ENCODING

The original data set includes 1,103,307 records and 44 features. After data cleaning and feature selection, we got 11,936 records and 7 features. By label encoding, we transform all the continuous and categorical data into letters

	ACTION	CLEARED	COLLATERALIZATION	TAXONOMY	TRADE_CONTINUATION	PRICE	AMOUNT
0	NEW	U	UC	Credit:Index:CDX:CDXIG	Trade	1	100
1	NEW	U	PC	Credit:Index:CDX:CDXIG	Trade	1	30
2	CORRECT	U	PC	Credit:Index:CDX:CDXIG	Trade	1	30
3	NEW	U	PC	Credit:Index:CDX:CDXIG	Novation	1.24	25
4	CORRECT	C	UC	Credit:Index:CDX:CDXHY	Trade	1	5
5	CORRECT	C	FC	Credit:Index:CDX:CDXIG	Trade	1	58

NE IG T (0,1] a m (90,100]

Sample of encoded text:

NE U UC IG T a m , NE U PC IG T a f , CR U PC IG T a f ...

We build a vocabulary consisting of 847 combinations of the 7 features. Then we map each possible value into number. We aggregate every 30 orders as one sequence, finally got 385 sequences.

Action

"NEW" : "NE"
"CORRECT" : "CR"
"CANCEL": "CC"

denotes the action we operate on the CDS index

Collateralization

"UC" : Uncollateralized
"FC" : Fully collateralized
"PC": Partially collateralized
"OC": One-way collateralized

reflects the risk of CDS index

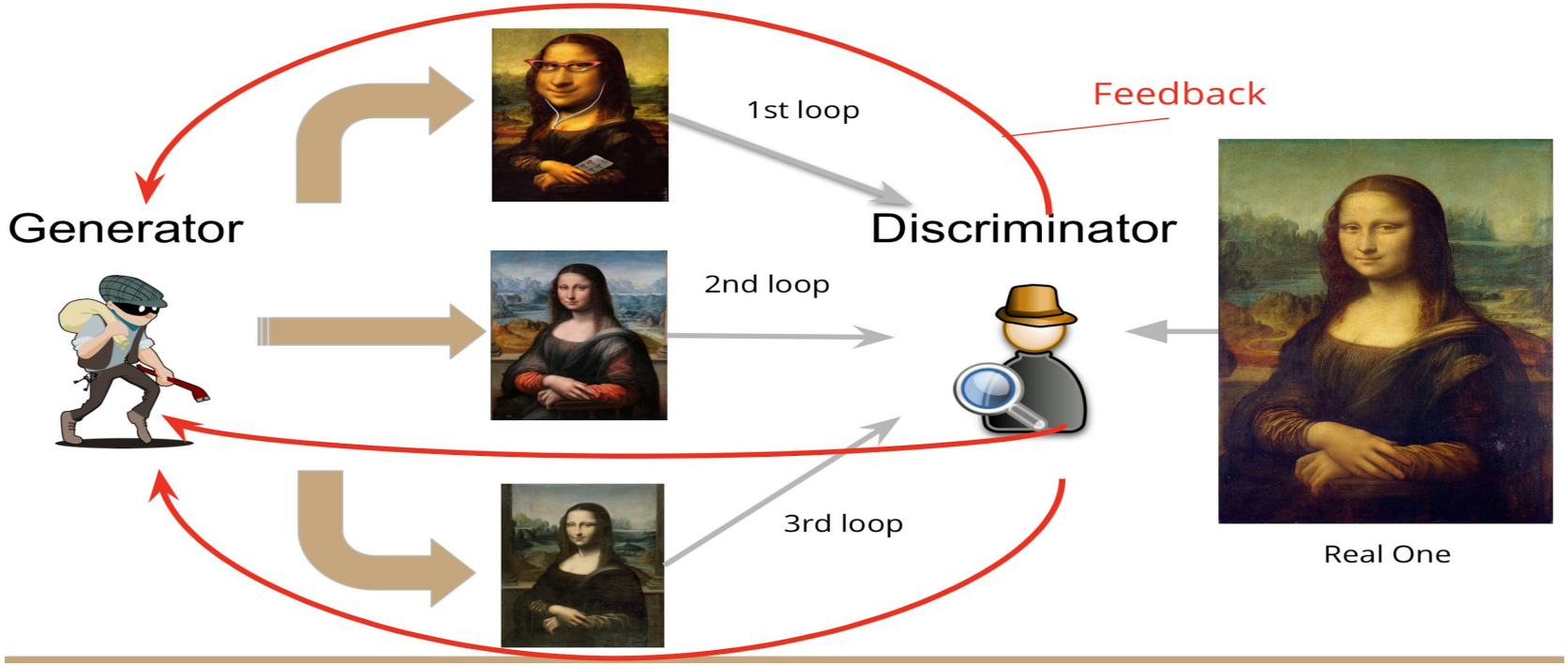
Trade Continuation

"Amendment" : A
"Novation" : N & "Trade": T
"Partial termination": P

denotes the post-execution trade continuation or life cycle

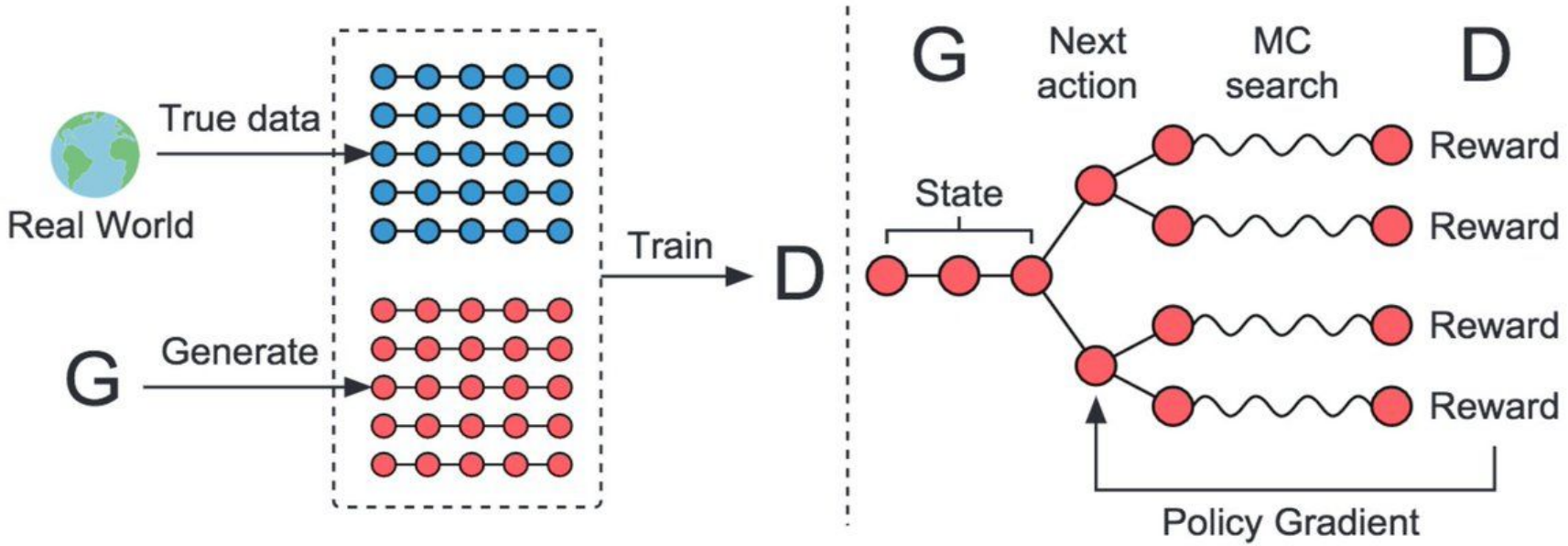
MODEL -- GAN

Generative Adversarial Network (GAN) has succeed in generating synthetic pictures that mimic the real one. We want to use this advantage to generate more financial data to solve the data lacking problem. In that way, we can use the generated data for better financial model building.



MODEL -- SeqGAN

We process our financial data as sequence data. However, the vanilla GAN has problems in generating discrete data. We applied the seqGAN built by Lantao Yu to remodel our generator and discriminator in a reinforcement learning framework.

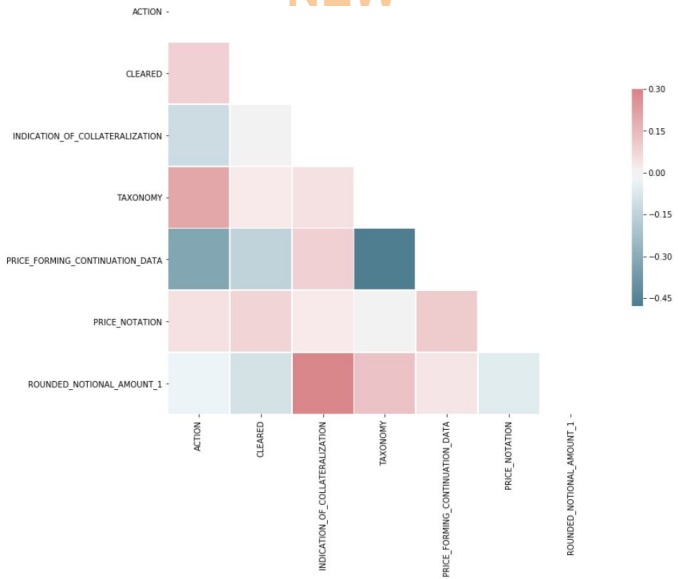


(https://arxiv.org/abs/1609.05473?context=cs.AI)

EVALUATION METRIC

To evaluate the quality of generated data, we need to design a metric. The first measurement is to check whether the correlation among different features still holds. We plot a correlation matrix between each pair of features.

NEW



OLD

The results show that the low correlation between different feature still holds

Second we want to see the multi-correlations. We do regression on continuous data and do classification on categorical data. Here we give an example of the regression of other 6 variables on price.

Coefficients	Estimated Std.	Pr > (| t |)	Significance
Intercept	6.2291	0.0333	*
Action Correct	3.4928	0.2276	
Action New	4.6757	0.1017	
Cleared	-4.2517	< 2E-16	***
Collateralization OC	-3.6778	0.1623	
Collateralization PC	-3.7982	2.65E-14	***
Collateralization UC	-3.1575	1.38E-09	***
Taxonomy	-2.9069	2.36E-13	***
Trade Continuation Novation	-2.4131	0.0230	*
Trade Continuation Partial Termination	22.8470	< 2E-16	***
Trade Continuation Trade	-0.6981	0.3133	

Significance Level:
*-significant,
**- very significant,
***-super significant

The accuracy of classifying ‘trade continuation’, ‘taxonomy’, and ‘cleared’ is 0.8572, 0.9317, and 0.8496 respectively.

This means that there exists high dependency among variables

Next we want to see whether the relative performance of each two algorithms (trained and tested) on the synthetic dataset is the same as their relative performance (when trained and tested) on the original dataset. We use five different models to classify the “cleared” feature. And then we use a metric called SRA to evaluate that.

Accuracy	Random Forest	Logistic Regression	Linear SVM	Multinomial Naive Bayes	Gaussian Naive Bayes
Original Data	0.87144	0.74539	0.69305	0.62353	0.57789
Generated Data	0.983	0.79265	0.7886	0.64165	0.3268

$$R_i = m_\tau(A_i(D_1), D_2)$$
$$S_i = m_\tau(A_i(D_1^G), D_2^G)$$

R_i, S_i Performance metric score
m_τ Performance metric on task τ
$A_i(D_1)$ Trained model from training data
D_1^G Training data from synthetic data
D_2^G Test data from synthetic data

$$SRA(G) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i} I((R_i - R_j) \times (S_i - S_j) > 0)$$

The SRA can be thought of as the (empirical) probability of a comparison on the synthetic data being ”correct”. SRA(“Cleared”) = 1

Other features:
SRA(“Action”) = 0.8, SRA(“Collateralization”) = 0.9, SRA(“Price forming”) = 0.9, SRA(“Taxonomy”) = 0.6
Based on the result, we can see the quality of the synthetic data is good in total.

Future work

Model | Hyperparameter tuning
Features | Underlying asset, execution venue, design new features
Application | Strategic design, simulation

Acknowledgement

J.P.Morgan

FUNG INSTITUTE FOR ENGINEERING LEADERSHIP
UC BERKELEY ENGINEERING

BERKELEY IEOR
INDUSTRIAL ENGINEERING & OPERATIONS RESEARCH

SHANGHAI JIAO TONG UNIVERSITY 1896

Guo Xin | Professor at UCB

Cheng-Ju Wu, Yang Nan | Ph.D student at UCB

Ian Goodfellow | Father of GAN

Lantao Yu | Primary author of SeqGAN paper