Customs Import Declaration Dataset

Chaeyoon Jeong, Sundong Kim, Jaewoo Park, Yeonsoo Choi

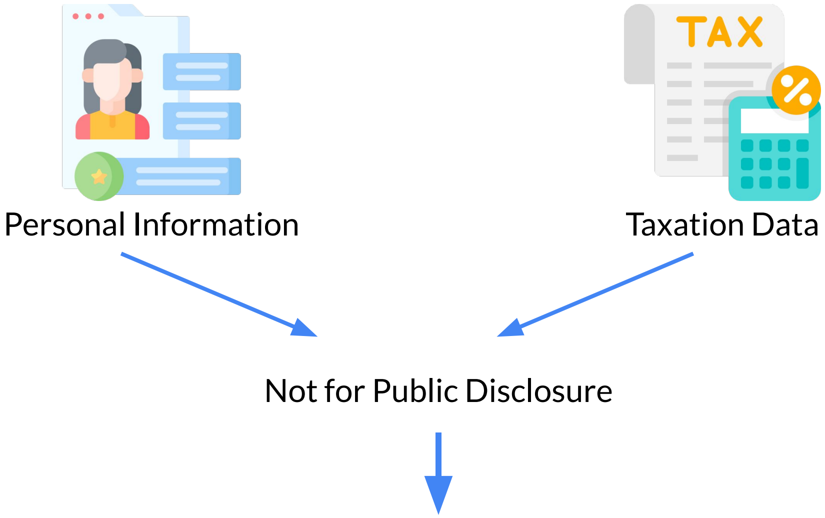
joungchaeyoon@gmail.com / lily9991@kaist.ac.kr
KAIST School of Computing, IBS Data Science Group

Contents

1. Introduction
2. Data Description
3. Data Generation
4. Data evaluation
5. Application - Fraud Detection
6. Summary

Data Opening and Privacy Concerns

Customs data contains sensitive information…



Personal Information

Taxation Data

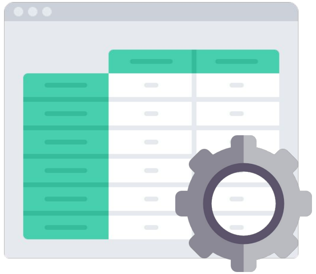
Not for Public Disclosure

Limited implementation of new technology (e.g. AI, Data Science)

Opening Synthetic Dataset

Synthetic customs import declaration dataset w/o privacy concerns

- Accessible data
- Facilitate customs research

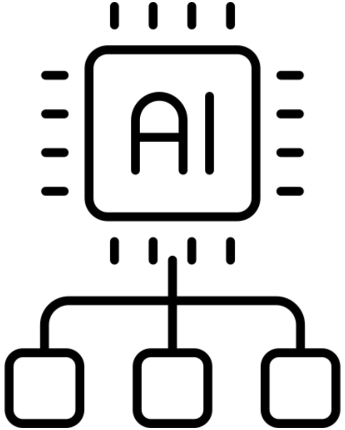


Synthetic Data

Can be found at: https://github.com/Seondong/Customs-Declaration-Datasets

Needs for Data Disclosure

Adoption of new technology is slow



Data utilization + AI

Data Description

Data Description

Data Outline

Tabular Format

54,000 rows (Declarations)

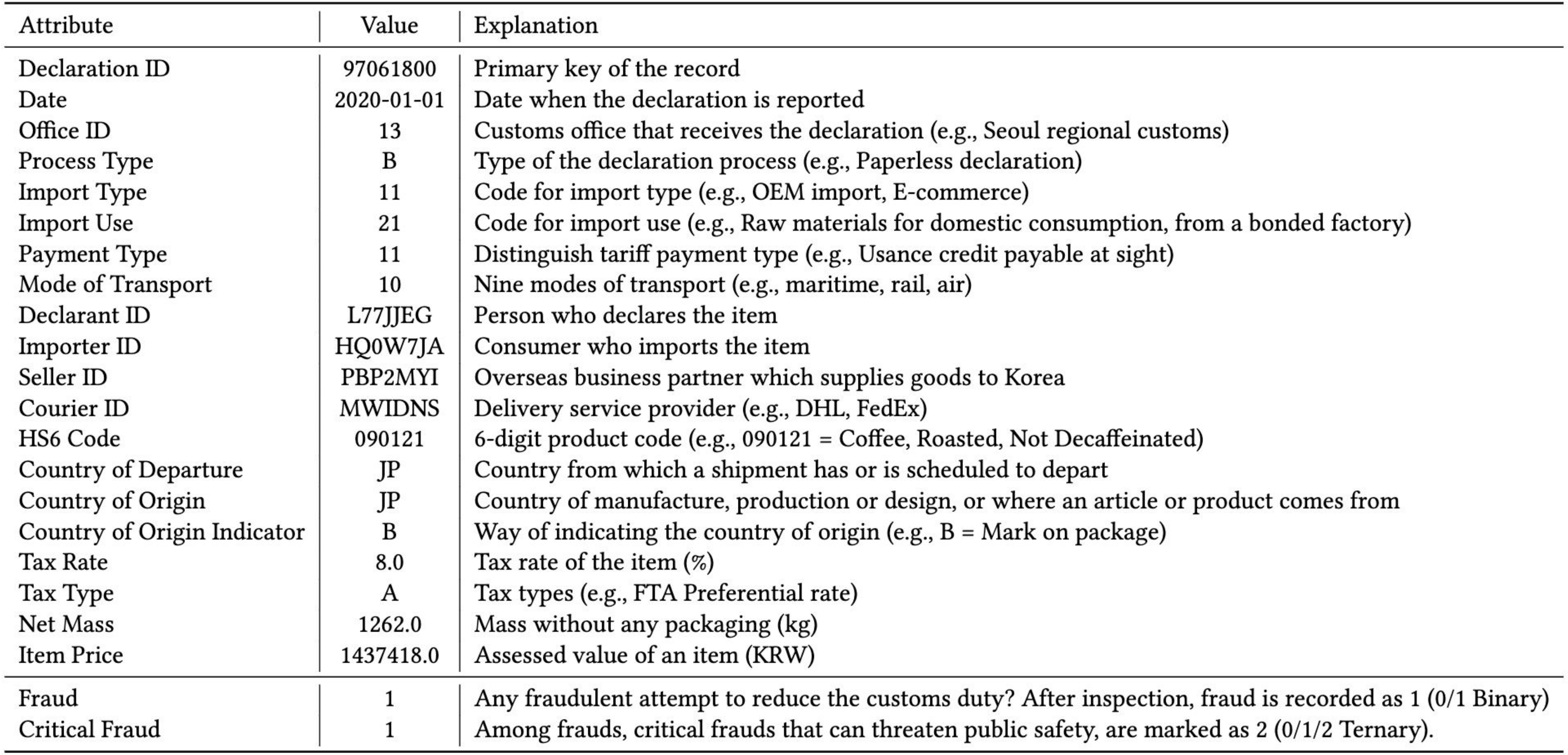
- Imports between January 2020 - June 2021

22 columns (Attributes)

- Data ID + Attributes + Inspection result

Declaration ID	Date	Office ID	Declarant ID	HS6 Code	Country of Departure	...	Net Mass (Kg)	Fraud	Critical Fraud
98902919	2020-01-01	40	A0ZY90B	850440	MY	...	108.0	0	0
45548205	2020-01-01	30	5PDPMM1	870899	CN	...	166.0	1	1
...	...	...	...	...	...	...	...	...	...

Data Outline



Attribute	Value	Explanation
Declaration ID	97061800	Primary key of the record
Date	2020-01-01	Date when the declaration is reported
Office ID	13	Customs office that receives the declaration (e.g., Seoul regional customs)
Process Type	B	Type of the declaration process (e.g., Paperless declaration)
Import Type	11	Code for import type (e.g., OEM import, E-commerce)
Import Use	21	Code for import use (e.g., Raw materials for domestic consumption, from a bonded factory)
Payment Type	11	Distinguish tariff payment type (e.g., Usance credit payable at sight)
Mode of Transport	10	Nine modes of transport (e.g., maritime, rail, air)
Declarant ID	L77JJEG	Person who declares the item
Importer ID	HQ0W7JA	Consumer who imports the item
Seller ID	PBP2MYI	Overseas business partner which supplies goods to Korea
Courier ID	MWIDNS	Delivery service provider (e.g., DHL, FedEx)
HS6 Code	090121	6-digit product code (e.g., 090121 = Coffee, Roasted, Not Decaffeinated)
Country of Departure	JP	Country from which a shipment has or is scheduled to depart
Country of Origin	JP	Country of manufacture, production or design, or where an article or product comes from
Country of Origin Indicator	B	Way of indicating the country of origin (e.g., B = Mark on package)
Tax Rate	8.0	Tax rate of the item (%)
Tax Type	A	Tax types (e.g., FTA Preferential rate)
Net Mass	1262.0	Mass without any packaging (kg)
Item Price	1437418.0	Assessed value of an item (KRW)
Fraud	1	Any fraudulent attempt to reduce the customs duty? After inspection, fraud is recorded as 1 (0/1 Binary)
Critical Fraud	1	Among frauds, critical frauds that can threaten public safety, are marked as 2 (0/1/2 Ternary).

Data Attributes

Declaration ID	Date	Office ID	Declarant ID	HS6 Code	Country of Departure	...	Net Mass (Kg)	Fraud	Critical Fraud
98902919	2020-01-01	40	A0ZY90B	850440	MY	...	108.0	0	0
45548205	2020-01-01	30	5PDPMM1	870899	CN	...	166.0	1	1
...	...	...	...	...	...	...	...	...	...

Data Attributes

Declaration ID	Date	Office ID	Declarant ID	HS6 Code	Country of Departure	...	Net Mass (Kg)	Fraud	Critical Fraud
98902919	2020-01-01	40	A0ZY90B	850440	MY	...	108.0	0	0
45548205	2020-01-01	30	5PDPMM1	870899	CN	...	166.0	1	1
...	...	...	...	...	...	...	...	...	...

Unique ID for each row

Primary Key

Data Attributes

Declaration ID	Date	Office ID	Declarant ID	HS6 Code	Country of Departure	...	Net Mass (Kg)	Fraud	Critical Fraud
98902919	2020-01-01	40	A0ZY90B	850440	MY	...	108.0	0	0
45548205	2020-01-01	30	5PDPMM1	870899	CN	...	166.0	1	1
...	...	...	...	...	...	...	...	...	...

19 essential attributes in the import declaration

ex) Type of product, Country of origin, Taxation type, etc

Data Attributes

Declaration ID	Date	Office ID	Declarant ID	HS6 Code	Country of Departure	...	Net Mass (Kg)	Fraud	Critical Fraud
98902919	2020-01-01	40	A0ZY90B	850440	MY	...	108.0	0	0
45548205	2020-01-01	30	5PDPMM1	870899	CN	...	166.0	1	1
...	...	...	...	...	...	...	...	...	...

Inspection Result

0: Normal / 1: Fraud / 2: Fraud with severe violation

Data Attributes

Declaration ID	Date	Office ID	Declarant ID	HS6 Code	Country of Departure	...	Net Mass (Kg)	Fraud	Critical Fraud
98902919	2020-01-01	40	A0ZY90B	850440	MY	...	108.0	0	0
45548205	2020-01-01	30	5PDPMM1	870899	CN	...	166.0	1	1
...	...	...	...	...	...	...	...	...	...

Categorical (Discrete) variable

Numerical (Continuous) variable

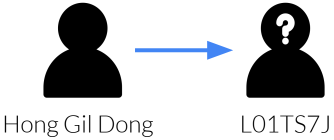
Data Generation

Removing sensitive information

To remove the possibility to retrieve the original data

1. Anonymization

   Convert each identity (Importer ID, Declarant ID, *etc.*) into code



2. Product classification code HS10 → HS6 Code

   Maintain only the international standard first 6 digits of HS code

9401.31-1000 : Swivel seats with variable height adjustment of wood, Covered with leather

↓

9401.31-1000 : Swivel seats with variable height adjustment of wood, Covered with leather

Maintaining Correlation Between attributes

To enhance the realism of the generated data

Join highly correlated attributes

→ Force the synthesizer to preserve the predetermined customs patterns between correlated attributes

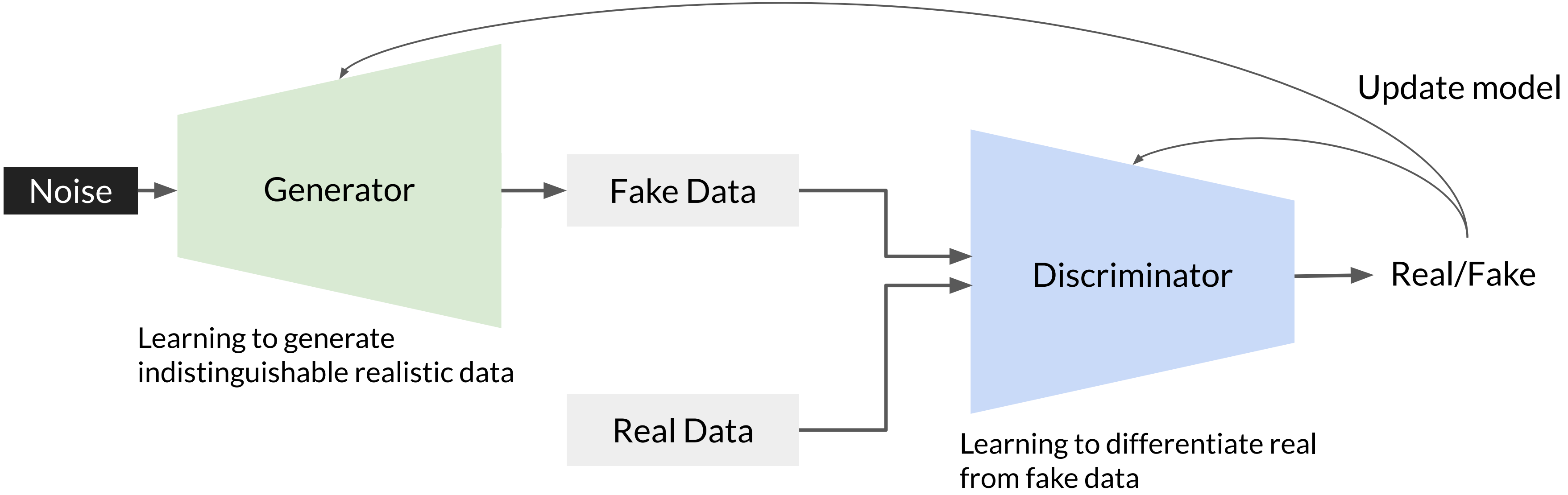
HS10 Code	Country of Departure	Country of Origin	Item Price	Tax Rate	Tax Type	Net Mass
4408909195	BE	BE	372254.4	0.0	FEU1	108.0
6907221000	CN	CN	375751.2	8.0	A	11352.0

↓

4408909195^BE^BE^372254.4^0.0^FEU1^108.0
6907221000^CN^CN^375751.2^8.0^A^11352.0

CTGAN: Conditional GAN for Tabular data

Train the model to mimic the source data



Learning to generate indistinguishable realistic data

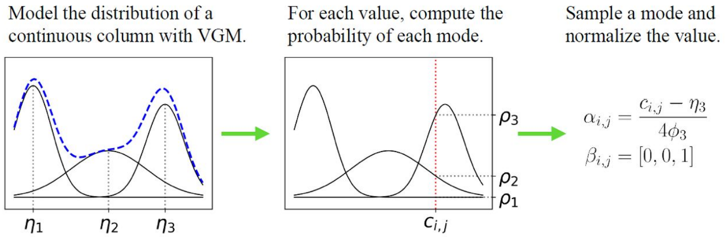
Learning to differentiate real from fake data

L Xu et. al., Modeling Tabular Data using Conditional GAN, NeurIPS 2019

CTGAN: Conditional GAN for Tabular data

1. Mode-specific Normalization

Convert continuous values into discrete values



Model the distribution of a continuous column with VGM.

For each value, compute the probability of each mode.

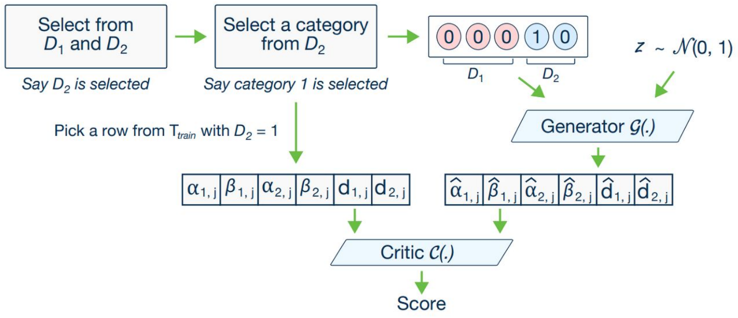
Sample a mode and normalize the value.

$$\alpha_{i,j} = \frac{c_{i,j} - \eta_3}{4\phi_3}$$

$$\beta_{i,j} = [0, 0, 1]$$

2. Training-by-sampling

Maintain the frequency of values existing in columns



Select from D_1 and D_2
Say D_2 is selected
Select a category from D_2
Say category 1 is selected
0 0 0 1 0
D_1
D_2
$z \sim \mathcal{N}(0, 1)$
Pick a row from T_{train} with $D_2 = 1$
Generator $\mathcal{G}(.)$
$\alpha_{1,j}$ $\beta_{1,j}$ $\alpha_{2,j}$ $\beta_{2,j}$ $d_{1,j}$ $d_{2,j}$
$\hat{\alpha}_{1,j}$ $\hat{\beta}_{1,j}$ $\hat{\alpha}_{2,j}$ $\hat{\beta}_{2,j}$ $\hat{d}_{1,j}$ $\hat{d}_{2,j}$
Critic $\mathcal{C}(.)$
Score

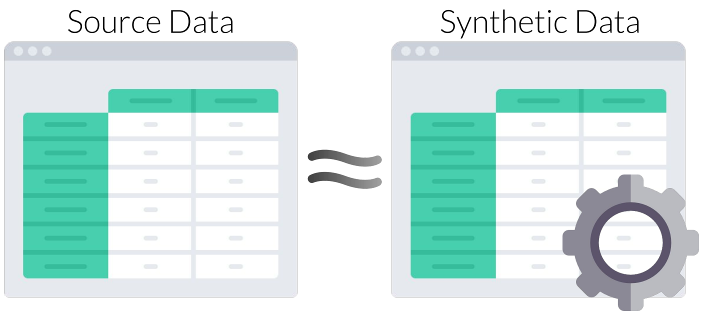
L Xu et. al., Modeling Tabular Data using Conditional GAN, NeurIPS 2019

Data Evaluation

Synthetic Data Quality Metrics

1. Similarity in data distribution

Does the synthetic data capture the distribution and correlation in real data? Is it realistic?



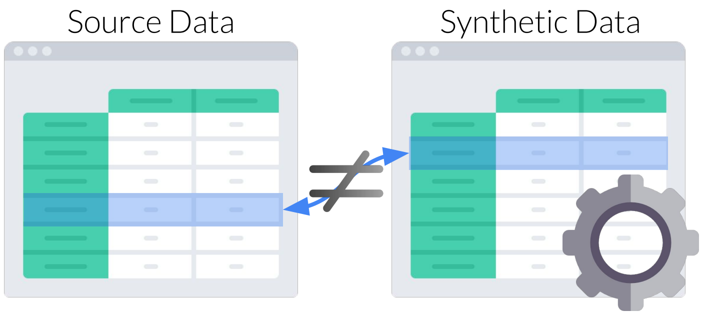
Source Data

Synthetic Data

- Column Shape (Distribution)
- Column Pair Trend (Correlation)
- Coverage (Presence)
- Boundary (Outliers)

2. Diversity of generated data

Is the synthetic data unique or does it copy the real rows?

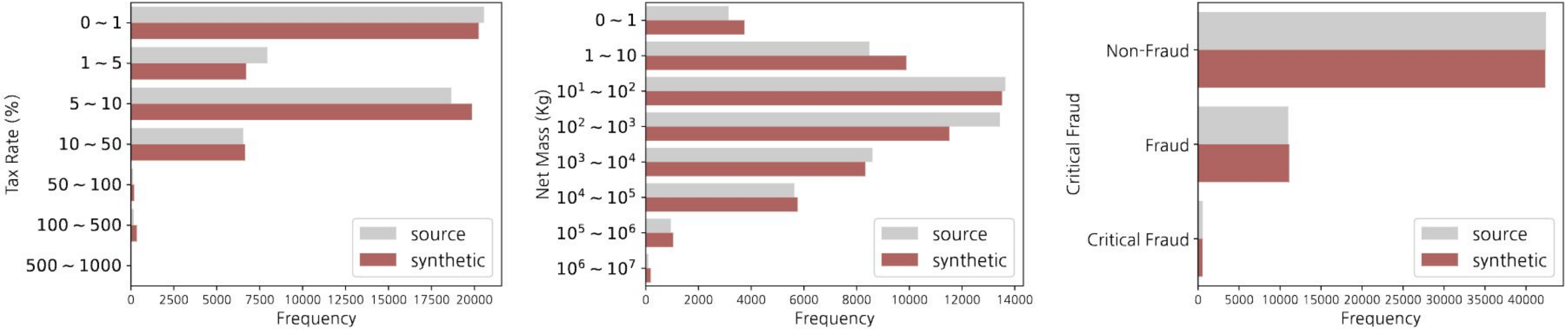


Source Data

Synthetic Data

- Diversity (Uniqueness)

Column Shape Similarity



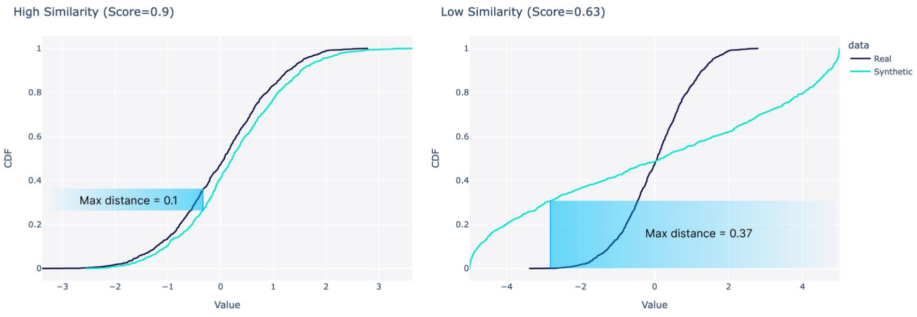
(a) Tax Rate **(b) Net Mass** **(c) Critical Fraud**

Distributions of representative features are similar between synthetic data and source data

Column Shape Similarity

1. Numerical - KS statistic

Complement of maximum distance between CDF



2. Categorical - Total variation distance

Complement of difference of proportion of each variable

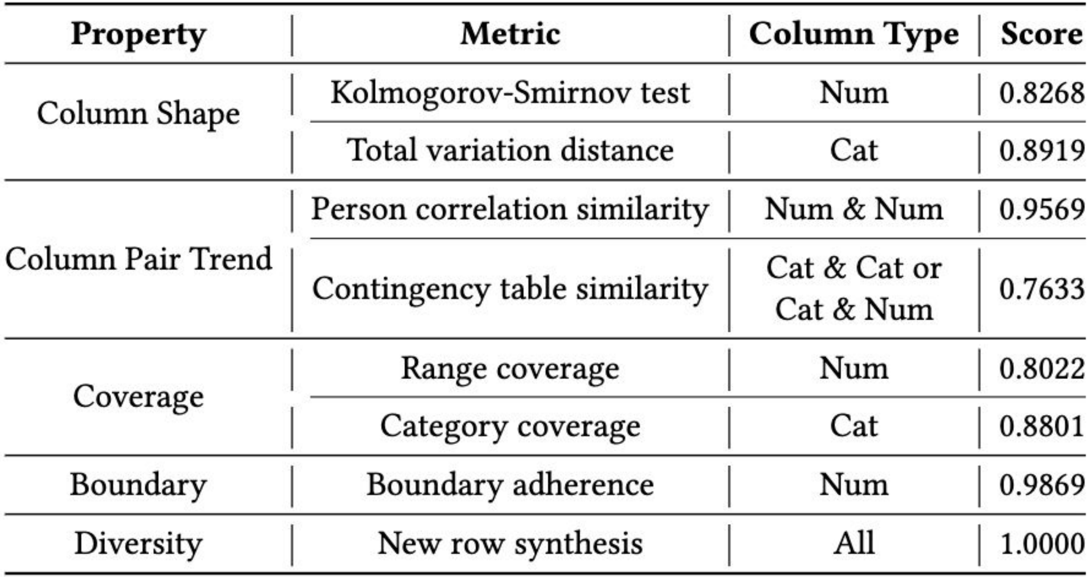
$$\text{score}(C) = 1 - \delta(f_r - f_s) = 1 - \frac{1}{2}\sum_{x \in C} |f_r(x) - f_s(x)|, \qquad (1)$$

discrete probability distribution of real and synthetic

Synthetic Data Metrics (SDMetrics), https://docs.sdv.dev/sdmetrics/

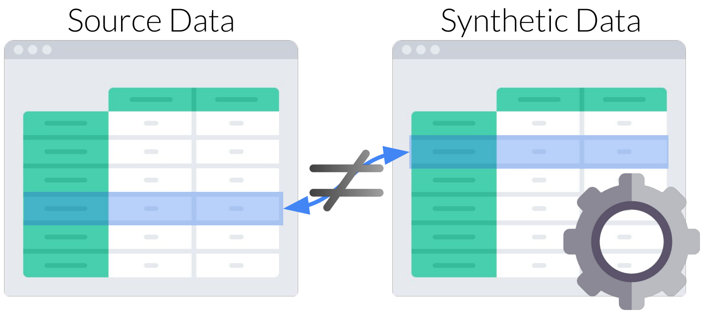
Synthetic Data Quality Metrics

A score close to 1.0 ⇒ synthetic data has good quality



Property	Metric	Column Type	Score
Column Shape	Kolmogorov-Smirnov test	Num	0.8268
	Total variation distance	Cat	0.8919
Column Pair Trend	Person correlation similarity	Num & Num	0.9569
	Contingency table similarity	Cat & Cat or Cat & Num	0.7633
Coverage	Range coverage	Num	0.8022
	Category coverage	Cat	0.8801
Boundary	Boundary adherence	Num	0.9869
Diversity	New row synthesis	All	1.0000

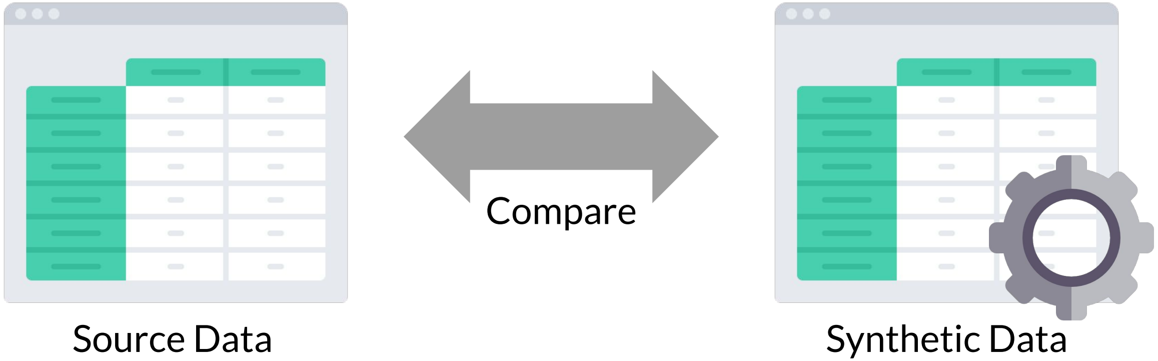
Diversity of Generated Data



Distributions of representative features are similar between synthetic data and source data

Metrics?

1.

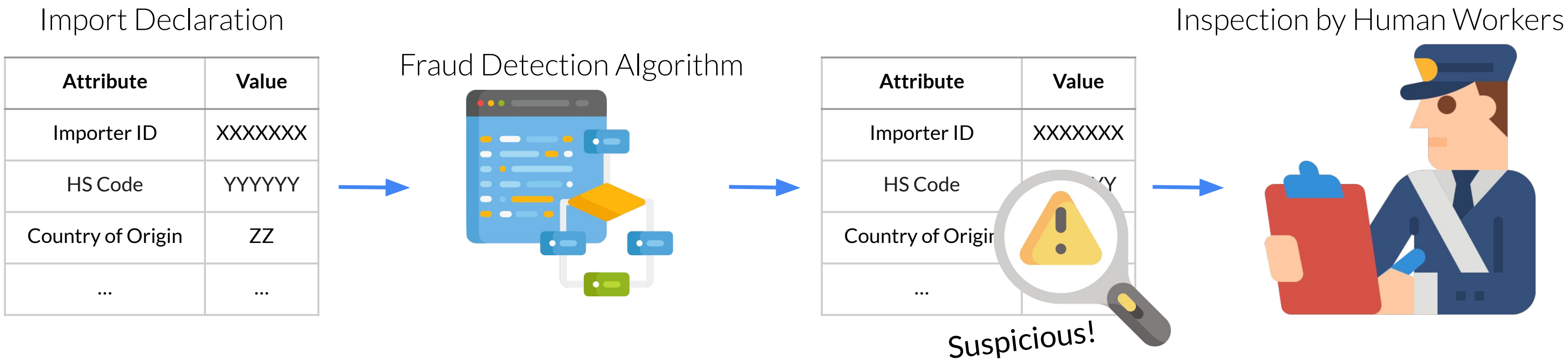


Application
- Fraud Detection

Fraud Detection

Detect suspicious imports

→ Streamline human resources for physical inspections



Import Declaration

Attribute	Value
Importer ID	XXXXXXX
HS Code	YYYYYY
Country of Origin	ZZ
...	...

Fraud Detection Algorithm

Attribute	Value
Importer ID	XXXXXXX
HS Code	YY
Country of Origi	
...	

Suspicious!

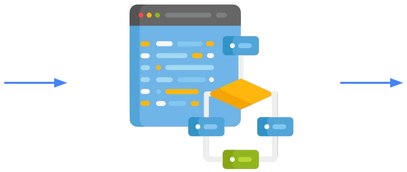
Inspection by Human Workers

Train Fraud Detection Model

Train binary classifier

Input : Import declaration data

Attribute	Value
Importer ID	XXXXXXX
HS Code	YYYYYY
Country of Origin	ZZ
…	…



Binary Classifier

Label: “Fraud”

0/1

Fraud Detection Evaluation

Precision@n%

- Measures the proportion of fraudulent items among the top-n% with highest model output
- Indicates the effectiveness in catching fraud among inspected cargos

Top-5% data among Test set
(Suspicious items)

	Model output
import 0013	0.999
import 6121	0.989
import 3302	0.976
...	...

Ground Truth

Fraud Label
1
0
0
...

Precision@5% = 0.3333

Performance Comparison

Inspection rate: n=5%

	Synthetic data	Source data
Logistic Regression	0.2759	0.3921
AdaBoost	0.3608	0.4902
Decision Tree	0.3561	0.5128
Random Forest	0.3608	0.5035
CatBoost	0.6698	0.5151
XGBoost	0.6745	0.5220
LightGBM	0.7783	0.5313

→ The synthetic data can be used as an open benchmark to develop advanced fraud detection algorithms.

Customs Import Declaration Data

An import declaration dataset synthesized with conditional tabular GAN

Advantages:

- Unrestricted disclosure
- Minimal identity risk
- Suitability for testing classification algorithms
- Facilitate research progress and effective trade control