based credit scoring model. By focusing on behavioural and transactional patterns, the model achieved up to 90% prediction accuracy, outperforming traditional approaches. A key insight is the transformative role of mobile money and GSM data as proxies for financial behaviour, enabling financial institutions to extend credit to the unbanked. However, the paper acknowledges significant challenges in data privacy, regulatory compliance, and data access—particularly as telecommunications providers are often reluctant to share datasets. This reinforces the value of synthetic data generation in research contexts, as demonstrated in our own study: without proprietary datasets, such as those used in this research, academic exploration would be severely constrained.

2.2 Synthetic Data in Financial Services

The growing reliance on data-driven decision-making in financial services is often hindered by limited access to high-quality datasets due to privacy concerns, regulatory restrictions, and competitive sensitivities. Synthetic data generation has emerged as a promising

solution to overcome these barriers, enabling model development, testing, and research without exposing sensitive customer information.

Authors Caliskan *et al.* (2023) explored the application of generative models—specifically Variational Autoencoders (VAE) and Conditional Tabular GANs (CTGAN)—to produce synthetic credit-related datasets. The aim of the study was to assess the models' ability to preserve the statistical properties of real loan offer data while ensuring privacy. Both approaches were evaluated using fidelity metrics, statistical similarity tests, and downstream machine learning performance comparisons. Results indicated that synthetic data generated using CTGAN was of a slightly better quality than that of the custom VAE model in visual, statistical and metric evaluation. This finding underscores CTGAN's suitability for generating high-quality synthetic financial data, making it a strong candidate for scenarios where access to real data is restricted due to privacy or proprietary constraints.

A recent study by Manasa *at al.*(2024) evaluated the application of advanced GAN architectures—TimeGAN, Conditional GAN (CGAN), and Wasserstein GAN (WGAN)—to generate high-quality synthetic financial data. The motivation stems from challenges in accessing real financial datasets due to privacy concerns, regulatory restrictions, and scarcity. By learning complex temporal and statistical patterns from real data, these GANs produce synthetic datasets closely matching real data in key metrics such as mean, variance, skewness, and kurtosis. The paper emphasizes the use of GAN-generated data for privacy preservation and data augmentation across multiple financial applications, including algorithmic trading, fraud detection, risk modelling, and credit scoring. Performance evaluations revealed minimal accuracy drops (less than 2%) compared to models trained on real data, supporting the feasibility of synthetic data as a replacement in sensitive contexts. While challenges such as mode collapse and training instability persist, architectures like WGAN and TimeGAN demonstrated improved stability and temporal feature capture. This work aligns closely with the current study's focus by reinforcing that well-trained GAN models can produce statistically and functionally comparable datasets for financial analytics, enabling research and innovation without compromising confidentiality.

Another comparative study by Ramachandra and Vaithiyanathan (2025) presented Fed-DPSDG-WGAN, a framework combining federated learning, differential privacy, and Wasserstein GANs to generate privacy-preserving synthetic financial data for loan default prediction. The system enabled multiple institutions to collaboratively train models without sharing raw data, ensuring compliance with privacy regulations. Within the process, CTGAN was applied to address class imbalance in the synthetic dataset — a common challenge in credit risk datasets where default cases are under-represented. This ensured the generated data remained representative of both majority and minority classes, improving downstream predictive model performance. Experimental results show that the approach maintained strong privacy guarantees while producing synthetic datasets that closely match real-world distributions, supporting effective loan default prediction without exposing sensitive records.

3 Research Methodology

This study followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, a widely adopted methodology for structuring analytical projects[5] This choice was made due to CRISP-DM's iterative, flexible nature, which is well-suited for both exploratory data analysis and rigorous model evaluation. The process comprised the following sequential stages: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

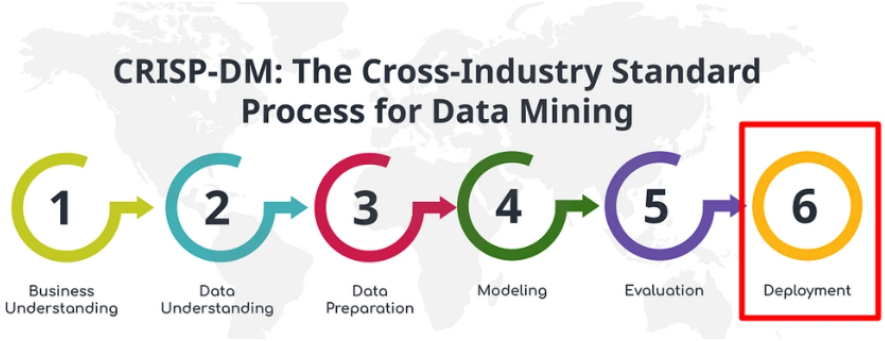


Figure 1: CRISP-DM Methodology

3.1 Business Understanding

The central objective guiding this study is to assess whether CTGAN-generated synthetic data could effectively replace original financial datasets to enable further research into alternative data credit scoring for credit-invisible individuals. The motivation stemmed from the limited availability of high-quality, real-world credit datasets due to privacy concerns and regulatory restrictions (e.g., GDPR). By focusing on the practical application of CTGAN in this domain, the study aligns with current industry needs surrounding privacy, data availability, and fairness in lending.

3.2 Data Understanding

The dataset used in this research was sourced from the Home Credit Default Risk dataset, available on Kaggle (Montoya *et al.*, 2018). Home Credit is an international non-bank financial institution specializing in instalment lending, particularly for individuals with little or no credit history. The dataset was initially shared as part of a Kaggle competition to encourage research in inclusive credit scoring. For this study, the test set from the main application data table was used. The dataset consists of 307,511 records and 122 variables, including demographic data, financial history, external risk scores, and behavioural patterns such as previous loan repayment records. The target variable in this dataset is TARGET, where a value of 1 indicates a default on a loan, and 0 indicates successful repayment. Initial exploration using pandas, matplotlib, and seaborn revealed a notable class imbalance as seen in figure 2: around 92% of the applicants belong to the non-default class, and the remaining 8% represent defaults.

[5] `https://www.datascience-pm.com/crisp-dm-2/:~:text=Some%20argue%20that%20it%20(%20CRoss%20Industry,really%20matters%20is%20how%20you%20implement%20it.`

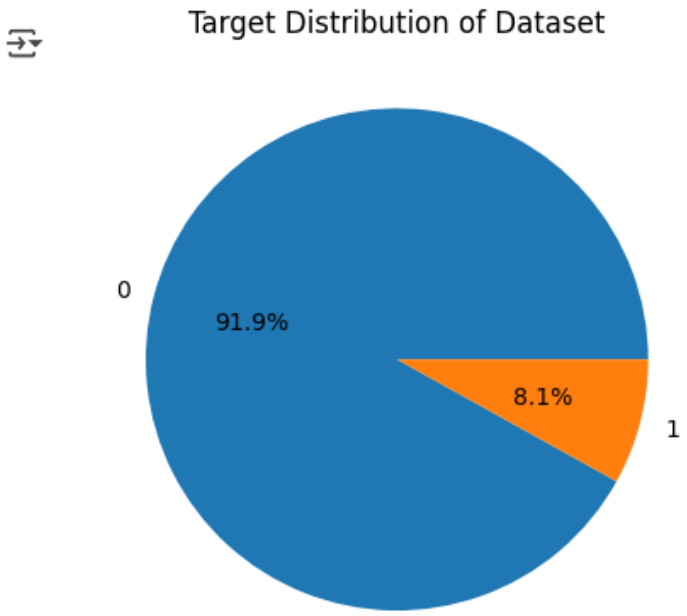


Figure 2: Target Distribution

3.3 Data Preparation

Data preprocessing was very minimal and structured into distinct steps to maintain the primary distribution and ensure compatibility with the CTGAN model and downstream machine learning pipelines.

- Data Reduction: The sample size was reduced to 150,000 records to ensure training was computationally feasible within resource limits while maintaining statistical representativeness
- Data Splitting: A 70/30 (105,000/45,000) train-test split was performed early in the pipeline to prevent data leakage. All preprocessing steps were restricted to the training set unless explicitly required otherwise.
- Feature Selection: To reduce dimensionality and enhance model performance, feature importance was evaluated using the SelectKBest method from sklearn.feature_selection with mutual_info_classif as the scoring function. This approach ranks features based on their estimated mutual information with the target variable — in this case, TARGET. Mutual information quantifies how much knowing the value of a feature reduces uncertainty about the target variable. A threshold score of 0.01 was set to eliminate features with minimal contribution to the classification task, resulting in a final selection of 46 key variables. This ensured that only the most relevant and informative features were retained for modelling.
- Missing Values: Missing values in selected features were handled using SimpleImputer from sklearn.impute library, applying median imputation for numerical features and most frequent imputation for categorical features to remain as close as possible to the original distribution.

- CTGAN Compatibilty: No data-altering preprocessing (e.g., normalization or scaling) was applied before training the CTGAN model, to preserve the original distribution. Specific CTGAN requirements were adhered to during formatting:
  - Categorical variables were encoded as strings or integers.
  - Continuous variables were cast as floats.
  - A discrete feature list was explicitly passed to the model for accurate learning.
- Synthetic Data Generation: CTGAN was chosen for its ability to model complex distributions in mixed-type tabular data and handle class imbalance through conditional sampling. Key hyperparameters included embedding size, batch size, and training epochs, were tuned to balance training stability and quality of generated samples.

3.4 Modelling

- Predictive Modelling: For evaluation three classification algorithms were chosen based on their efficiency in previous related work: logistic regression (LR), Random Forest(RF) and XGBoost (XGB). Three training datasets were compared:
  - Baseline - Real training data
  - Synthetic - CTGAN-generated synthetic only data
  - Combined - Real + Synthetic training data

3.5 Evaluation

The evaluation strategy combined fidelity, utility, and fairness assessments:

- Data Fidelity:
  - Kolmogorov–Smirnov (KS) statistic to quantify distributional similarity between real and synthetic features.
  - Pairwise correlation (Pearson and Spearman) similarity comparing correlation matrices from real vs synthetic datasets.
- Predictive Ability:
  - Train on Synthetic, Test on Real (TSTR) framework to assess cross-domain generalisation.
  - Standard metrics on the real test set: ROC-AUC, PR-AUC, accuracy, precision, recall, and F1-score.
- Fairness Analysis: Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), and Average Odds Difference (AOD) computed for protected attributes such as income type, education type, and housing type.
- Feature Importance Stability: Spearman rank correlation of feature importances between models trained on real vs synthetic data.

3.6 Deployment

The final deployment phase of the CRISP-DM framework, which involves integrating the developed model or system in a live production environment was intentionally omitted in this study.

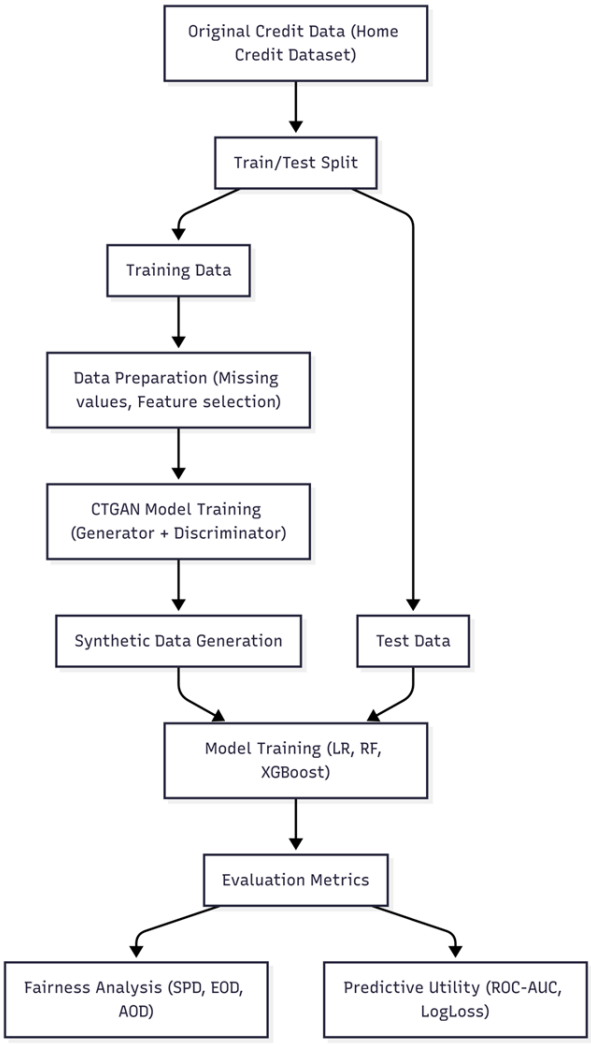


Figure 3: Research Methodology

4 Design Specification and Implementation

This project investigates whether synthetic tabular datasets, generated by the Conditional Tabular Generative Adversarial Network (CTGAN), can serve as a viable replacement for original credit scoring datasets in research on alternative data credit scoring for credit-invisible populations. The design emphasizes preserving the original statistical properties of the dataset while enabling robust model evaluation on real-world credit risk prediction tasks. The system pipeline comprises three primary stages: minimal preprocessing, synthetic data generation with CTGAN, and downstream predictive modelling with three classifiers, followed by a rigorous evaluation of model performance and data fidelity.

A deliberate choice was made to apply only the necessary preprocessing steps before training the CTGAN model. This decision was based on the aim of preserving the original data distribution and structure as much as possible, avoiding transformations that could introduce artificial bias or distort feature relationships. Handling of missing data was performed using SimpleImputer, with median replacement for numerical features and most frequent category replacement for categorical features. Class imbalance was not addressed in preprocessing so that no synthetic oversampling or under-sampling altered the base distribution of the target variable prior to CTGAN training. This approach ensures that any differences between real and synthetic data distributions arise from the generative process rather than from pre-generation manipulation.

4.1 Conditional Tabular Generative Adversarial Network (CTGAN)

CTGAN (Conditional Tabular GAN) is a specialized generative model designed to synthesise realistic mixed-type tabular data, overcoming the limitations of traditional GANs in handling non-continuous, non-Gaussian feature spaces (Xu et al., 2019). Like all GANs, CTGAN is grounded in the adversarial learning framework first introduced by Goodfellow et al. (2014), where a generator G and a discriminator D are trained simultaneously in a minimax game: G aims to produce samples identical to the real data, while D learns to correctly classify samples as real or synthetic.

Two key innovations make CTGAN suitable for credit-scoring data. First, it employs mode-specific normalisation for continuous variables. These are modelled as mixtures of Gaussians, allowing the generator to learn complex, multi-modal shapes — a critical capability when modelling financial variables such as income, ratios, and age, which often deviate significantly from normality. This approach addresses a common shortcoming in conventional GANs, which assume uni-modal Gaussian distributions (Mirza and Osindero, 2014). Second, it uses conditional sampling over discrete columns: at each step, a discrete column and one of its categories are sampled to form a condition vector concatenated to the generator's input. This targeted conditioning creates adequate coverage of minority categories.

Loss Function

During training, each network learns how to improve after each training iteration (epoch) by the loss function. The generator and discriminator each maintain their own loss values, updated iteratively over the course of training. At each epoch, both networks aim to minimise their respective loss functions, but with opposing objectives: the discriminator seeks to maximize its ability to distinguish between real and synthetic data, while the generator attempts to produce synthetic data that the discriminator classifies as real.

$$\mathcal{L}_D = \frac{1}{m}\sum_{i=1}^{m}[D(\mathrm{x}'^{(i)}) - D(\mathrm{x}^{(i)})]$$
$$\mathcal{L}_G = -\frac{1}{m}\sum_{i=1}^{m}[D(\mathrm{x}'^{(i)})] + H$$

Figure 4: Loss Function

Let x denote an original sample, x′ a synthetic sample and D(x) the discriminator's output given the real data and D(x′) for the synthetic data. H is the cross entropy score and is positive.

The discriminator's loss encourages D(x) to be high for real samples and D(x′) low for synthetic. In the opposite, the generator's loss is minimised when D(x′) is high indicating that synthetic sample are being classified as real[6](Xu *et al.*, 2019).

Loss Value Scenarios

- Positive Generator Loss - Zero Discriminator Loss: Reflects an intial phase where both networks are under optimised with the generator creating poor quality synthetic data and the discriminator assuming which is real or not.
- Increasing Negative Generator Loss - Zero Discriminator Loss: As training progresses, there is improved synthetic data quality from the generator but slight improvement from the discriminator.
- Generator Loss Stabilizes at Negative - Zero Discriminator Loss: The final stage, where the generator has successfully generated synthetic data the discriminator cannot reliably differentiate from the original data [7].

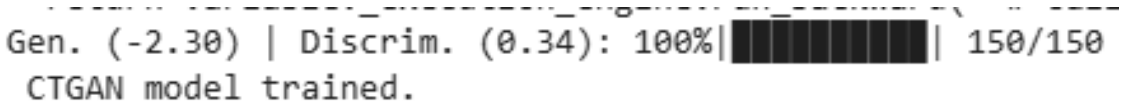


Figure 5: CTGAN Model Loss Values

The resulting loss values at the end of training the CTGAN model is reported in figure 5 with a Generator loss -2.30 of and Discriminator loss of 0.34 after 150 epochs. This result translates to the generator producing high-quality synthetic samples and the discriminator being challenged at differentiating between real and fake.

In this implementation, the CTGAN model was configured with:

- epochs=150 — the number of full passes of the Generator and Discriminator over the training data.
- generator_dim=(256, 256) — the size of output samples for each of the residuals. A Residual Layer will be created for each one of the values provided.

[6]https://github.com/sdv-dev/SDV/discussions/980
[7]https://sdv.dev/blog/interpreting-ctgan-progress/

- discriminator_dim=(256, 256) — the size of the output samples for each one of the Discriminator layers. A Linear Layer will be created for each one of the values provided.
- generator_lr=0.0002 and discriminator_lr=0.0002 — learning rates for the generator and discriminator respectively.
- verbose=True — enabling training progress outputs [8].

The final outputs from this stage were:

- A trained CTGAN object.
- A synthetic dataset matching the training set in schema and number of rows.

```
NAME_HOUSING_TYPE

 Original Data: ['House / apartment' 'Municipal apartment' 'With parents'
 'Rented apartment' 'Office apartment' 'Co-op apartment']

 Synthetic Data: ['House / apartment' 'Office apartment' 'With parents'
 'Municipal apartment' 'Rented apartment' 'Co-op apartment']

FLAG_OWN_REALTY

 Original Data: ['Y' 'N']

 Synthetic Data: ['Y' 'N']

NAME_EDUCATION_TYPE

 Original Data: ['Secondary / secondary special' 'Incomplete higher' 'Higher education'
 'Lower secondary' 'Academic degree']

 Synthetic Data: ['Higher education' 'Secondary / secondary special' 'Incomplete higher'
 'Academic degree' 'Lower secondary']

NAME_TYPE_SUITE

 Original Data: ['Unaccompanied' 'Other_A' 'Children' 'Spouse, partner' 'Family' 'Other_B'
 'Group of people']

 Synthetic Data: ['Unaccompanied' 'Spouse, partner' 'Family' 'Other_A' 'Other_B' 'Children'
 'Group of people']
```

Figure 6: Original vs synthetic categorical features and its values

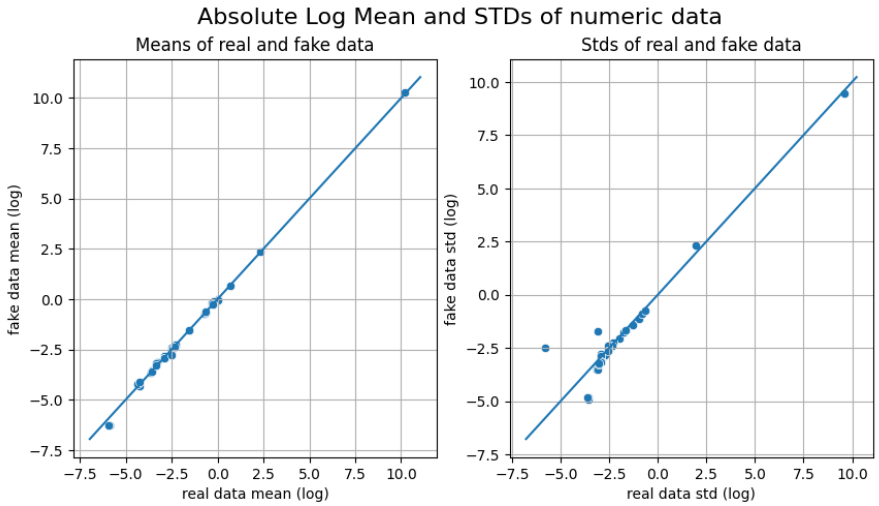
In figure 6, we can see that CTGAN generated synthetic values for categorical features that completely mirrored the ones of each original dataset.

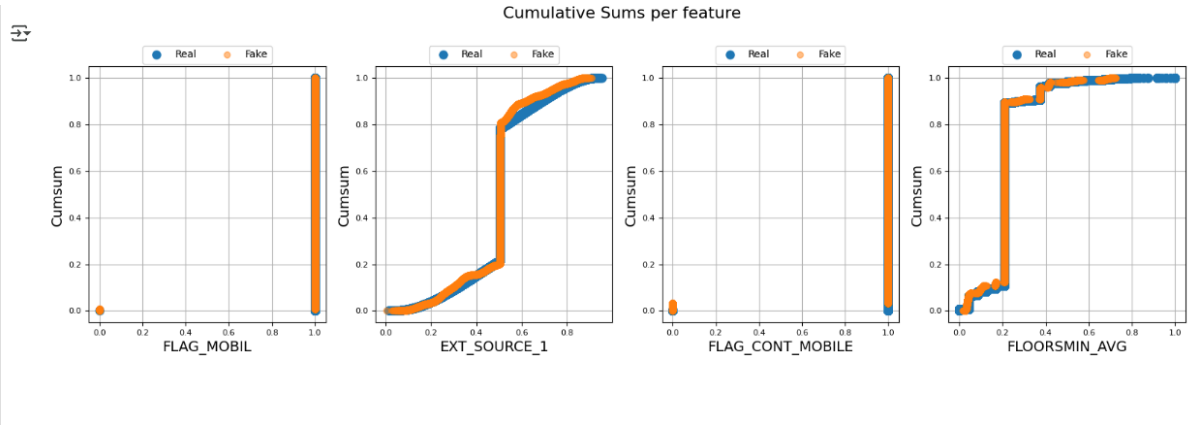
4.2 Synthetic vs Original Quality Comparison

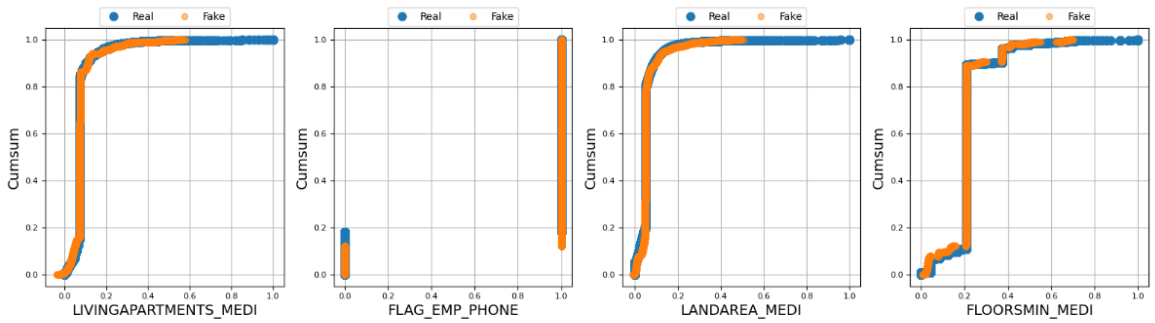
The TableEvaluator library was used to compare each feature's empirical distribution and to visualise pairwise relationships between columns. The feature-by-feature comparison as highlited in figure 7, delivers for every column, overlaid histograms/density plots (for continuous variables) or bar charts (for categorical variables), allowing a direct visual inspection of how closely synthetic distributions trace their real counterparts. The tool

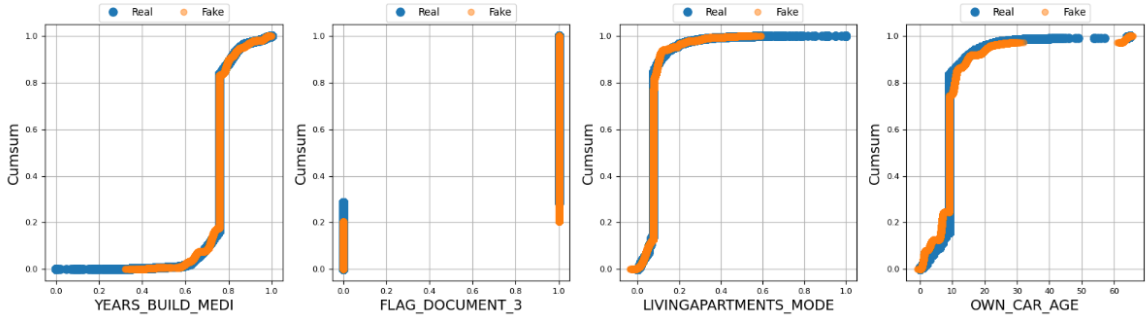
[8] https://docs.sdv.dev/sdv/single-table-data/modeling/synthesizers/ctgansynthesizer

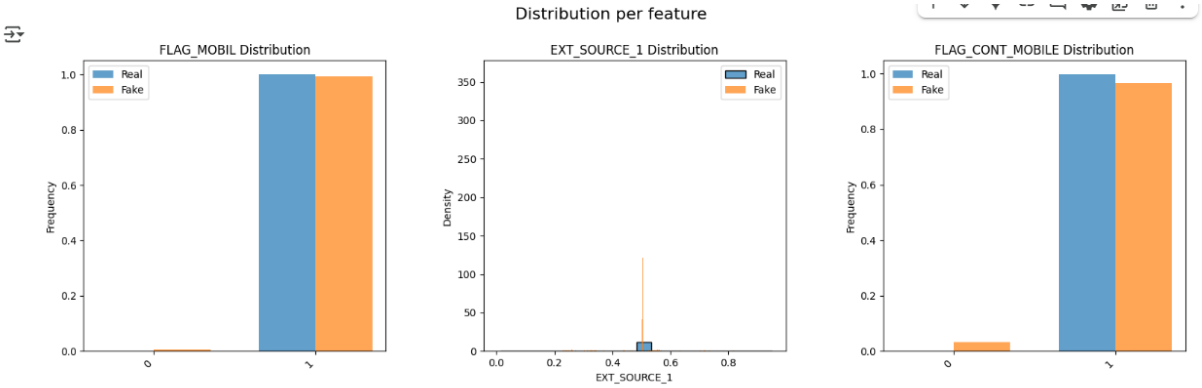
also provides correlation heatmaps for both real and synthetic tables, which is crucial in credit data. In this project, these plots showed strong alignment on most features, with a small number of variables exhibiting modest deviations in tail behaviour or category prevalence—typical pressure points for tabular GANs.

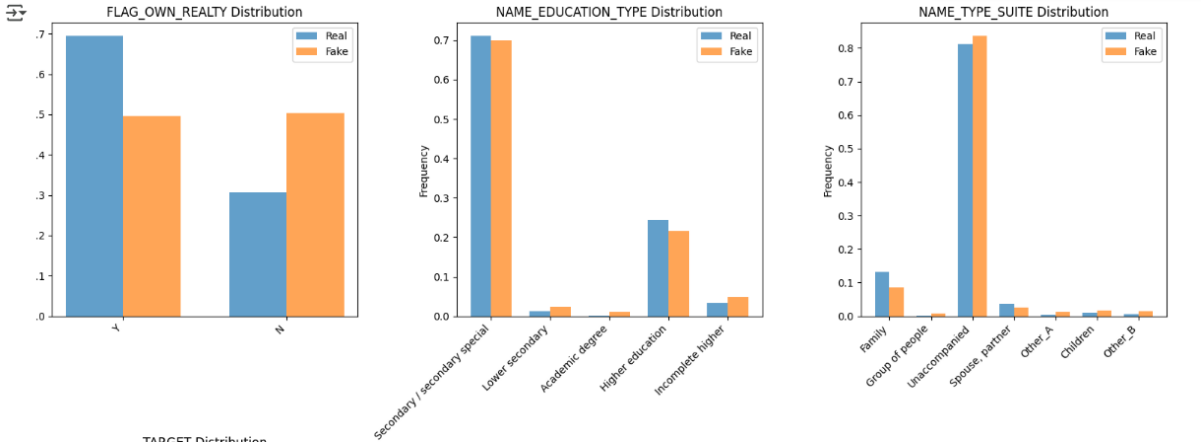


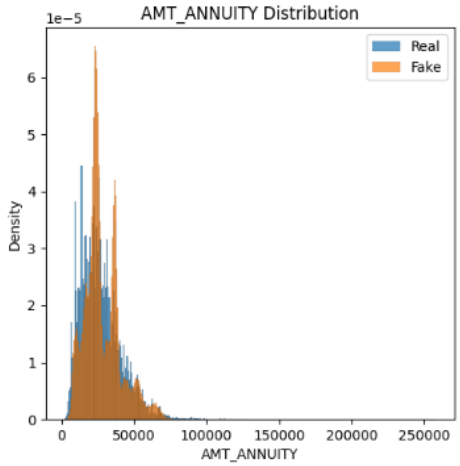


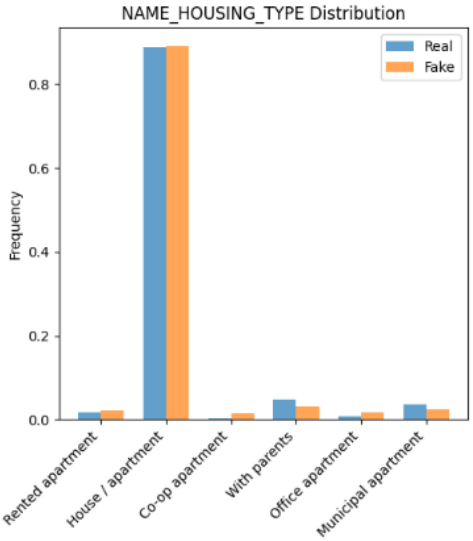


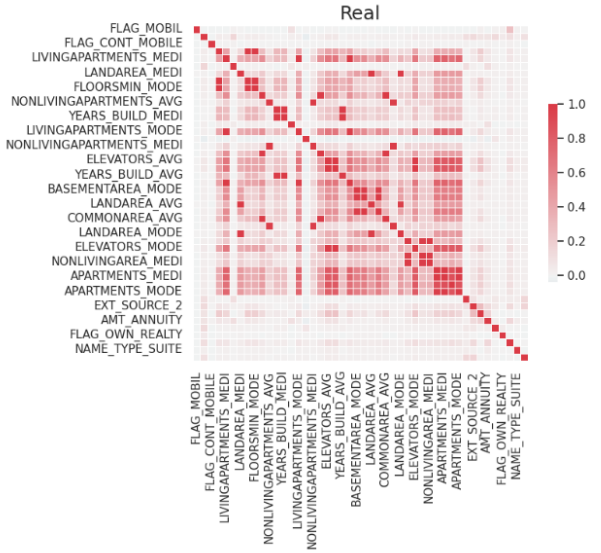

Real
Fake
Cumsum
YEARS_BUILD_MEDI
FLAG_DOCUMENT_3
LIVINGAPARTMENTS_MODE
OWN_CAR_AGE


Distribution per feature
FLAG_MOBIL Distribution
EXT_SOURCE_1 Distribution
FLAG_CONT_MOBILE Distribution
Real
Fake
Frequency
Density
EXT_SOURCE_1

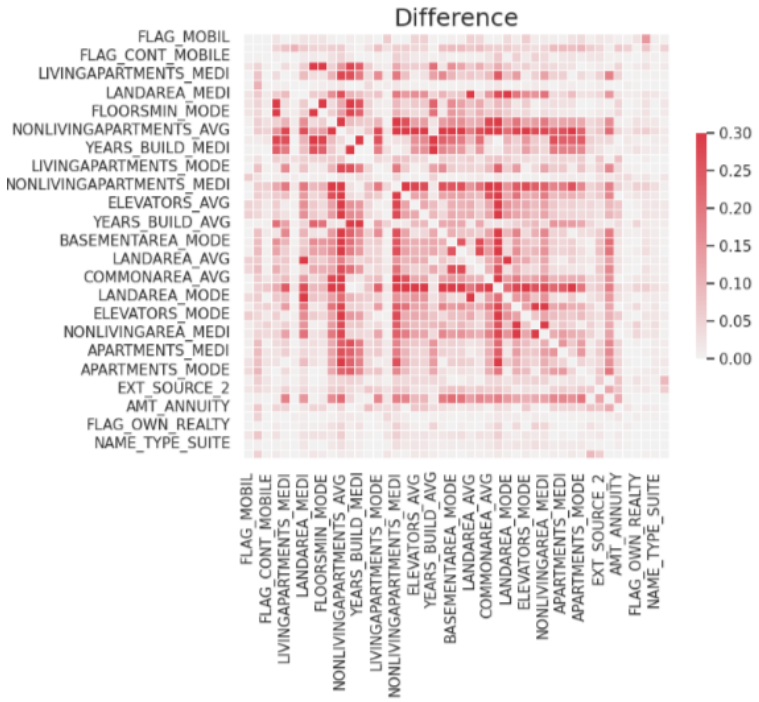

FLAG_OWN_REALTY Distribution
NAME_EDUCATION_TYPE Distribution
NAME_TYPE_SUITE Distribution
Real
Fake
Frequency
Y
N
Secondary / secondary special
Lower secondary
Academic degree
Higher education
Incomplete higher
Family
Group of people
Unaccompanied
Spouse, partner
Other_A
Children
Other_B

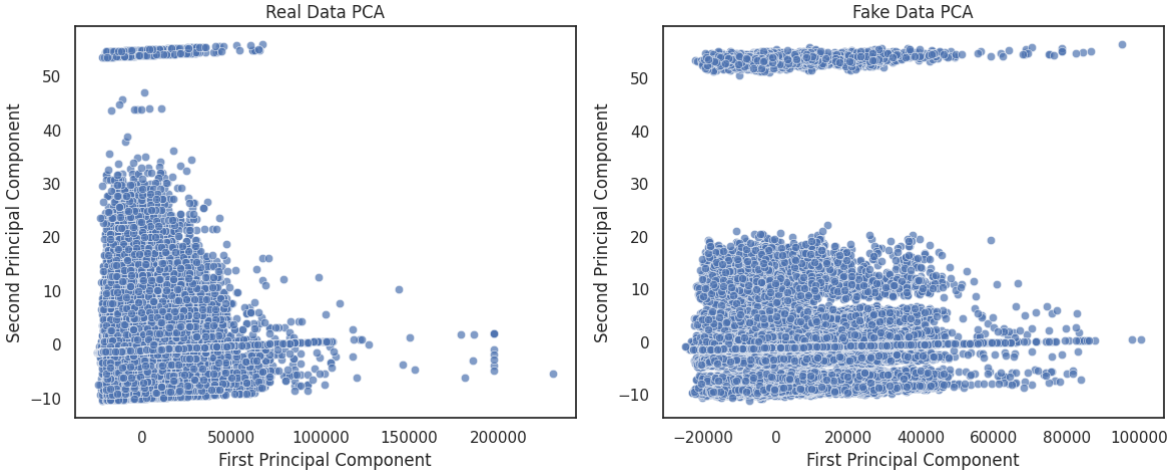

AMT_ANNUITY Distribution
1e−5
Real
Fake
Density
AMT_ANNUITY
0
50000
100000
150000
200000
250000


NAME_HOUSING_TYPE Distribution
Real
Fake
Frequency
Rented apartment
House / apartment
Co-op apartment
With parents
Office apartment
Municipal apartment


Real
FLAG_MOBIL
FLAG_CONT_MOBILE
LIVINGAPARTMENTS_MEDI
LANDAREA_MEDI
FLOORSMIN_MODE
NONLIVINGAPARTMENTS_AVG
YEARS_BUILD_MEDI
LIVINGAPARTMENTS_MODE
NONLIVINGAPARTMENTS_MEDI
ELEVATORS_AVG
YEARS_BUILD_AVG
BASEMENTAREA_MODE
LANDAREA_AVG
COMMONAREA_AVG
LANDAREA_MODE
ELEVATORS_MODE
NONLIVINGAREA_MEDI
APARTMENTS_MEDI
APARTMENTS_MODE
EXT_SOURCE_2
AMT_ANNUITY
FLAG_OWN_REALTY
NAME_TYPE_SUITE
1.0
0.8
0.6
0.4
0.2
0.0


Fake
FLAG_MOBIL
FLAG_CONT_MOBILE
LIVINGAPARTMENTS_MEDI
LANDAREA_MEDI
FLOORSMIN_MODE
NONLIVINGAPARTMENTS_AVG
YEARS_BUILD_MEDI
LIVINGAPARTMENTS_MODE
NONLIVINGAPARTMENTS_MEDI
ELEVATORS_AVG
YEARS_BUILD_AVG
BASEMENTAREA_MODE
LANDAREA_AVG
COMMONAREA_AVG
LANDAREA_MODE
ELEVATORS_MODE
NONLIVINGAREA_MEDI
APARTMENTS_MEDI
APARTMENTS_MODE
EXT_SOURCE_2
AMT_ANNUITY
FLAG_OWN_REALTY
NAME_TYPE_SUITE
1.0
0.8
0.6
0.4
0.2
0.0


Difference
FLAG_MOBIL
FLAG_CONT_MOBILE
LIVINGAPARTMENTS_MEDI
LANDAREA_MEDI
FLOORSMIN_MODE
NONLIVINGAPARTMENTS_AVG
YEARS_BUILD_MEDI
LIVINGAPARTMENTS_MODE
NONLIVINGAPARTMENTS_MEDI
ELEVATORS_AVG
YEARS_BUILD_AVG
BASEMENTAREA_MODE
LANDAREA_AVG
COMMONAREA_AVG
LANDAREA_MODE
ELEVATORS_MODE
NONLIVINGAREA_MEDI
APARTMENTS_MEDI
APARTMENTS_MODE
EXT_SOURCE_2
AMT_ANNUITY
FLAG_OWN_REALTY
NAME_TYPE_SUITE
0.30
0.25
0.20
0.15
0.10
0.05
0.00


Real Data PCA
Second Principal Component
First Principal Component
Fake Data PCA
Second Principal Component
First Principal Component

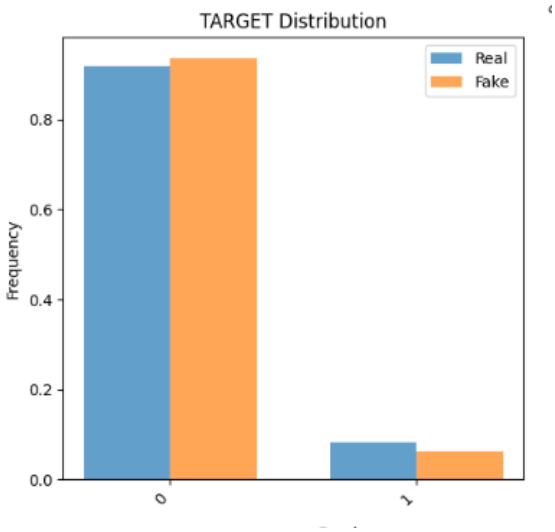


Figure 7: Overlay of empirical distribution between real vs fake data

Quantitative comparison The TableEvaluator evaluate() method was then used to summarise similarity between the real and synthetic tables with a compact set of scores. The run as shown in figure 8 reported:

```
Classifier F1-scores and their Jaccard similarities::
    real   fake
0 0.9893 0.9893

Privacy results:
                                    result
Duplicate rows between sets (real/fake)  (0, 0)
nearest neighbor mean                    2.4292
nearest neighbor std                     2.5357

Miscellaneous results:
                               Result
Column Correlation Distance RMSE  0.1002
Column Correlation distance MAE   0.0732

Results:
                                                result
Basic statistics                                0.9957
Correlation column correlations                 0.8882
Mean Correlation between fake and real columns  0.7253
1 - MAPE Estimator results                      0.9893
Similarity Score                                0.8996
{'Basic statistics': 0.995724654391117,
 'Correlation column correlations': 0.8881631907915041,
 'Mean Correlation between fake and real columns': 0.7252828005925392,
 '1 - MAPE Estimator results': 0.9892602347504849,
 'Similarity Score': 0.8996077201314112}
```

Figure 8: TableEvaluator summary results

- Basic statistics: 0.9957 — near-perfect alignment of univariate statistics (e.g., means, standard deviations, skew), indicating that marginal distributions were learned very closely.
- Correlation column correlations: 0.8882 — high agreement between column-wise correlation structures in real vs synthetic data; important for preserving model-relevant relationships.
- Mean correlation between fake and real columns: 0.7253 — an average measure of dependency preservation; the value reflects good but not perfect recovery of multivariate structure, consistent with the visual heatmaps.

- 1 MAPE estimator results: 0.9893 — high agreement when evaluating a simple estimator trained to map features to the target, suggesting that predictive signals present in the real data are substantially captured by the synthetic sample.
- Similarity Score: 0.8996 — an overall, composite similarity measure landing in the "strong" range for tabular synthesis.

The evaluation also printed privacy checks (e.g., Duplicate rows between sets (real/fake): (0, 0)) and Classifier F1-scores and their Jaccard similarities: (real 0.9893; fake 0.9893). Together, these results provide evidence that the CTGAN sample retained the key statistical signatures and label relationships needed for downstream modelling, while not memorising rows from the source dataset.

4.3 Predictive Modelling

To assess the usefulness of the generated data, three classifiers were selected; Logistic Regression, Random Forest and XGBoost.Each model was implemented within a Pipeline containing the same preprocessing steps as used for the real dataset, ensuring consistent feature transformation between training regimes. Hyperparameters were tuned using stratified k-fold cross-validation with randomised search to mitigate bias from fixed parameter choices.

All implementation were conducted in Python within Google Colab. Core packages included scikit-learn for preprocessing and modelling, CTGAN for synthetic data generation, TableEvaluator for fidelity assessment, pandas/NumPy for data handling, and Matplotlib/Seaborn for visualisation.

5 Evaluation

This section analyses the experimental findings, focusing on the degree to which CTGAN-generated synthetic data replicates the statistical and predictive properties of the original credit-scoring dataset, as well as the fairness implications of its use in downstream classification models. The evaluation framework follows two complementary dimensions:

- Fidelity assessment, measuring the statistical similarity between synthetic and real datasets.
- Utility assessment, comparing model performance and fairness metrics when trained on synthetic data versus the original data.

5.1 Fidelity Assessment

The statistical integrity of the synthetic dataset was evaluated using the TableEvaluator library, which conducts a feature-by-feature comparison between the original and synthetic datasets.

For continuous variables, the Kolmogorov–Smirnov (KS) statistic (presented in figure 9) was used. KS statistic is a non-parametric test used to compare two probability distributions. It quantifies the distance between their cumulative distribution functions (CDFs) and helps determine if the samples come from the same distribution (Cui et al.,

2025). A small p-value(typlically less than 0.05) indicates a compelling difference between the distributions and a large p-value suggests otherwise[9].

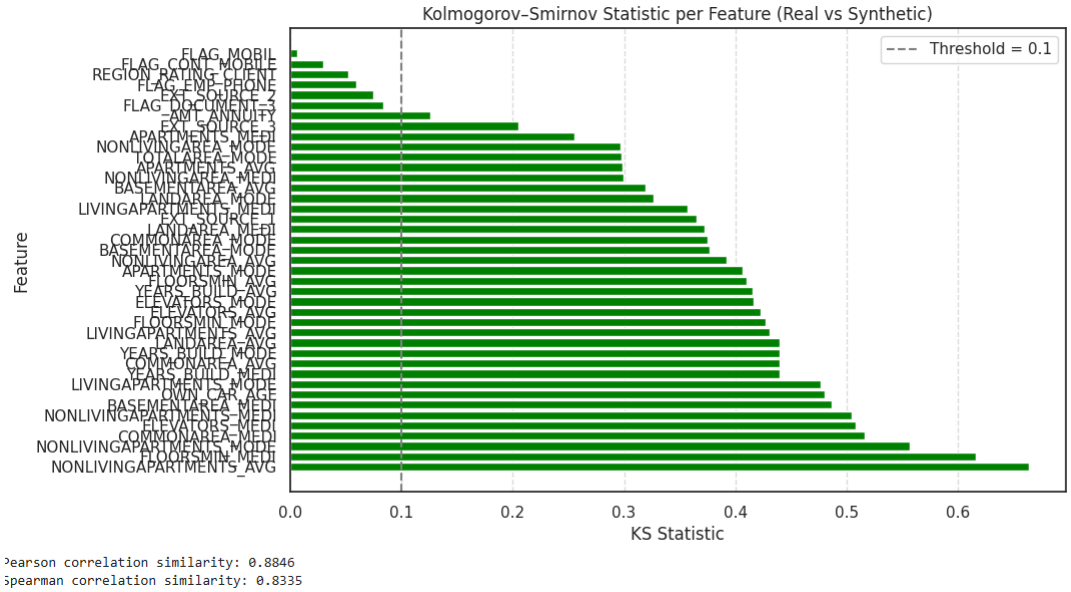


Figure 9: Kolmogorov–Smirnov (KS) statistic results

The Pearson correlation similarity also known as the Pearson correlation coefficient is used to measure the linear relationship between two datasets with a value of 1 being strong positive linear relationship, 0 being no relationship and -1 indicating negative relationship (Mana and Sasipraba, 2021). Between the real and synthetic data the result was 0.8846 as seen in figure 10.

A Spearman correlation similarity score of 0.8335, demmonstrates a strong monotonic i.e strength and direction relationship between two ranked variables [10].

```
Pearson correlation similarity: 0.8846
Spearman correlation similarity: 0.8335
```

Figure 10: Pearson Spearman Correlation Similarity results

In addition the feature-importance rank correlation (Spearman) between real and synthetic datasets was 0.2164. This fair correlation implies that while CTGAN replicated many feature-importance relationships observed in the original dataset, some re-ordering of predictive relevance occurred, potentially impacting model interpretability.

5.2 Predictive Utility

Predictive ability was assessed by training Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB) classifiers on three datasets:

[9]https://medium.com/@data-overload/the-kolmogorov-smirnov-test-56e18ba621de

[10]https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide-2.php

- Baseline: Real data only
- Combined: Real + synthetic data
- Synthetic: Synthetic data only

	Dataset	Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	LogLoss	Brier
0	Baseline	LR	0.6774	0.1536	0.6656	0.2496	0.7317	0.2091	0.6112	0.2108
1	Baseline	RF	0.7197	0.1600	0.5831	0.2511	0.7191	0.2038	0.5708	0.1928
2	Baseline	XGB	0.6847	0.1566	0.6639	0.2534	0.7402	0.2198	0.5926	0.2034
3	Combined	LR	0.6461	0.1442	0.6868	0.2383	0.7257	0.2051	0.6357	0.2221
4	Combined	RF	0.6438	0.1412	0.6725	0.2333	0.7178	0.1997	0.6262	0.2183
5	Combined	XGB	0.6565	0.1466	0.6766	0.2410	0.7310	0.2140	0.6110	0.2122
6	Synthetic	LR	0.5890	0.1234	0.6719	0.2086	0.6812	0.1679	0.6638	0.2359
7	Synthetic	RF	0.9195	0.7500	0.0017	0.0033	0.6626	0.1600	0.3091	0.0840
8	Synthetic	XGB	0.6228	0.1219	0.5933	0.2023	0.6459	0.1589	0.6253	0.2204

Figure 11: Model Predictive Results

As dictated in figure 11, Synthetic-only training generally resulted in lower ROC-AUC (0.6459-0.6812), though tree-based models (RF, XGB) retained more of their predictive performance than Logistic Regression.

LogLoss and Brier Score are metrics that assess the quality of the model's predicted probabilities, not just the final classification outcome. LogLoss (Logarithmic Loss): This metric penalizes models that are confident in incorrect predictions more heavily than those that are less confident. A lower LogLoss indicates better calibrated probabilities (closer to the true probability)(Vovk, 2015). The brier Score Loss metric calculates the mean squared difference between the predicted probability and the actual outcome (0 or 1). A lower Brier score indicates more accurate probability predictions(Hoessly, 2025). Both LogLoss and Brier Score provide insights into how reliable the model's probability outputs are, which is important for downstream tasks that might rely on these probabilities rather than just the predicted class label.

5.3 Fairness Metrics

Fairness in credit scoring is critical given the regulatory, ethical, and reputational implications of biased decision-making(Bellamy et al., 2019).In this study, fairness was evaluated using three group fairness metrics across the following protected (attributes that separate a population into groups that have differences in terms of benefits received) attributes education type, housing type, and income type

- Average Odds Difference (AOD) — Measures the average difference in false positive rates (FPR) and true positive rates (TPR) between privileged and unprivileged groups. Values near zero indicate fairness while values further from zero indicate disparity in classification outcomes. Positive values suggest a bias favoring the privileged group, and negative values indicate the opposite.

- Equal Opportunity Difference (EOD) — Measures the difference in TPR between privileged and unprivileged groups. A value of zero means both groups have equal access to positive outcomes.

- Statistical Parity Difference (SPD) — Measures the difference in the proportion of positive predictions between groups. Zero denotes fairness while deviations indicate a higher or lower likelihood of favorable outcomes for a particular group, regardless of actual qualification.

	Attribute	Model	Baseline AOD	Combined AOD	Synthetic AOD	Baseline EOD	Combined EOD	Synthetic EOD	Baseline SPD	Combined SPD	Synthetic SPD
0	NAME_EDUCATION_TYPE	LR	0.2489	0.3504	0.4295	0.2757	0.2711	0.4212	0.2536	0.4569	0.4344
1	NAME_EDUCATION_TYPE	RF	0.3275	0.3995	0.0011	0.3246	0.3574	0.0021	0.3578	0.4677	0.0002
2	NAME_EDUCATION_TYPE	XGB	0.3306	0.3068	0.4324	0.2417	0.1929	0.4151	0.4454	0.4477	0.4680
3	NAME_HOUSING_TYPE	LR	0.2562	0.5121	0.5987	0.3109	0.5761	0.6396	0.2151	0.4709	0.5719
4	NAME_HOUSING_TYPE	RF	0.3391	0.3001	0.0010	0.4351	0.3701	0.0019	0.2691	0.2516	0.0002
5	NAME_HOUSING_TYPE	XGB	0.3079	0.3448	0.5637	0.4026	0.3751	0.6242	0.2387	0.3311	0.5210
6	NAME_INCOME_TYPE	LR	0.5559	0.5962	0.6606	0.7244	0.7498	0.7760	0.4189	0.4715	0.5668
7	NAME_INCOME_TYPE	RF	0.4951	0.5147	0.0026	0.6535	0.5841	0.0050	0.3665	1.0000	0.0003
8	NAME_INCOME_TYPE	XGB	0.5501	0.5755	0.5787	0.7207	0.7336	0.6802	0.4114	0.4471	0.4963

Figure 12: Fairness Evaluation Results

Figure 12 depicts the fairness metrics results which reveal substantial and persistent disparities across protected attributes. For instance, an SPD of 0.5668 for income type in the synthetic-only Logistic Regression model indicates that privileged income groups were over 56% more likely to receive favourable credit predictions, irrespective of actual repayment likelihood. Similarly, an EOD of 0.7760 suggests a 77.6% higher true positive rate for privileged income groups compared to their unprivileged counterparts. Inconsistencies are also evident for housing type (e.g., SPD = 0.5210 in synthetic-only XGBoost) and education type (SPD = 0.4680 in synthetic-only XGBoost), demonstrating that bias is not limited to a single attribute. These results are consistent with the original dataset's bias patterns, confirming that CTGAN preserves — and in some cases amplifies — existing inequities.

6 Conclusion

This research set out to evaluate the extent to which CTGAN-generated synthetic credit-scoring data could serve as a viable alternative to real datasets, particularly in the context of developing and testing alternative data credit scoring models for credit-invisible populations.

The findings demonstrate that CTGAN effectively captures the statistical dependencies of the original dataset, with high Pearson and Spearman correlation similarities and reasonable alignment in feature distributions. Predictive modelling results show that tree-based classifiers trained on synthetic data can achieve competitive ROC-AUC scores compared to those trained on real data, indicating preserved non-linear patterns essential to credit scoring.

However, moderate reductions in performance, particularly in logistic regression models, and a moderate feature-importance rank correlation (0.2164) suggest that synthetic

datasets are not perfect replicas. Most critically, fairness analysis reveals that CTGAN tends to replicate, rather than mitigate, existing biases in the original data. This means that while synthetic data provides clear advantages in terms of privacy and data accessibility, it cannot be assumed to produce fairer models without deliberate bias mitigation measures.

From a research perspective, the study agrees that CTGAN generated synthetic data is a valuable tool for reproducible experimentation, particularly in contexts where real data access is limited by regulation or privacy concerns. From an industry perspective, CTGAN can support rapid prototyping, stress-testing, and algorithm development without risking exposure of sensitive customer information. However, for production-grade applications, further fidelity improvement and bias control will be necessary.

Ultimately, CTGAN-generated synthetic data should be viewed as a potent supplement rather than a complete replacement for original data. When combined with fairness improvements, it holds strong potential to advance financial inclusion by enabling broader research and innovation in credit scoring for the underserved populations.

7 References

Agarwal, S., Alok, S., Ghosh, P., and Gupta, S. (2023) 'Financial Inclusion and Alternate Credit Scoring: Role of Big Data and Machine Learning in Fintech', Indian School of Business. http://dx.doi.org/10.2139/ssrn.3507827.

Djeundje, V. B., Crook, J., Calabrese, R. and Hamid, M. (2021) 'Enhancing credit scoring with alternative data', Expert Systems with Applications, 163. `https://doi.org/10.1016/j.eswa.2020.113766.`

Jordon, J., Yoon, J. and Van Der Schaar, M. (2018) 'PATE-GAN: Generating synthetic data with differential privacy guarantees', in International conference on learning representations.

Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K., (2019). 'Modeling tabular data using conditional GAN'. Advances in neural information processing systems (32), NeurIPS Proceedings. `https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78Paper.pdf.`

Hlongwane, R., Ramaboa, K.K.K.M., and Mongwe, W. (2024) 'Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data'. PLoS ONE 19(5): e0303566. https://doi.org/10.1371/journal.pone.0303566.

Mfanasibili Ngwenya. (2024) "Credit Scoring in Africa: Employing Logistic Regression on Alternative Data", 2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Male, Maldives, 2024, pp. 1-6. doi: 10.1109/ICECCME62383.2024.1079601110.1109/ICECCME62383.2024.10796011.

Caliskan, H., Yayla, O.F. and Genc, Y. (2023)' A Comparative Analysis of Synthetic Data Generation with VAE and CTGAN Models on Financial Credit Loan Offer Data', in 2023 8th International Conference on Computer Science and Engineering (UBMK) (pp. 212-217).

Manasa, V., Varshney, N., Al-Jawahry, H.M., Kumar, B.S., Prabhakar, P.E. and Kumar, B.A. (2024) 'Leveraging Generative Adversarial Networks for Synthetic Financial Data

Generation'. in 2024 International Conference on IoT, Communication and Automation Technology (ICICAT) (pp. 1409-1414).

Ramachandra, P. and Vaithiyanathan, S. (2025). 'Fed-DPSDG-WGAN: Differentially Private Synthetic Data Generation for Loan Default Prediction via Federated Wasserstein GAN'. IEEE Access. Doi: 10.1109/ACCESS.2025.3552487.

Montoya, A. inversion, KirillOdintsov, and Kotek, M. (2018) Home Credit Default Risk. Available at: https://kaggle.com/competitions/home-credit-default-risk, 2018. Kaggle.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) 'Generative adversarial nets'. *Advances in neural information processing systems*, 27.

Mirza, M. and Osindero, S. (2014) 'Conditional generative adversarial nets'.

Cui, E.H., Li, Y. and Liu, Z. (2025) 'The Kolmogorov-Smirnov Statistic Revisited'.

Mana, S.C. and Sasipraba, T. (2021) 'Research on cosine similarity and pearson correlation based recommendation models', *Journal of Physics: Conference Series* (Vol. 1770, No. 1, p. 012014).

Vovk, V. (2015) 'The fundamental nature of the log loss function', *in Fields of logic and computation II*: Essays dedicated To Yuri Gurevich on the Occasion of His 75th Birthday (pp. 307-318). Cham: Springer International Publishing.

Hoessly, L. (2025) 'On misconceptions about the Brier score in binary prediction models'.

Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. and Nagar, S. (2019)' AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias'. *IBM Journal of Research and Development*, 63(4/5), pp.4-1.