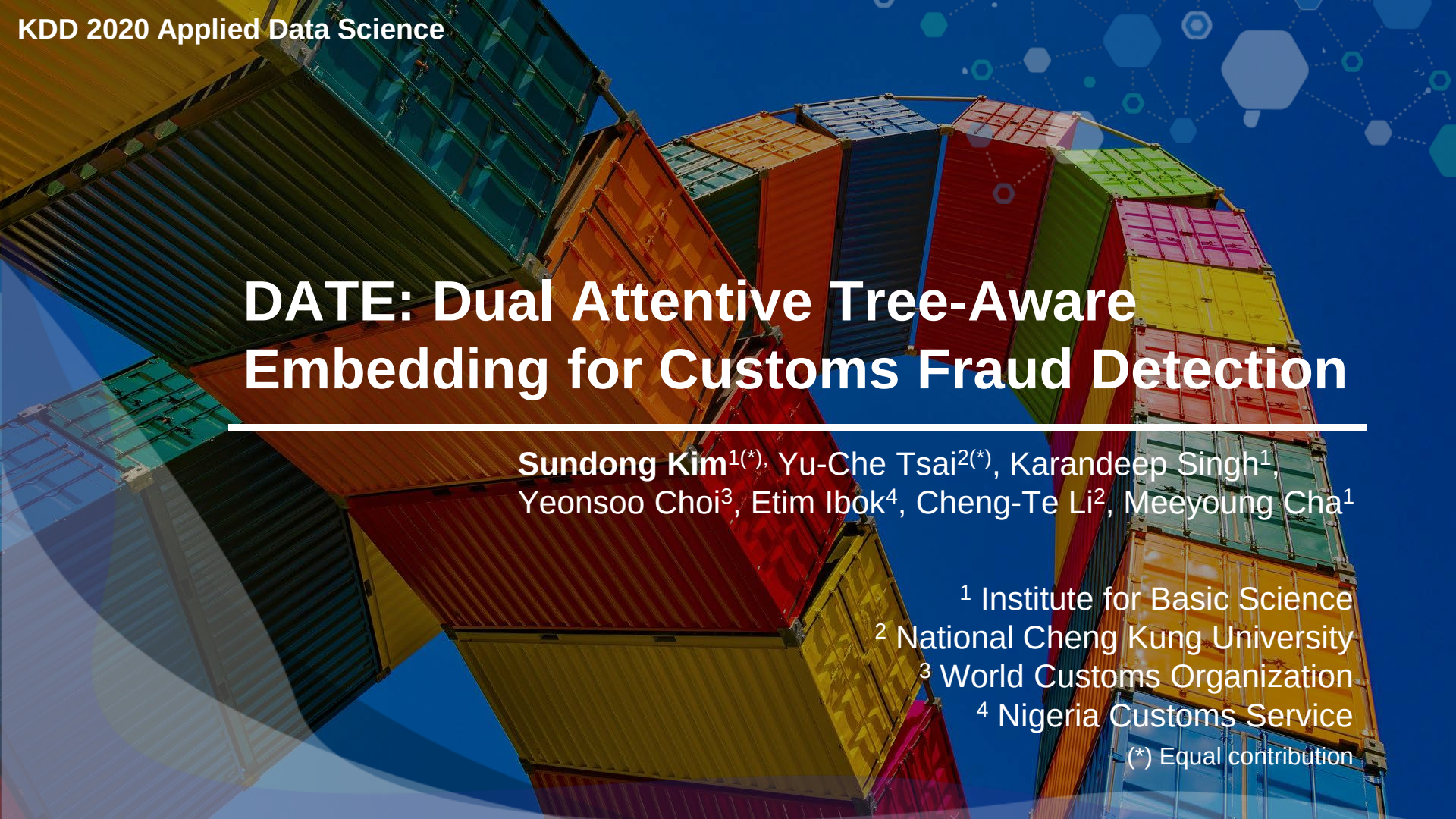


DATE: Dual Attentive Tree-Aware Embedding for Customs Fraud Detection

Sundong Kim[1(*)], Yu-Che Tsai[2(*)], Karandeep Singh[1], Yeonsoo Choi[3], Etim Ibok[4], Cheng-Te Li[2], Meeyoung Cha[1]

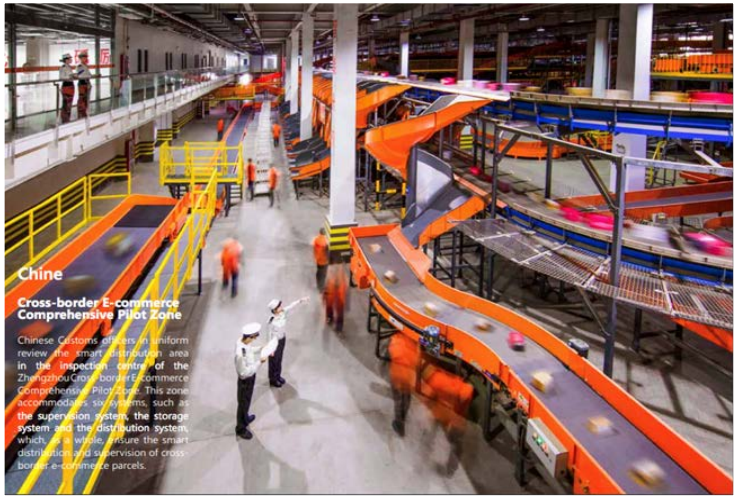
[1] Institute for Basic Science
[2] National Cheng Kung University
[3] World Customs Organization
[4] Nigeria Customs Service

(*) Equal contribution

Customs

- Customs are government authorities responsible for controlling the flow of goods and passengers across borders and collecting customs duties and taxes from traders.

Source: Official website of the General Administration of Customs of China



Amounts of Global Trades

- According to the World Customs Organization (WCO), customs organizations cleared $19.7 trillion worth of imports, 1.4 billion passengers, and collected 30% of tax revenue globally in 2018.

Control less, but better
→ **Customs Selection**

Customs Fraud and its types

- Catch-all term for frauds committed by companies and individuals trying to evade payments that are owed to the federal government in connection with the importation of goods into the country.

Examples of Customs Fraud

CUSTOMSNEWS

HEADLINES | CUSTOMS | FINANCE | REGULATIONS | ANTI-SMUGGL

ANTI-SMUGGLING

HOT ront of anti-drugs Border controls tightened to prevent illegal crossings

Make false declaration for tax fraud of billion VND

09:31 | 29/05/2020

Like Share Be the first of your friends to like this.

VCN- Being permitted to declare and take responsibility for import-export goods, but some enterprises havemade false declarations of name, code of goods for tax fraud and have been detected and hadtax arrears collected by customs.

- Hanoi: Focusing on coordination of competent forces for anti-smuggling efforts
- Risk of fraud when quota for sugar imported from ASEAN countries lifted

A Chinese shipment with false declaration seized by Ho Chi Minh Customs Department on May 15. Photo: T.H

CUSTOI

HEADLINES

ANTI-SMUGGLING

HOT ront of anti-drugs

Make false
billion VND

09:31 | 29/05/2020

Like Share Be t

VCN- Being permitte
goods, but some ent
goods for tax fraud a
customs.

Hanoi: Focusing on c

Risk of fraud when q

A Chinese shipm

Customs uncovers P34.7-M cigarettes declared as furniture from China

Published May 8, 2020, 4:10 PM

By Betheena Unite

Instead of furniture, boxes of "Two Moon" cigarettes were found inside two containers from China at the Port of Manila Friday, the Bureau of Customs said.

(BOC / MANILA BULLETIN)

CUSTON

HEADLINES

ANTI-SMUGGLING

ront of anti-drugs

Make false d
billion VND

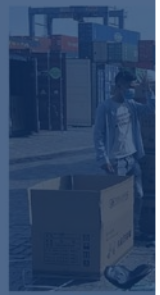
09:31 | 29/05/2020

Like Share Be the

VCN- Being permitted
goods, but some enter
goods for tax fraud an
customs.

Hanoi: Focusing on co

Risk of fraud when quo



A Chinese shipme

06/03/2020

ICE HSI Baltimore seizes over 14,000 unapproved COVID-19 treatment capsules, several unapproved test kits

Customs Fraud and its types

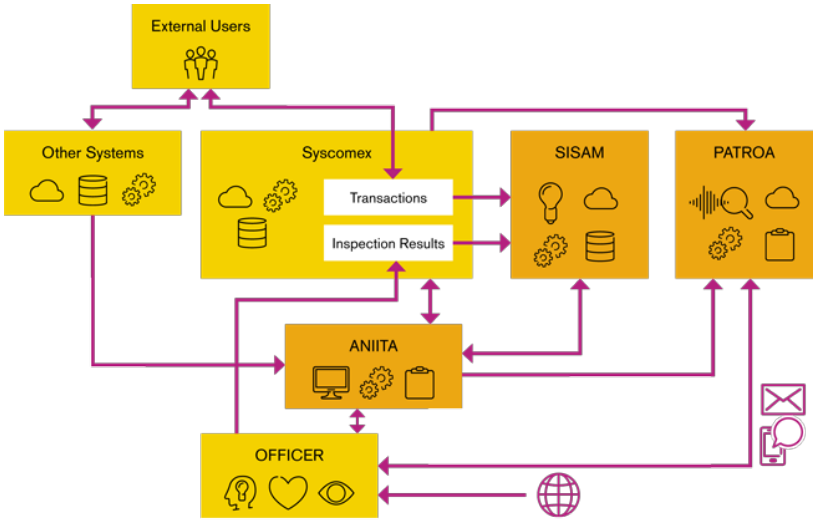
Type of Frauds	Illicit motives	Our Scope
Undervaluation of trade goods	To avoid ad-valorem customs duty, or conceal illicit financial flows from exporters	Yes
Misclassification of HS code	To get a lower tariff rate applied or trade prohibited goods by avoiding restriction	
Manipulation of origin country	To get a preferential tariff rate under a free trade agreement	
Smuggling without declaration	To trade prohibited goods by avoiding restriction and customs duties	No
Overvaluation of trade goods	To disguise illicit financial flows as legitimate trade payment from importers	

TV (HS 852859, 8% duty)
PC Monitor (HS 852852, 0% duty)

Smuggling

E-Clearance and Risk Management Solutions

- Some members have successfully adopted ML in their e-clearance system
- Many administrations are still planning or requesting international support



Brazil's new integrated risk management solutions

The UNI-PASS system of Republic of Korea

Supporting Customs With Data Analytics

Other articles in this Edition >>

Flash Info

BACUDA: supporting Customs with data analytics

By the WCO Secretariat

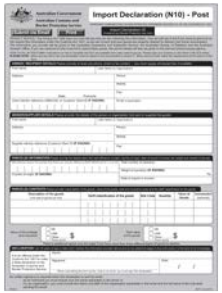
WCO Members asked the Secretariat to place a new focus on the development of guidance and capacity to support the use of data analytics. As one of the responses, a team of experts was put in place under a project called BACUDA. The project's name is an acronym, which stands for "BAnd of CUstoms Data Analysts." It is also a Korean word that means "to change." Indeed, the aim of the project is to help Customs administrations in embracing analytical tools and methodologies, a major move for many.

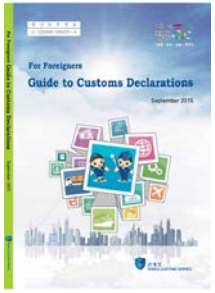
BACUDA team members are all data experts with whom the Secretariat has been collaborating for some years. They are Customs officials in charge of risk management, statistics and IT systems, as well as professional economists and data scientists with an academic background in computer science. Data scientists of various nationalities from the Institute of Basic Science (IBS), the Korea Advanced Institute of Science and Technology (KAIST), and the National Cheng Kung University (NCKU) are involved in the project and leading the development of state-of-the-art algorithms. However, any qualified data experts working in Customs administrations or in academia may join the BACUDA team.

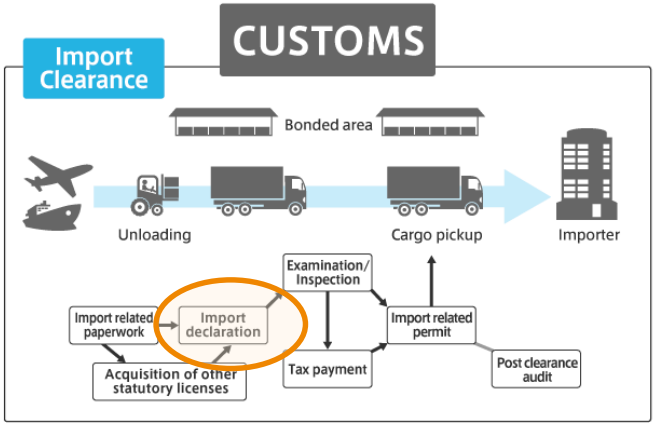
https://mag.wcoomd.org/magazine/wco-news-91-february-2020/bacuda/

Format of Import Declarations

ASYCUDA
Automated SYstem for CUstoms DAta







Type	Variable	Description	Example
Features	*sgd.id*	An individual numeric identifier for Single Goods Declaration (SGD).	SGD347276
	sgd.date	The year, month and day on which the transaction occurred.	13-11-28
	importer.id	An individual identifier by importer based on the tax identifier number (TIN) system.	IMP364856
	declarant.id	An individual identification number issued by Customs to brokers.	DEC795367
	country	Three-digit country ISO code corresponding to transaction.	USA
	office.id	The customs office where the transaction was processed.	OFFICE91
	tariff.code	A 10-digit code indicating the applicable tariff of the item based on the harmonised system (HS).	8703232926
	quantity	The specified number of items.	1
	gross.weight	The physical weight of the goods.	150kg
	fob.value	The value of the transaction excluding, insurance and freight costs.	$350
	cif.value	The value of the transaction including the insurance and freight costs.	$400
	total.taxes	Tariffs calculated by initial declaration.	$50
Prediction Target	***illicit***	Binary target variable that indicates whether the object has fraud.	1
	revenue	Amount of tariff raised after the inspection, only available on some illicit cases.	$20

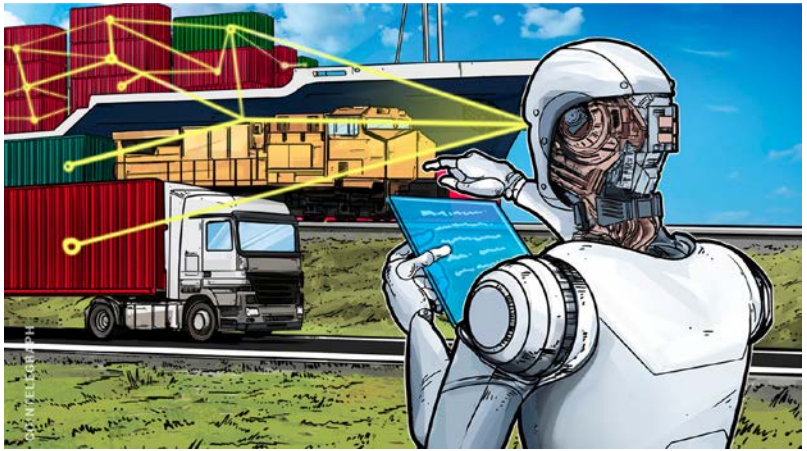
Customs Selection by Detecting Frauds - Input Features

Features

10,000 items

Item	Declared Price	Importer
Ferrari 488 Spider	$50,000	John
Porsche 911	$30,000	Jane
	...	...
Hyundai Elantra	$15,000	Kim
	...	...
Water 2Lx12	$6	Wang

Goal: Select 10% of items that requires inspection



Customs Selection by Detecting Frauds - Making predictions

Features

Prediction

Item	Declared Price	Importer	Predicted Illicit prob	Ranking (Priority)
Ferrari 488 Spider	$50,000	John	0.99	1
Porsche 911	$30,000	Jane	0.98	2
	...	...	...	...
Hyundai Elantra	$15,000	Kim	0.01	1000
	...	...	...	...
Water 2Lx12	$6	Wang	0.00	10000

Goal: Select 10% of items that requires inspection

Customs Selection by Detecting Frauds - Inspection Results

Features **Prediction** **Inspection results and labels obtained**

Item	Declared Price	Importer	Predicted Illicit prob	Ranking (Priority)	Inspection Results	Illicit	Revenue
Ferrari 488 Spider	$50,000	John	0.99	1	Under-invoiced, Original price: $350,000, Rate applied: 30%, Additional duties: $90,000	1	$90,000
Porsche 911	$30,000	Jane	0.98	2	Original price: $130,000, 30% rate -> Surtax applied for $100k	1	$30,000
	...	...	...	...	...	...	...
Hyundai Elantra	$15,000	Kim	0.01	1000	CLEARED – Right price for the used car, Odometer: 25300km,	0	$0
	...	...	...	...	...	...	...
Water 2Lx12	$6	Wang	0.00	10000	CLEARED	0	$0

Evaluation Metrics

- **Precision@n%**: How many inspected items are illicit?
- **Recall@n%**: How many inspected items are screened out of all frauds?
- **Revenue@n%**: How much worth that inspected items have?
- **AUC, F1-score**: Metric for overall prediction results

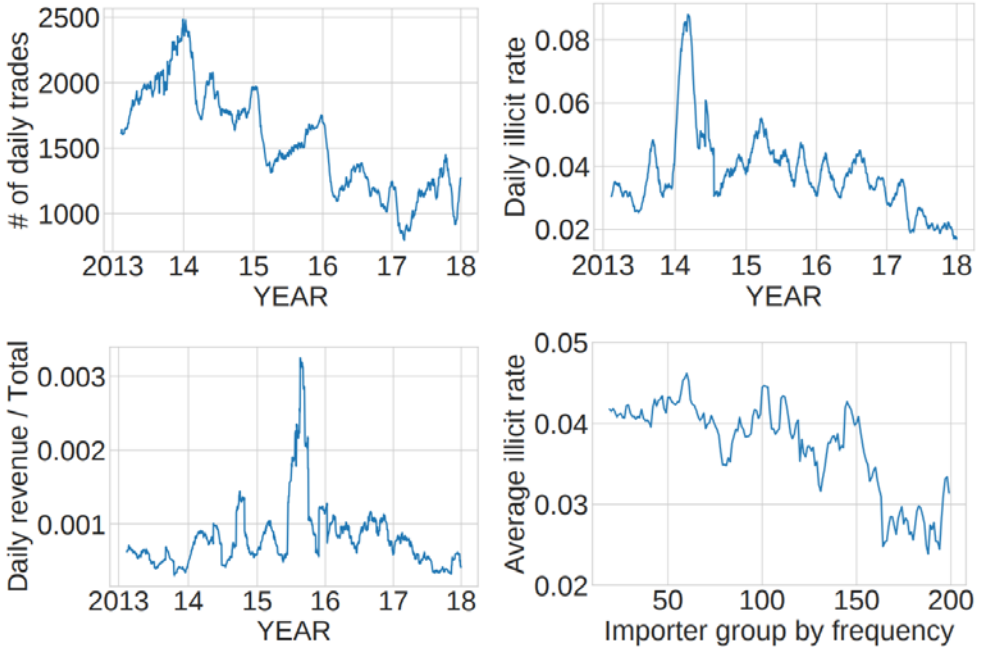
Features | Prediction | Compare with inspection results

Item	Declared Price	Importer	Predicted Illicit prob	Ranking (Priority)	Inspection Results	Illicit	Revenue
Ferrari 488 Spider	$50,000	John	0.99	1	Under-invoiced, Original price: $350,000, Rate applied: 30%, Additional duties: $90,000	1	$90,000
Porsche 911	$30,000	Jane	0.98	2	Original price: $130,000, 30% rate -> Surtax applied for $100k	1	$30,000
	...	...	...	...	...	...	...
Hyundai Elantra	$15,000	Kim	0.01	1000	CLEARED – Right price for the used car, Odometer: 25300km,	0	$0
	...	...	...	...	...	...	...
Water 2Lx12	$6	Wang	0.00	10000	CLEARED	0	$0

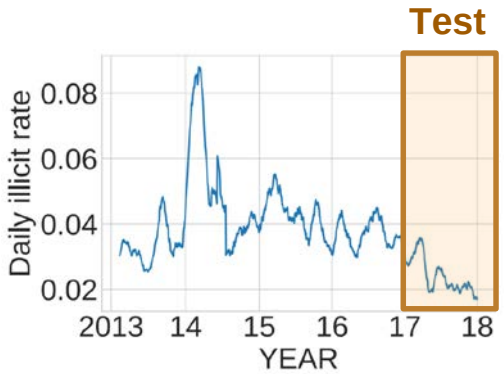
Statistics of Nigeria Customs

Currently achieving
100% inspection rate

More than 1.9M import trade flows
from 2013 to 2017



Experimental Settings



- Offline testing: Use five-year imports

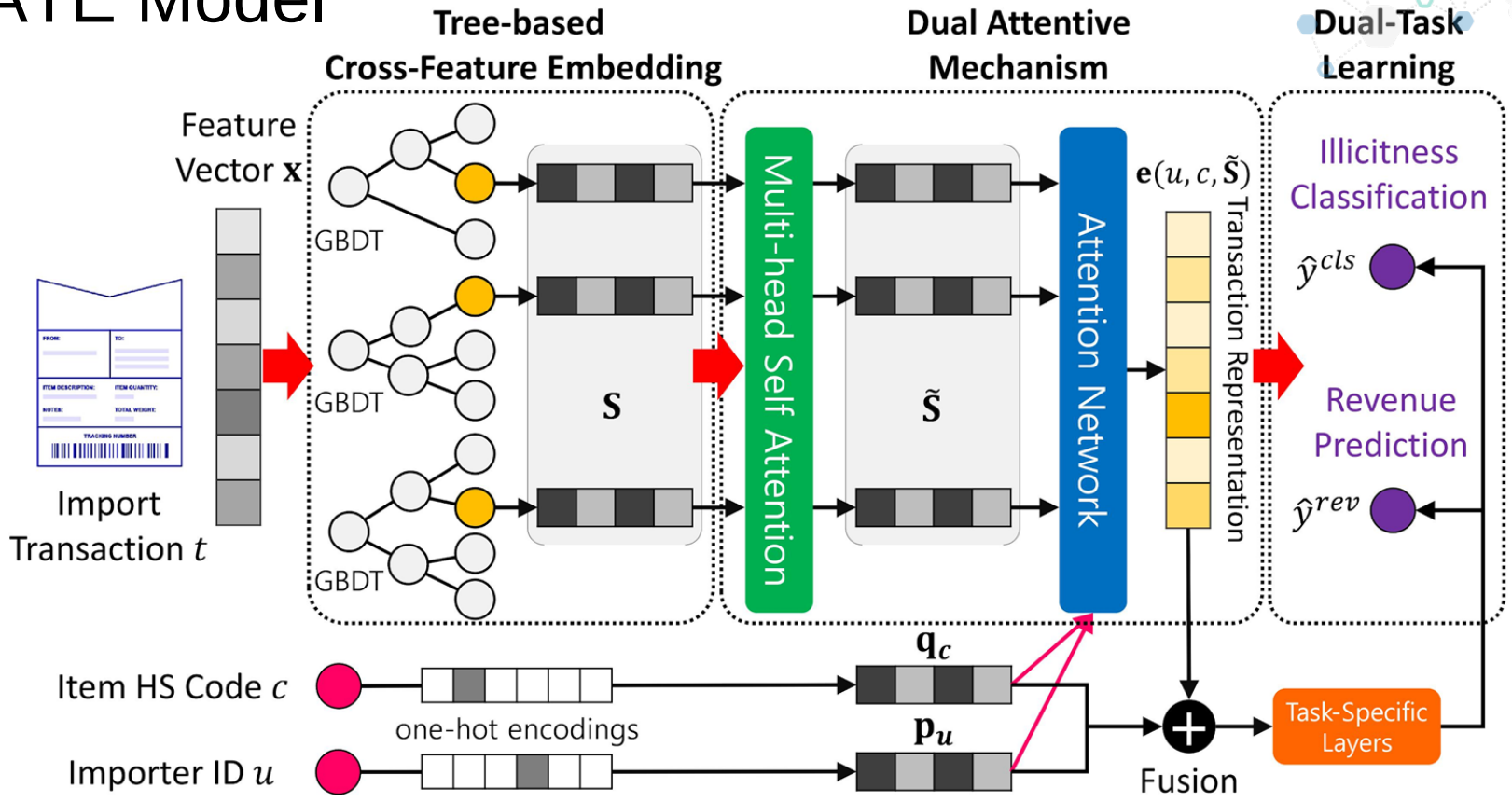
- Online testing & Deployment plan: The test will be conducted in the two major ports in Nigeria - Tincan and Onne port

  - 1st phase: Receive weekly data and provide prediction results to NCS. The model's predictions would be matched against the corresponding inspection results (Ongoing – From Mar 2020 to June 2020)

  - 2nd phase: Deploy our model in a live system, inform the model's predictions to officers from NCS before inspections and check whether they perform better in detecting frauds

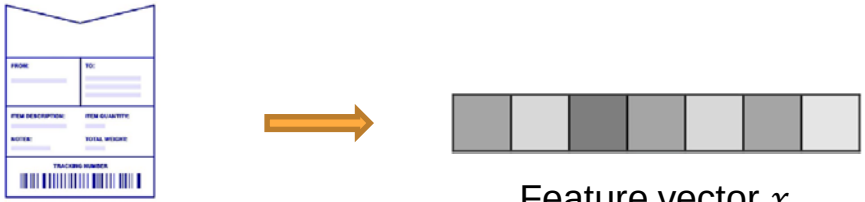
  - 3rd phase: Reduce the number of inspections based on the model's prediction, we can measure: 'average clearance time' and 'reduced cost for inspection'

DATE Model


Tree-based
Cross-Feature Embedding
Dual Attentive
Mechanism
Dual-Task
Learning
Feature
Vector x
GBDT
GBDT
GBDT
S
Multi-head Self Attention
S̃
Attention Network
e(u, c, S̃)
Transaction Representation
Illicitness
Classification
ŷ^cls
Revenue
Prediction
ŷ^rev
Import
Transaction t
Item HS Code c
one-hot encodings
Importer ID u
q_c
p_u
Fusion
Task-Specific
Layers

Feature Engineering

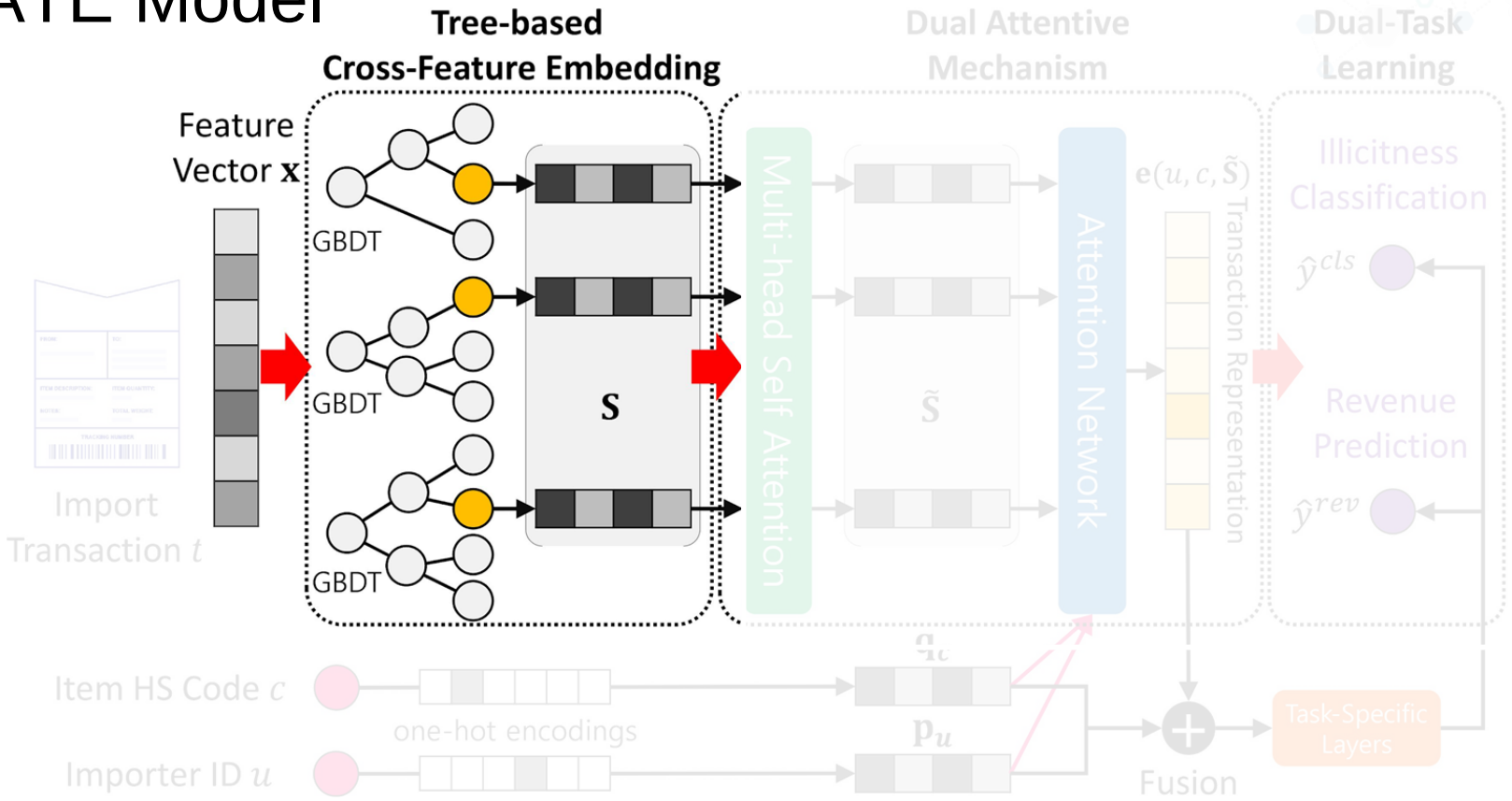
- Each transaction consists of several numeric features and categorical features
  - Numeric features: Quantity, Gross weight, FOB value, CIF value, Total taxes, Temporal features
  - Categorical features: Importer ID, Declarant ID, Tariff code (HS code), Country, Office ID
- Non-linear relationships: Unit value, Value/kg, Unit tax, Face ratio
- Converted some categorical features to binary variables by quantifying its risk indicators (Importer ID, Declarant ID, HS code, Countries of origin)



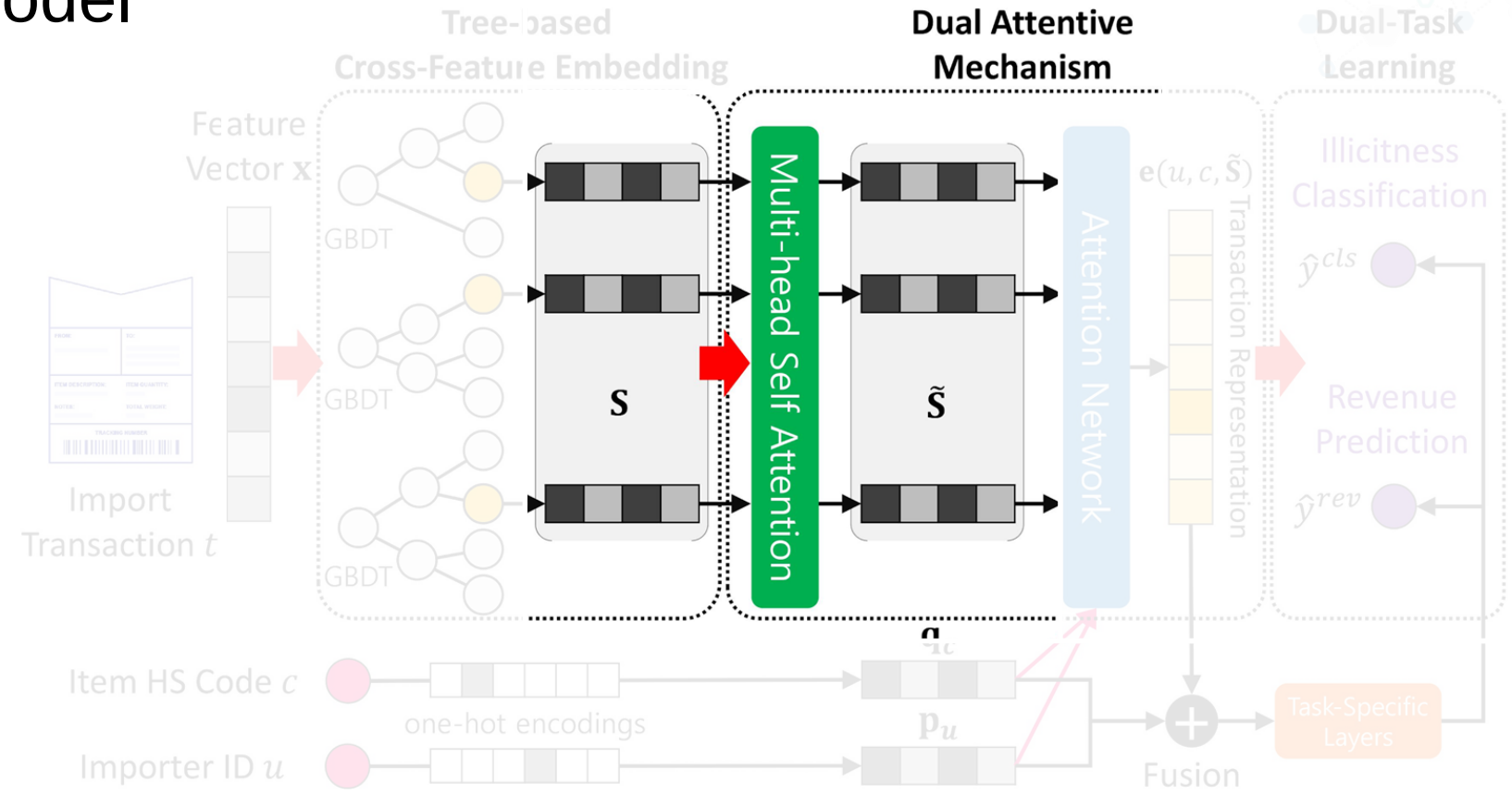
Import Transaction t

Feature vector x

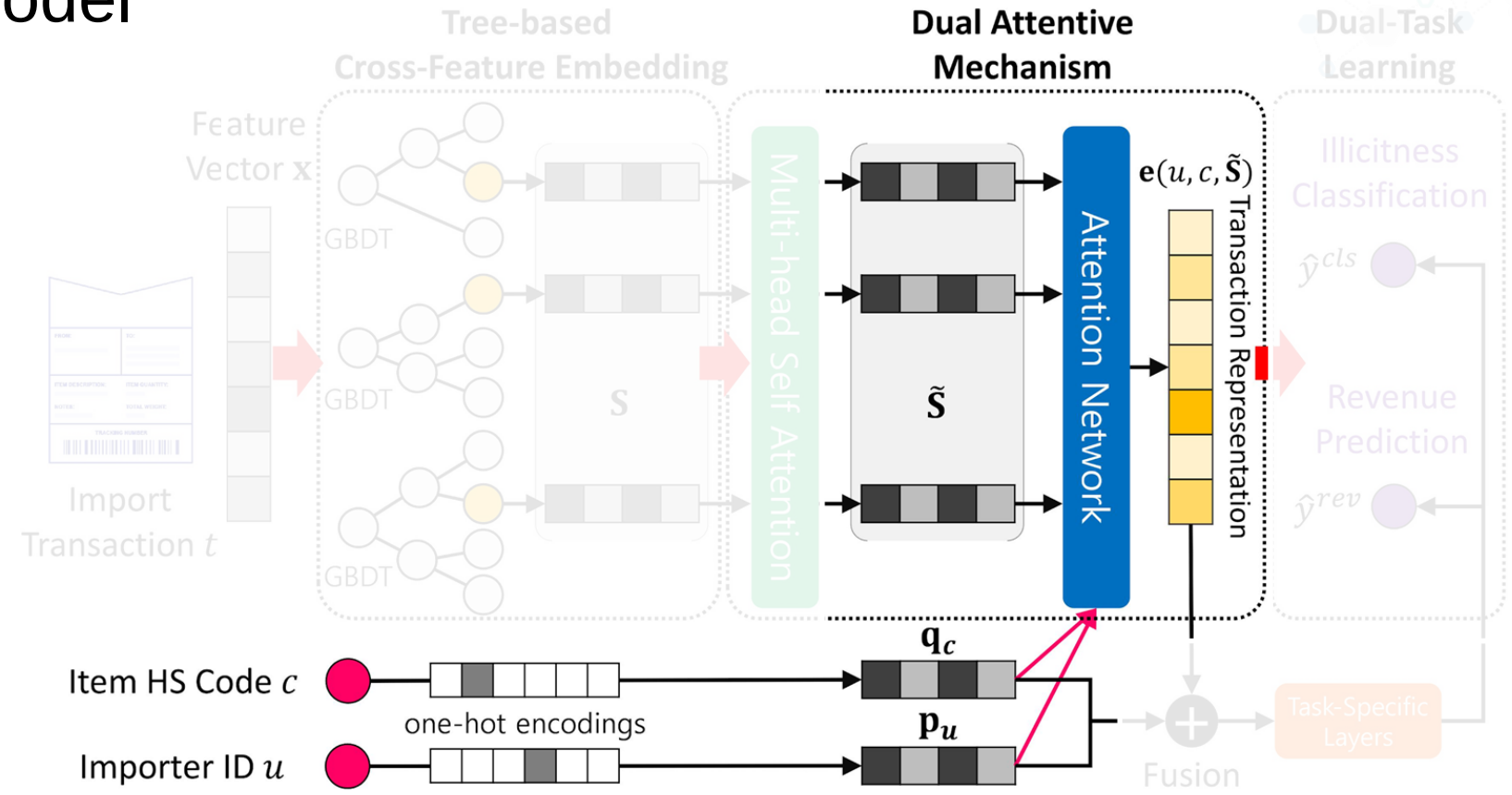
DATE Model


Tree-based
Cross-Feature Embedding
Feature
Vector x
GBDT
GBDT
GBDT
S
Multi-head Self Attention
Dual Attentive
Mechanism
Attention Network
Transaction Representation
Dual-Task
Learning
Illicitness
Classification
Revenue
Prediction
Import
Transaction t
Item HS Code c
one-hot encodings
Importer ID u
Fusion
Task-Specific
Layers

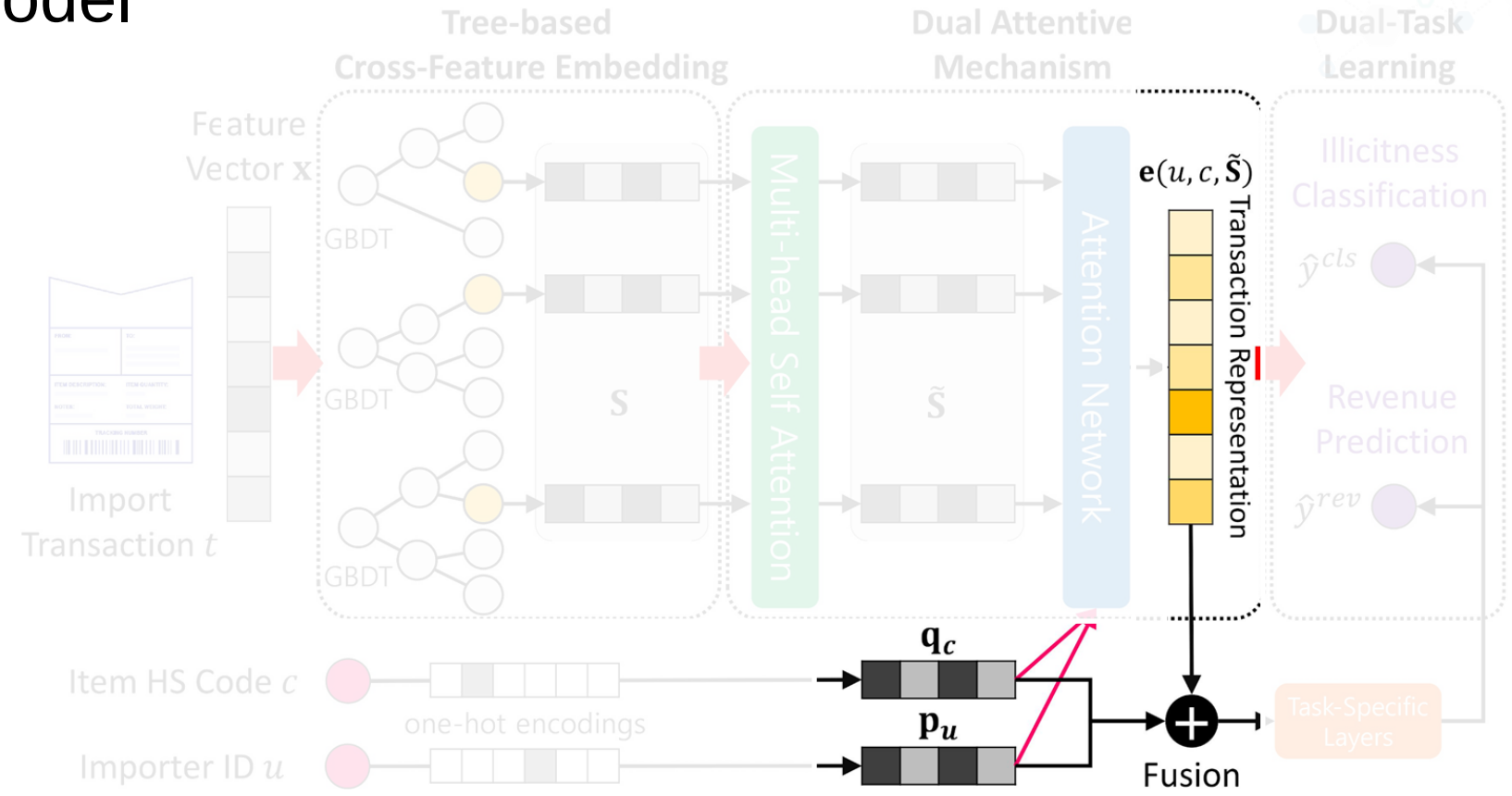
Model


Tree-based
Cross-Feature Embedding
Dual Attentive
Mechanism
Dual-Task
Learning
Feature
Vector x
GBDT
GBDT
GBDT
S
Multi-head Self Attention
S̃
Attention Network
e(u, c, S̃)
Transaction Representation
Illicitness
Classification
ŷ^cls
Revenue
Prediction
ŷ^rev
Import
Transaction t
Item HS Code c
one-hot encodings
Importer ID u
p_u
Fusion
Task-Specific
Layers

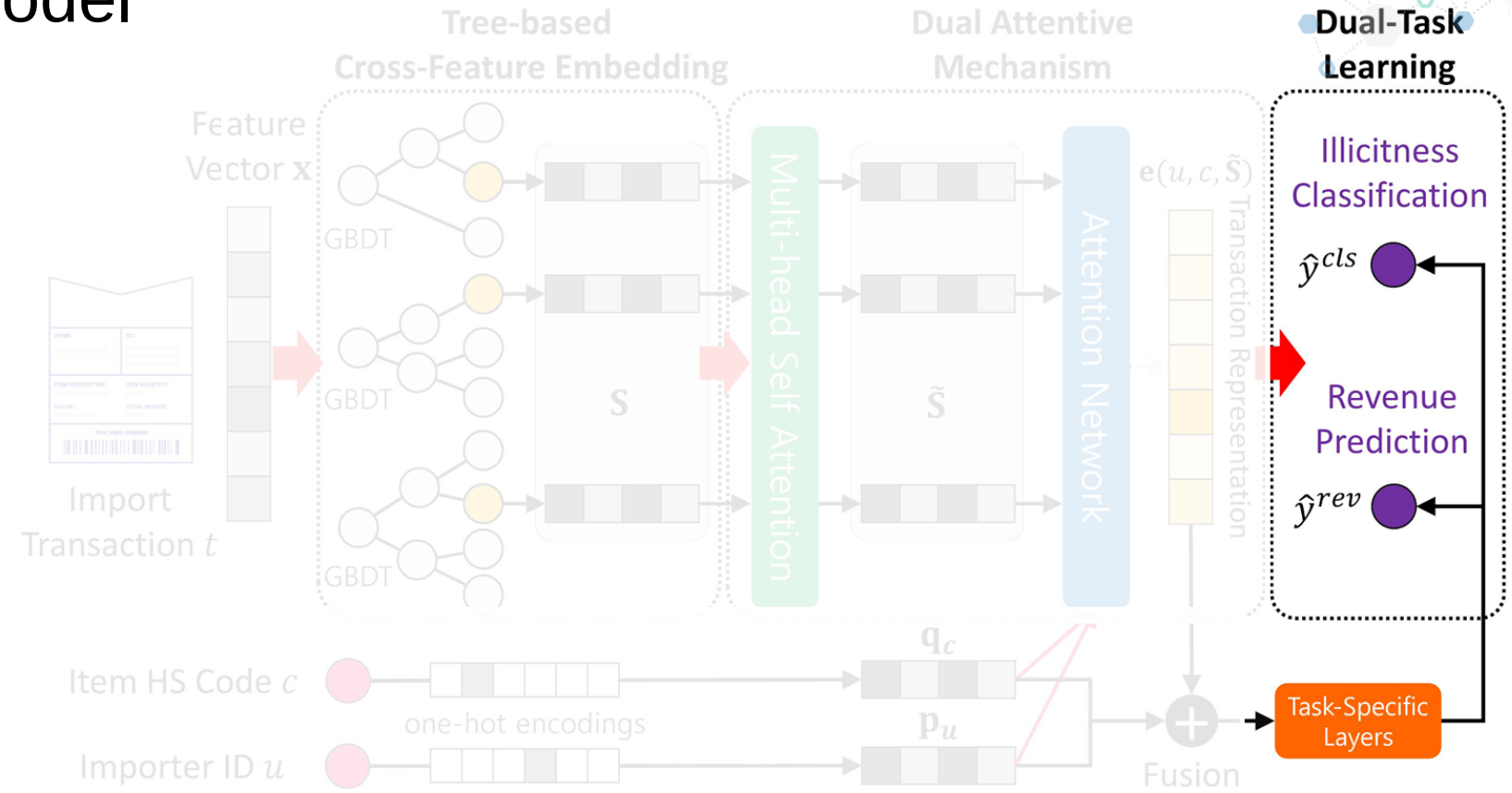
Model


Tree-based
Cross-Feature Embedding
Dual Attentive
Mechanism
Dual-Task
Learning
Feature
Vector x
GBDT
GBDT
GBDT
S
Multi-head Self Attention
Attention Network
e(u, c, S̃)
Transaction Representation
Illicitness
Classification
Revenue
Prediction
Import
Transaction t
Item HS Code c
one-hot encodings
Importer ID u
q_c
p_u
Fusion
Task-Specific
Layers

Model


Tree-based
Cross-Feature Embedding
Dual Attentive
Mechanism
Dual-Task
Learning
Feature
Vector x
GBDT
GBDT
GBDT
S
Multi-head Self Attention
Attention Network
e(u, c, S̃)
Transaction Representation
Illicitness
Classification
Revenue
Prediction
Import
Transaction t
Item HS Code c
Importer ID u
one-hot encodings
q_c
p_u
Fusion
Task-Specific
Layers

Model



$$\mathcal{L}_{DATE} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{rev} + \lambda\|\Theta\|^2$$

Results – Comparison with Baselines

- Price/Importer: Select in the order of the declared price/fraud rate of importers
- IForest (Liu et al, 2008): Tree-based anomaly detection algorithm
- GBDT (Chen et al, 2016): XGBoost with cross features, trained on binary label y^{cls}
- GBDT+LR (He et al, 2014): Logistic regression based on cross features
- TEM (Wang et al, 2018): Tree-enhanced model with attentions.
- **DATE_{CLS}, DATE_{REV}**: Custom selection with $\hat{y}^{cls}, \hat{y}^{rev}$ of DATE, respectively

	n = 1% (Selecting top 1%)			n = 2%			n = 5%			n = 10%			Overall	
Model	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.	AUC	F1
Price	2.75%	1.23%	15.17%	2.23%	1.99%	20.64%	2.06%	4.60%	34.95%	2.30%	10.28%	50.98%	67.57%	7.81%
Importer	11.43%	5.10%	4.36%	9.41%	8.39%	7.56%	6.47%	14.43%	13.18%	5.22%	23.31%	30.31%	59.20%	9.10%
IForest	5.61%	2.50%	14.30%	6.19%	5.52%	23.14%	5.66%	12.62%	40.62%	5.12%	22.85%	54.14%	66.89%	5.28%
GBDT	90.01%	40.15%	24.59%	66.16%	59.04%	38.89%	32.19%	71.80%	57.20%	17.58%	78.42%	66.86%	93.38%	63.69%
GBDT+LR	90.95%	40.40%	27.18%	72.94%	65.09%	44.22%	35.02%	78.11%	63.77%	18.72%	83.54%	73.77%	94.82%	68.76%
TEM	88.72%	39.59%	39.48%	74.70%	66.43%	58.48%	37.39%	83.41%	78.58%	19.91%	88.54%	85.02%	96.52%	70.55%
DATE_{CLS}	**92.66%**	**41.33%**	44.97%	**80.79%**	**72.05%**	67.14%	**38.77%**	**86.49%**	84.35%	**20.24%**	**90.29%**	89.03%	**96.79%**	**75.32%**
DATE_{REV}	82.25%	36.63%	**49.29%**	79.93%	71.22%	**68.48%**	38.74%	86.41%	**84.57%**	20.11%	89.74%	**89.2%**	95.66%	75.23%

Results – Ablation Studies

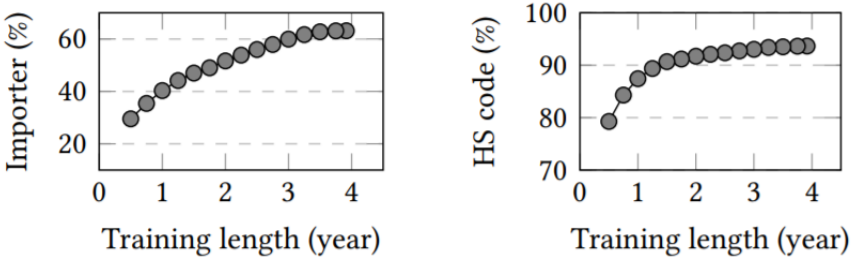
- **Full Model:** DATE_{CLS} using all components
- **w/o MSA** (Multi-head self attention): Ignore the relations between cross-features
- **w/o FHI** (Fusion with HS & Importer ID): Ignore the relations with two features
- **w/o DTL** (Dual-task learning): Train the model only on the binary labels
- **w/o AN** (Attention network): Treat each cross feature equally

	$n = 1\%$		$n = 5\%$		$n = 10\%$	
Model	Pre.	Rev.	Pre.	Rev.	Pre.	Rev.
Full Model	**92.66%**	**44.97%**	**38.77%**	**84.35%**	**20.24%**	**89.03%**
w/o MSA	91.83%	41.22%	38.47%	82.64%	20.17%	86.34%
w/o FHI	89.89%	27.91%	36.03%	80.39%	19.12%	78.79%
w/o DTL	90.72%	35.11%	37.57%	78.46%	19.74%	85.25%
w/o AN	91.58%	40.20%	38.11%	80.54%	19.02%	87.69%

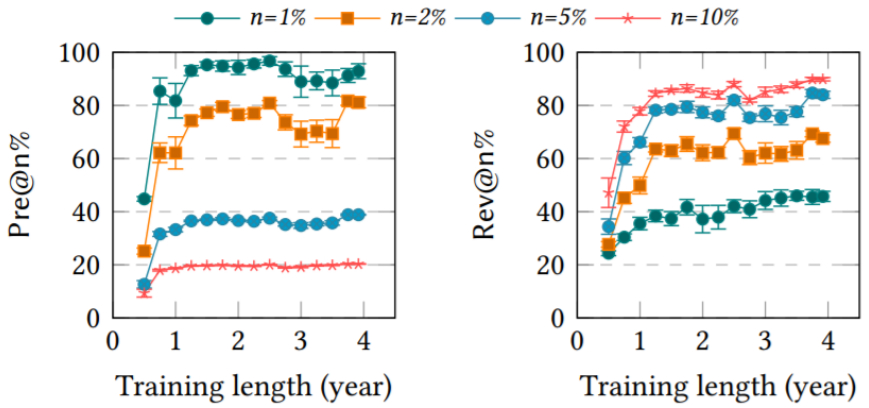
Results – Effects on Training Length

- Most recent k months are used for training
- Performance increases rapidly until about two years of past data are secured.

Coverage of two test attributes in training set



Performance by changing training length



Results – Performance on Subgroups

- We break down the test set into several subgroups (Importer, HS code)
- $IMP_{[0]}$ denotes the new importers, $IMP_{(0,\ 100]}$ denotes the group of importers who appeared, but less than or equal to 10 times.

	$n = 1\%$			$n = 5\%$			
Subgroup	Pre.	Rec.	Rev.	Pre.	Rec.	Rev.	Illicit rate
$Imp_{[0]}$	100.00%	39.27%	37.87%	45.72%	89.94%	87.41%	2.51%
$Imp_{(0,\ 10]}$	98.84%	41.47%	35.60%	42.44%	89.18%	81.57%	2.37%
$Imp_{(10,\ 50]}$	98.86%	37.25%	35.44%	46.43%	87.46%	80.43%	2.65%
$Imp_{(50,\ 250]}$	91.72%	46.16%	38.64%	32.76%	82.31%	76.35%	1.99%
$Imp_{(250,\ \infty)}$	72.76%	45.24%	53.10%	23.80%	73.90%	79.39%	1.60%
$HS_{[0]}$	97.09%	27.52%	29.28%	51.21%	72.83%	70.66%	3.51%
$HS_{(0,\ 312]}$	91.50%	41.00%	53.74%	35.40%	79.35%	88.15%	2.23%
$HS_{(312,\ 1781]}$	96.20%	44.48%	40.15%	39.45%	91.22%	84.08%	2.16%
$HS_{(1781,\ 8714]}$	98.19%	65.65%	54.79%	28.31%	94.67%	82.35%	1.49%
$HS_{(8714,\ \infty)}$	99.81%	55.47%	56.21%	33.16%	92.16%	96.00%	1.17%

Results – Qualitative Studies

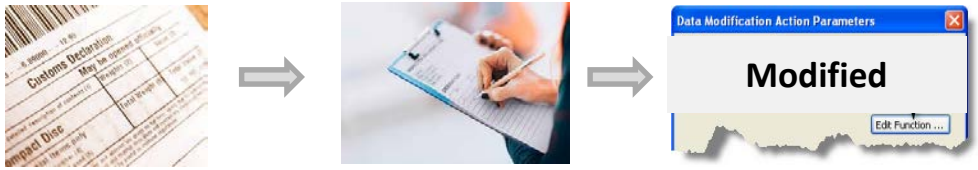
- To show that DATE has some potential to achieve **human-level interpretability**
- Select the top-2 significant cross features based on the highest attention scores

Comparison of illicit and licit transaction with respect to their corresponding cross features (CF) with highest attention score

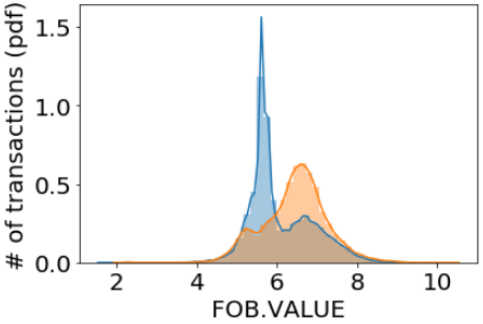
	Illicit case	Licit case
Item	Used TOYOTA VENZA, $16,863	Used TOYOTA CAMRY, $4,673
CF 1	risk.importer=0 & tax.ratio<43.7% & gross.weight<3327.43 & fob.value>$1,366	12.2%<tax.ratio<16.8% & face.ratio>62.5%
CF 2	value/kg>$2 & cif.value>$1,912 & risk.(office,importer)=0 & tax.ratio <0.18%	risk.HS.origin=0 & value/kg<$2 & cif.value>$1,640 & risk.(office,importer)=0
$\hat{y}^{cls}$	0.9849	0.0001

Results – Avoiding Potential Leakage

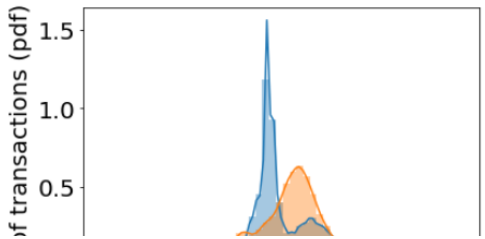
- **Problem**: Some initial feature values are adjusted after the inspection, which might affect the performance



- **Solution**: Rescale average price by multiplying with $X = \mathrm{N}(0.6, 0.1^2)$




```
illicit
0    9.386096e+06
1    1.559226e+07
Name: FOB.VALUE, dtype: float64
```



Rescaling	Pre@2%	Pre@5%	Rev@1%	Rev@2%	Rev@5%	Rev@10%
None	81.31%	38.77%	44.55%	66.56%	83.98%	89.30%
Stochastic	78.48%	36.96%	42.38%	63.04%	77.11%	83.90%

Conclusions

What is the DATE model? DATE is a customs selection model that ranks trade flows in the order of fraud risk and maximizes customs revenue

Is it interpretable? Yes, because of its decision rule from GBDT and weights from the attention mechanism.

Is it effective? Yes, we confirm the superiority of DATE over state-of-the-art models including TEM, and it is robust against noise in input data

Will it be used in Customs? Yes, we open-sourced the code and tutorials, and we are actively participating in seminars by WCO and its BACUDA initiative. Several countries joined BACUDA, and we are testing our algorithm in Nigeria Customs.

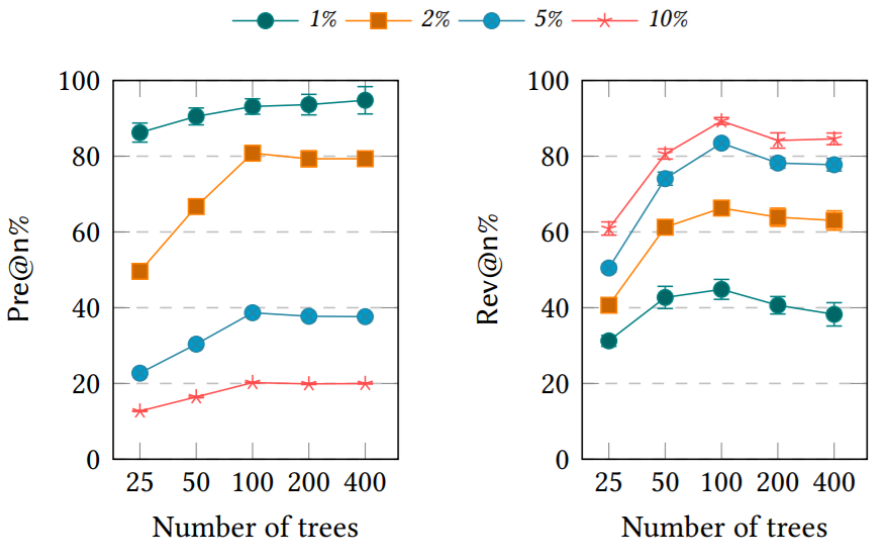
Thank you

Codes and tutorials are available on:
http://bit.ly/kdd20-date

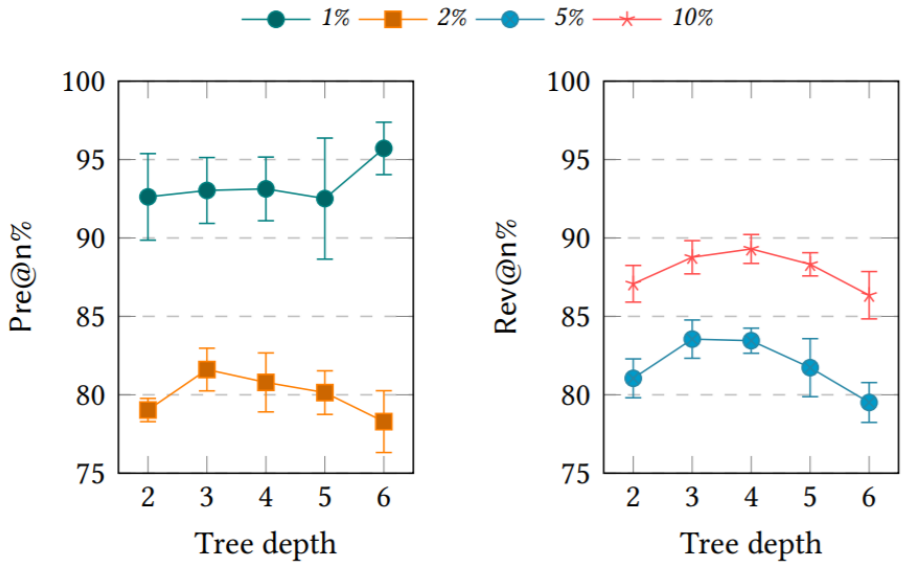
Promotional video:
https://bit.ly/kdd20-date-promo-video

Results – Hyperparameter Analysis

Tree numbers in GBDT



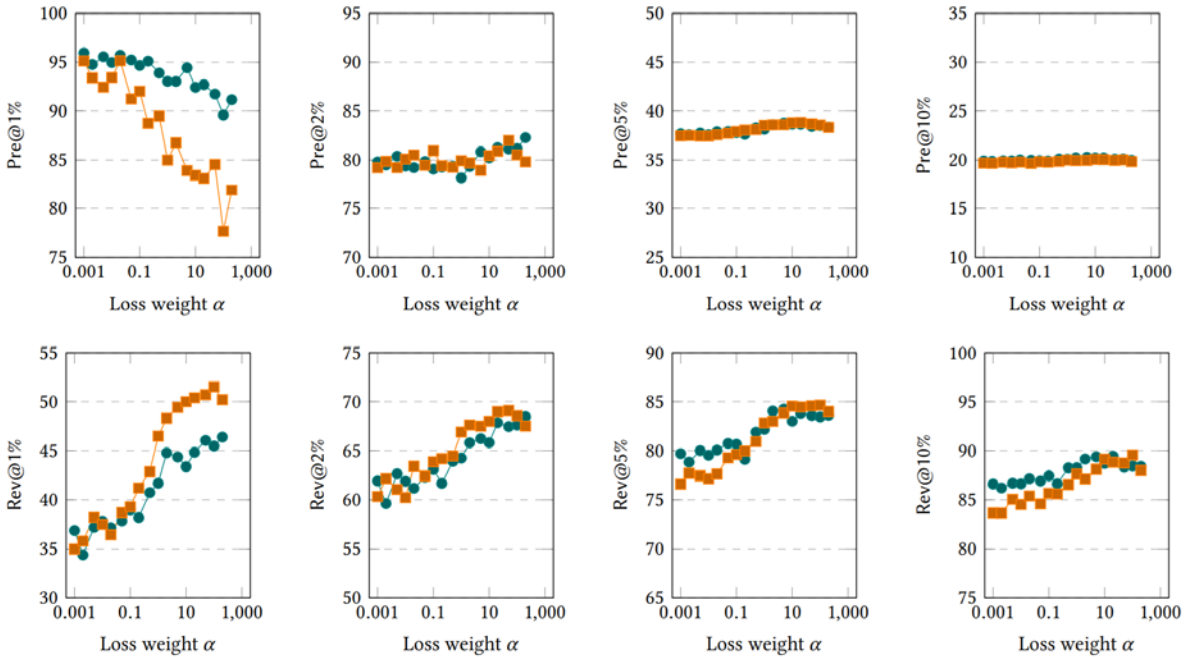
Tree depth



Results – Hyperparameter Analysis

Effect by controlling loss weight α

$$\mathcal{L}_{DATE} = \mathcal{L}_{cls} + \alpha\mathcal{L}_{rev} + \lambda\|\Theta\|^2$$



Embedding dimension