SynthRank: Synthetic Data Generation of Individual's Financial Transactions Through Learning to Ranking

Waylon Li Mengyu Wang Carsten Maple Tiejun Ma

University of Edinburgh, University of Warwick, The Alan Turing Institute

Overview

- SynthRank is a task-oriented pipeline for synthetic financial transaction data generation.
- SynthRank leverages LEarn-TO-Rank (LETOR) to extract relationship between transaction records and labels for generating synthetic features.
- SynRank can balance privacy preservation (against inference attacks), utilities for the risky trader prediction task, and feature similarity.

Dataset

- Contains 13,607,120 trading records from November 2003 to July 2014.
- Contains sensitive information of individuals.

Discrete	Age	Age group	Mobile	Device
	Segment	Transaction size	MarketCluster	Trading market preference
	accountid	The account ID of the trader	Period	Buckets of 20 trades per account
Continuous	AVGPTS3_20	P&L in Points >= 3 in the past 20 trades	PerFTSE20	Share of trades placed
	ProfitRate20	Average profit rate of the past 20 trades	NextTotalPL_GBP20	P&L for the next 20 trades
	WinTradeRate20	Average winning rate in past 20 trades	NextTotalPL_GBP	P&L for the next 100 trades
	SharpeRatio20	Sharpe ratio of the past 20 trades	WinningRate	Winning rate of the next 100 trades
	ProfitxDur20	Interaction of ProfitRate20 and DurationRate20	SharpNext100	Sharpe ratio during the next 100 trades
	PassAvgReturn	Avg. return	TotalTrades	Total trades made by a trader
	AvgShortSales20	Share of short positions in the past 20 trades	DurationRate20	Average time (winning vs losing position open)
	AvgOpen20	Average of the P&L among past 20 trades	DurationRatio20	Mean duration / std.dev. duration of past 20 trades
	OrderCloseRate20	% of trades closed by an order in past 20 trades.	TradFQ20	The number of daily trades on average for the trader
	NumTrades	Accumulated until the past 20 trades		

Table 1. Features in the dataset.

Task-oriented Generation: What Is The Task?

Task: Risky trader prediction in Contracts for Difference (CFDs).

- 10% of the £1.2 trillion trades annually on the London Stock Exchange are related to CFDs [1].
- Crucial for maintaining market integrity and reducing risks for market makers.

Primary objective: determine the hedging strategy for trade j. (hedge if the future return of the trader is in the top 1%)

Methodology

Key idea:

- **Transformation to Ranking Problem:** The dataset is transformed into a ranking problem by defining ranking groups (*qid*) and using the risk label as a binary relevance label.
- **Ranking Group Allocation As Noise:** Diverse synthetic financial transactions are generated by varying group allocations and using the ranking model to create a ranking score embedding.

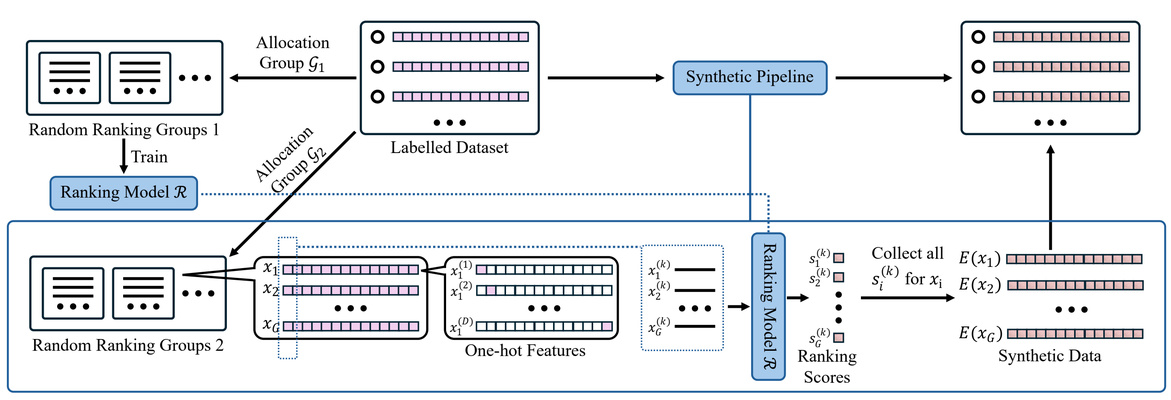

Figure 1. Illustration of SynthRank pipeline.

Functionalities:

- **Flexible and Controllable Generation:** (1) Unlimited synthetic data generation by varying the ranking group allocations. (2) Capability to generate data with specific labels, including pure non-risky and pure-risky group.
- **Privacy-preserved Generation:** (1) The generated synthetic data is difficult to infer sensitive information from due to concealed ranking group allocation and ranking model. (2) Enhanced security against inference attacks, making it unfeasible to deduce original values from synthetic data.

Evaluation 1: Maximising Synthetic Data Utility

Data Source	RF Classifier F_1	RF Classifier P&L	RF Classifier AUC	MLP Classifier F_1	MLP Classifier P&L	MLP Classifier AUC
Original	0.511	127.738	0.837	0.522	127.968	0.838
TVAE	0.498	127.060	0.732	0.509	127.539	0.648
CTGAN	0.510	127.833	0.781	0.531	124.721	0.742
CopulaGAN	0.499	127.203	0.775	0.517	125.907	0.729
RTF	0.499	127.098	0.824	0.498	127.131	0.805
LMART(XGB:pair)	0.548	128.185	**0.863**	0.543	131.207	**0.862**
LMART(XGB:ndcg)	0.558	128.985	0.849	0.536	129.858	0.848
LMART(XGB:map)	**0.567**	129.160	0.858	0.560	128.451	0.845
LMART(LGBM)	0.566	**131.269**	0.857	**0.567**	**131.547**	0.856

Table 2. Performance comparison of RF and MLP models on the original dataset versus synthetic datasets generated using different benchmarking models.

Evaluation 2: Privacy Preservation

Privacy Inference Attack: Attackers trained on synthetic data will try to infer actual sensitive features by some known insensitive features in the actual data.

Privacy Against Inference (PAI) score: $(1 - Acc)$, where Acc is the accuracy of an attacker in inferring true sensitive information.

Model	RF	KNN	MLP
TVAE	0.4510	0.5155	0.4414
CTGAN	0.4374	0.5021	0.4311
CopulaGAN	0.4288	0.4850	0.4293
RTF	0.3973	0.1442	0.4324
SynthRank	**0.9453**	**0.9453**	**0.9117**

Table 3. Average PAI scores for synthetic data from baseline models and SynthRank with LambdaMART (LGBM).

Evaluation 3: Feature Similarity

Feature Type	TVAE	CTGAN	CoGAN	RTF	SynthRank
Continuous	0.188	0.147	0.152	**0.129**	0.153
Discrete	0.082	0.097	0.145	**0.003**	0.460
All	0.128	0.119	0.148	**0.058**	0.326

Table 4. Average KS test statistics of continuous and discrete features. Notably, "CoGAN" stands for "CopulaGAN".

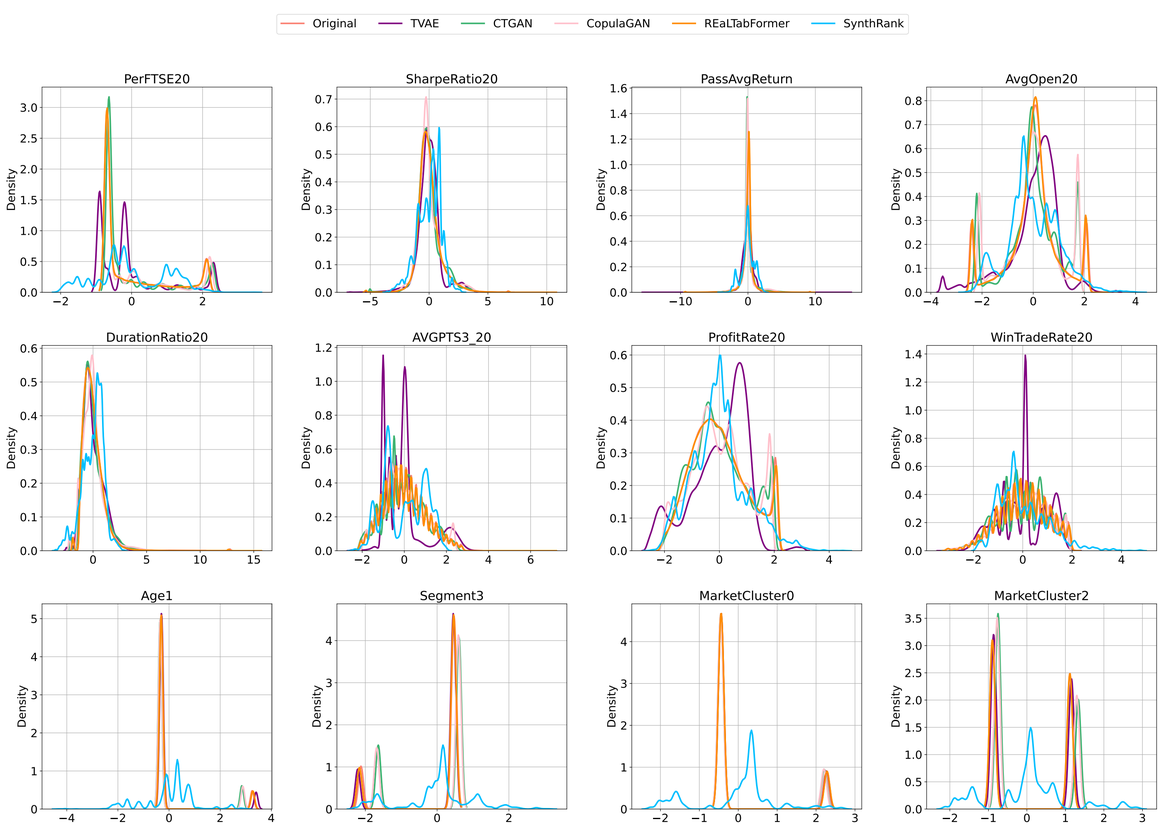

Figure 2. KDE Plots of 12 Selected Features from Original and Synthetic Datasets.

Key Findings

- According to Table 2, SynthRank shows consistent improvement on maximising the utility for the risky trader prediction task compared to all the baseline models.
- In Table 3, SynthRank consistently outperforms the baseline models across all three attacker models.
- Evident by Table 4 and Figure 2, SynthRank fall shorts in mimicking the discrete features since it transform them into continuous but it demonstrates its ability to capture the peaks of the original discrete distributions.

References

[1] T. Ma, P.A.F. Fraser-Mackenzie, M. Sung, A.P. Kansara, and J.E.V. Johnson. Are the least successful traders those most likely to exit the market? a survival analysis contribution to the efficient market debate. *European Journal of Operational Research*, 299(1):330–345, 2022.

[2] B. Zhao, A. Agrawal, C. Coburn, H. Asghar, R. Bhaskar, M. Kaafar, D. Webb, and P. Dickinson. On the (in)feasibility of attribute inference attacks on machine learning models. In *2021 IEEE European Symposium on Security and Privacy (EuroSamp;P)*, pages 232–251, Los Alamitos, CA, USA, sep 2021. IEEE Computer Society.