IMI BIGDataAIHub | Scotiabank
Big Data & Artificial Intelligence Case
Team 16 - SignalWeave

Scotiabank BIGDataAIHUB Institute for Management & Innovation UNIVERSITY OF TORONTO MISSISSAUGA

Team Introduction

Shadi Abdul-Sater

MBiotech, HBSc

Owais Amir

MBiotech, BHSc

Anthony Pede

MBiotech, HBSc

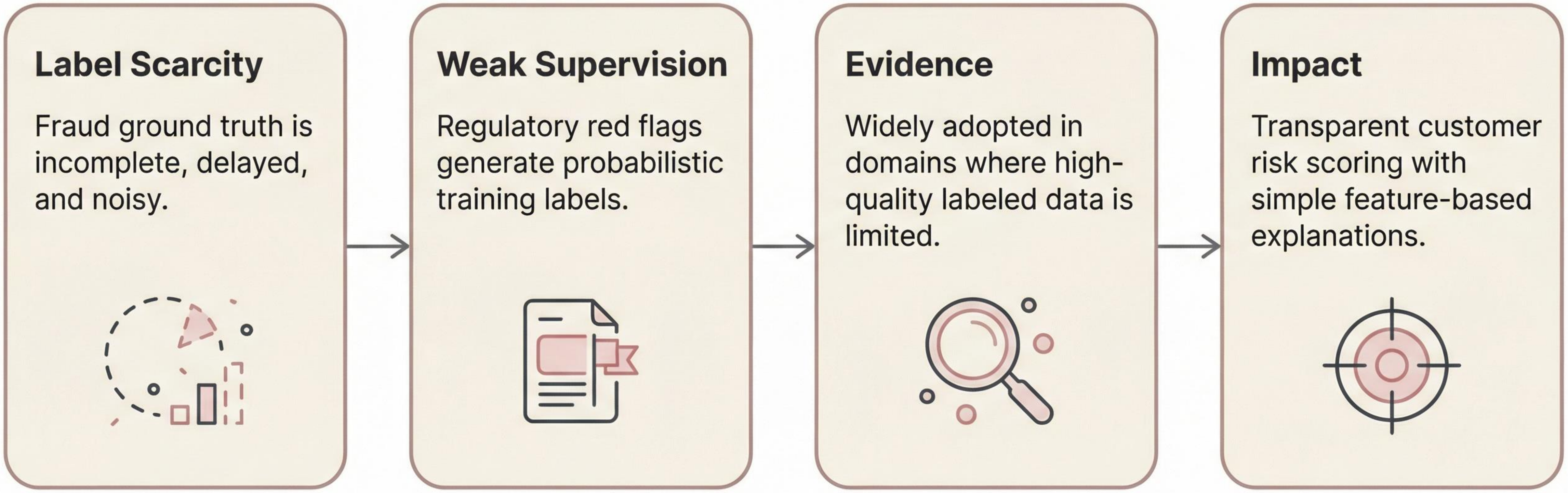
Alexander Turco

MSc, BSc

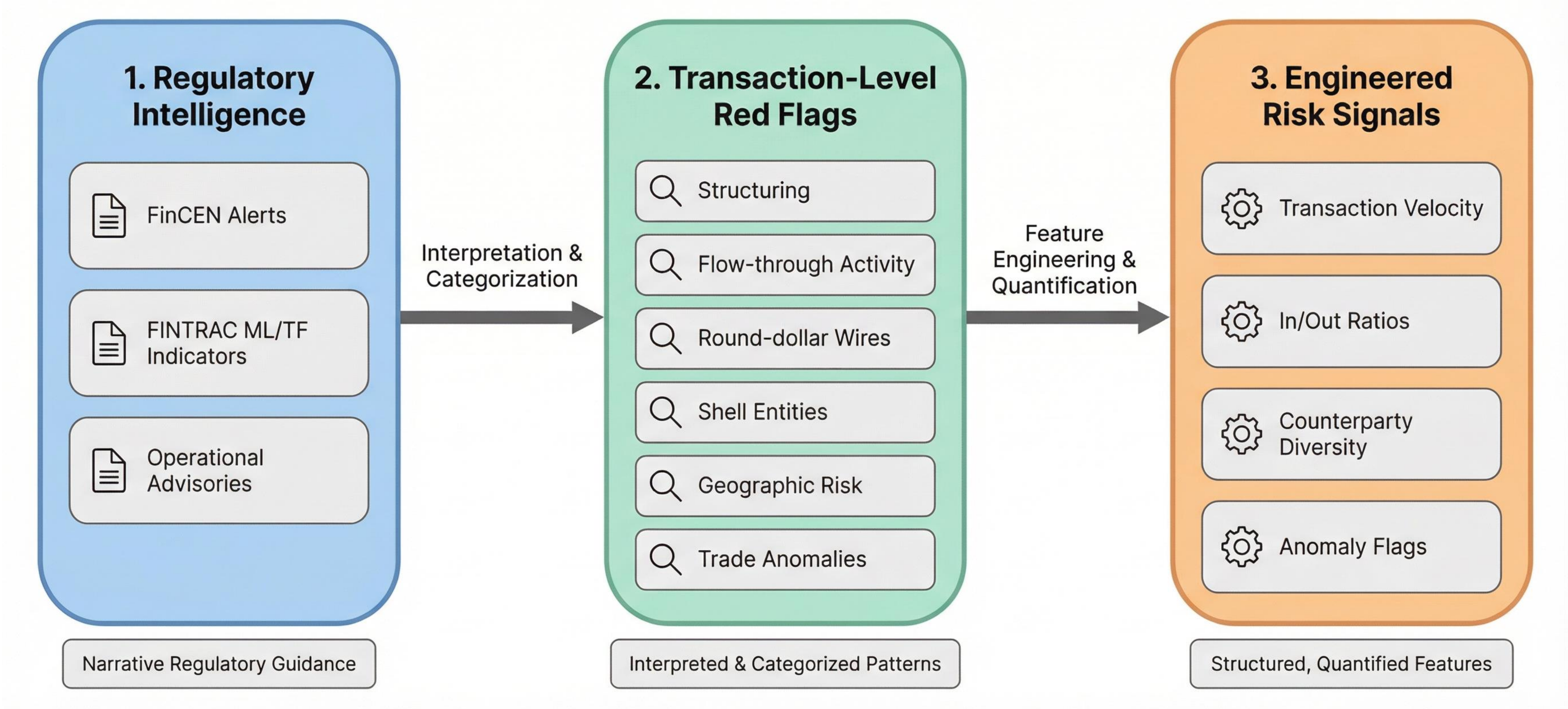
Mingjie Zhao

MSc, BSc

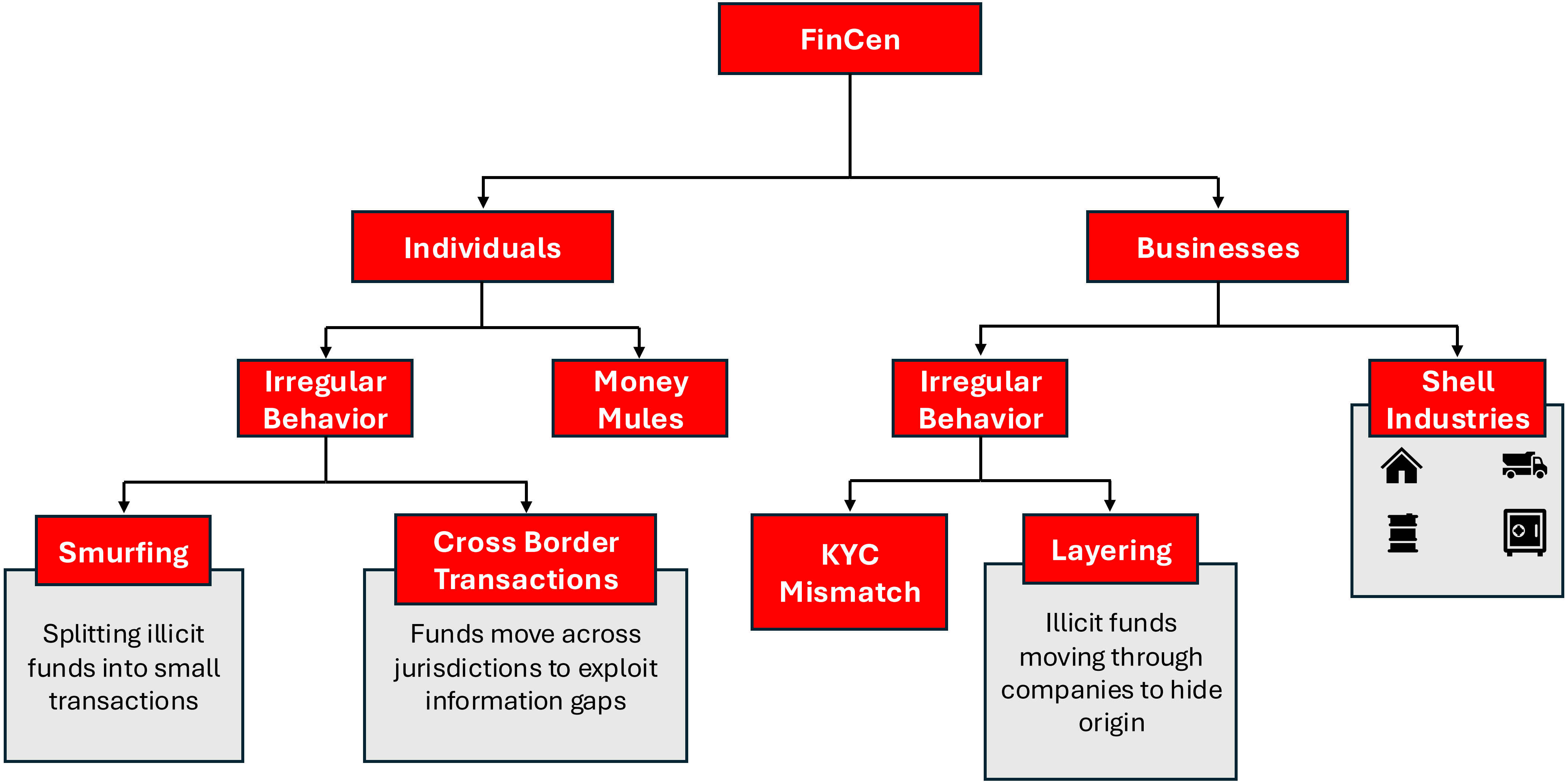
Executive Summary



Task One: Building an AML Knowledge Library



Introduction to AML



Source(s): FinCen & FINTRAC

Features Gathered From Literature Mining

120 Total Features

Individual

10

Features

- Gender
- Age
- Marital status
- Etc.

Business

6

Features

- Industry
- Employee count
- Sales
- Etc.

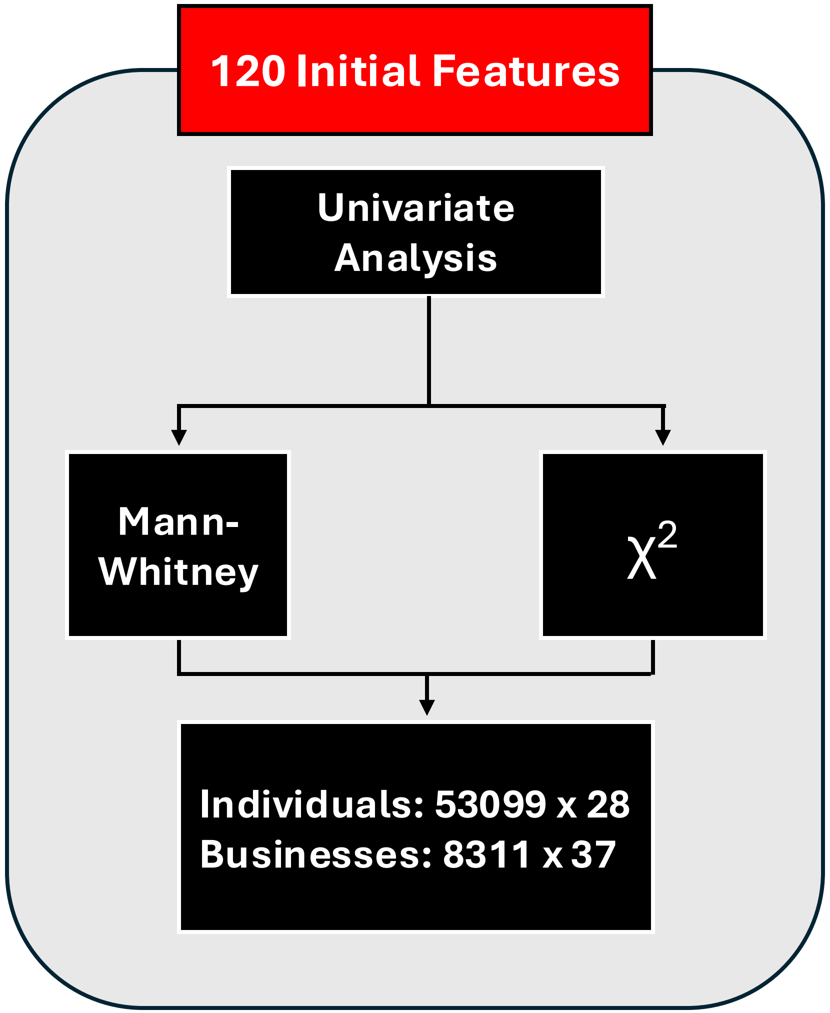
Both

104

Features

- Demographic
- Channel-specific
- Aggregated
- Transaction freq
- Etc.

Refining Features List (Preprocessing)

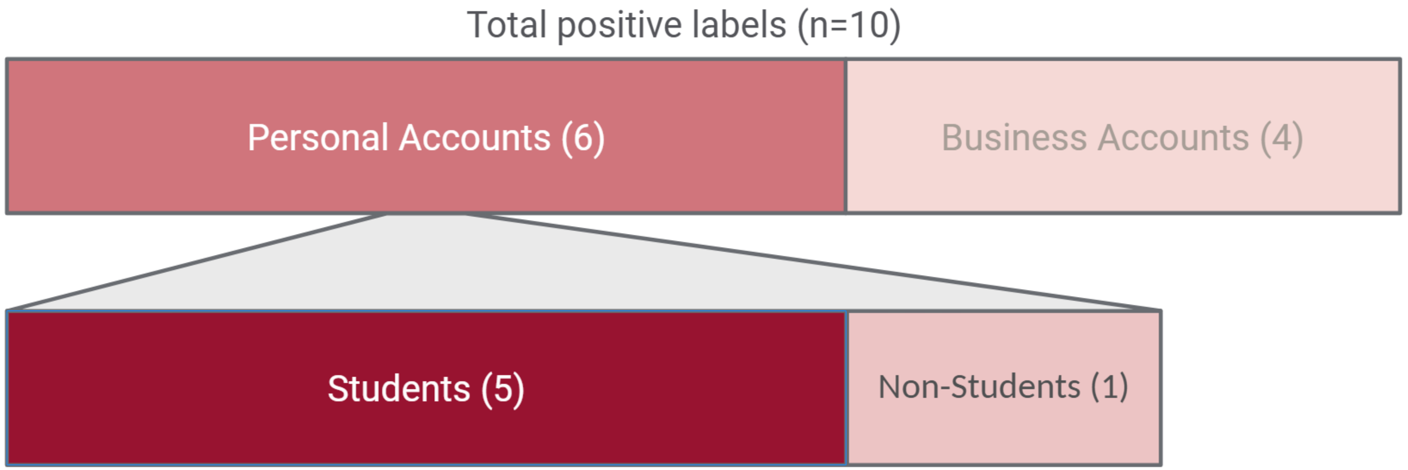


Colinear features were manually removed

Certain feature combinations were used to develop snorkel labelling factors

Task Two: Building a Money Laundering Detection Model

Why Supervised Learning Fails

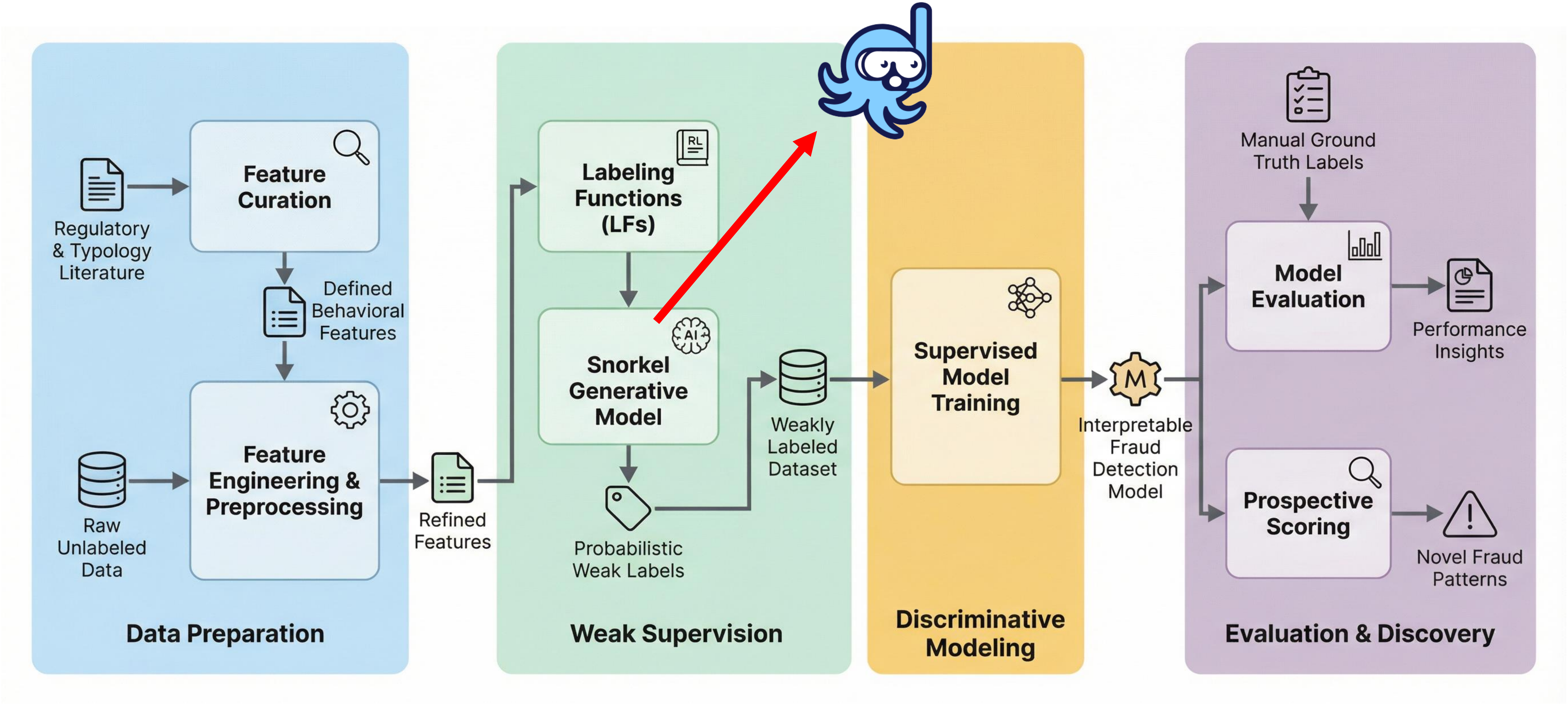


A supervised model would incorrectly learn: Student = High Risk.

Label Scarcity

Fraud ground truth is incomplete, delayed, and noisy.

Introduction to Snorkel AI (Weak Supervision)



A Deeper Dive into Labelling Functions

Probabilistic LF Aggregation

Snorkel treats true labels as unobserved and learns LF accuracy from agreement and conflict patterns, assigning higher probabilistic weight to more reliable signals

LF	Polarity	Coverage	Overlaps	Conflicts
Short hold time	[1]	0.079545	0.079545	0.056818
Low channel diversification	[1]	0.022727	0.022727	0
Transaction same amount	[1]	0.045455	0.045455	0.011364
Infrequent transactions	[0]	0.204545	0.204545	0.034091
Long hold time	[0]	0.204545	0.136364	0
Low debit transfers	[0]	0.204545	0.181818	0.022727
No recent transactions	[0]	0.136364	0.136364	0.022727
Low var transaction	[0]	0.454545	0.295455	0.045455
High sale high debit	[1]	0.011364	0.011364	0.011364
Frequent transaction short hold time	[1]	0.011364	0.011364	0.011364
High var emt	[1]	0.034091	0.011364	0

Developing Separate Models for Individuals and Businesses

Business model

Regulatory & Typology Literature
(FinCEN, FINTRAC, Trade-Based Money Laundering alerts)

↓

Business-Specific Engineered Features

- Industry risk score
- Revenue-to-employee anomalies
- Shell company indicators
- Trade-based transaction patterns
- Beneficial ownership complexity
- Corporate flow-through activity
- Geographic trade exposure

↓

Snorkel Labeling Functions
(Business-specific red flag rules)

↓

Weakly Labeled Business Dataset
(Probabilistic labels generated by Snorkel)

↓

XGBoost Business Risk Model

Output: Business AML Risk Score

Individual model

Regulatory & Typology Literature
(FinCEN, FINTRAC, structuring and personal transaction typologies)

↓

Individual-Specific Engineered Features

- Transaction velocity
- Structured deposits
- Round-amount frequency
- Income vs activity mismatch
- Peer-group anomaly score
- Counterparty diversity
- High-risk jurisdiction exposure

↓

Snorkel Labeling Functions
(Individual-specific red flag rules)

↓

Weakly Labeled Individual Dataset
(Probabilistic labels generated by Snorkel)

↓

XGBoost Individual Risk Model

↓

Individual AML Risk Score

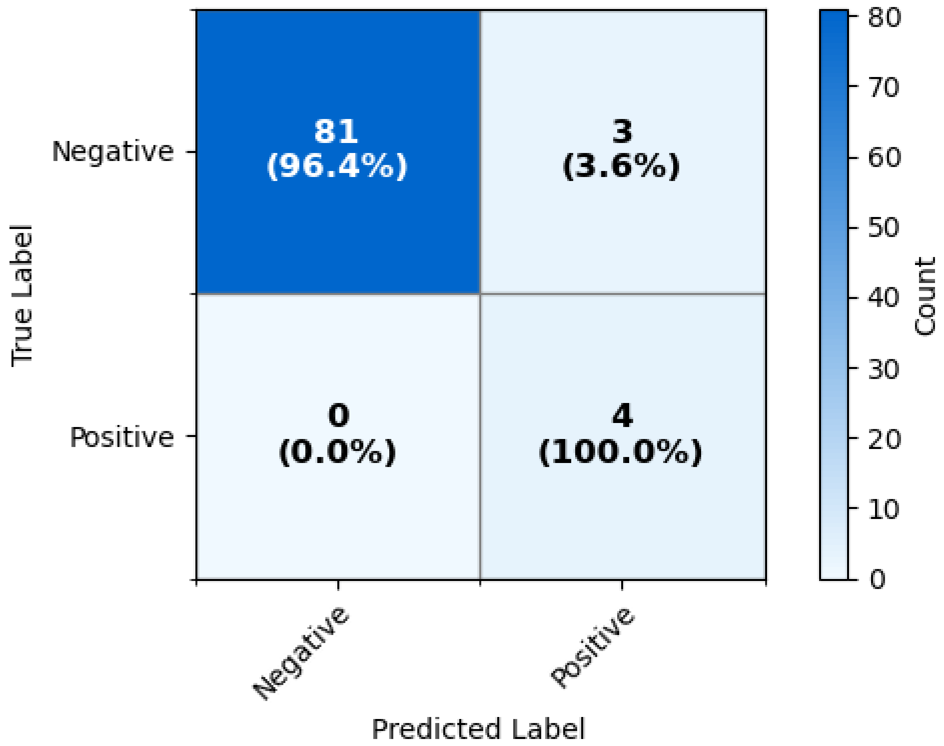
Model Metrics

Business Model Performance

0.57 Precision

1.0 Recall

0.74 MCC

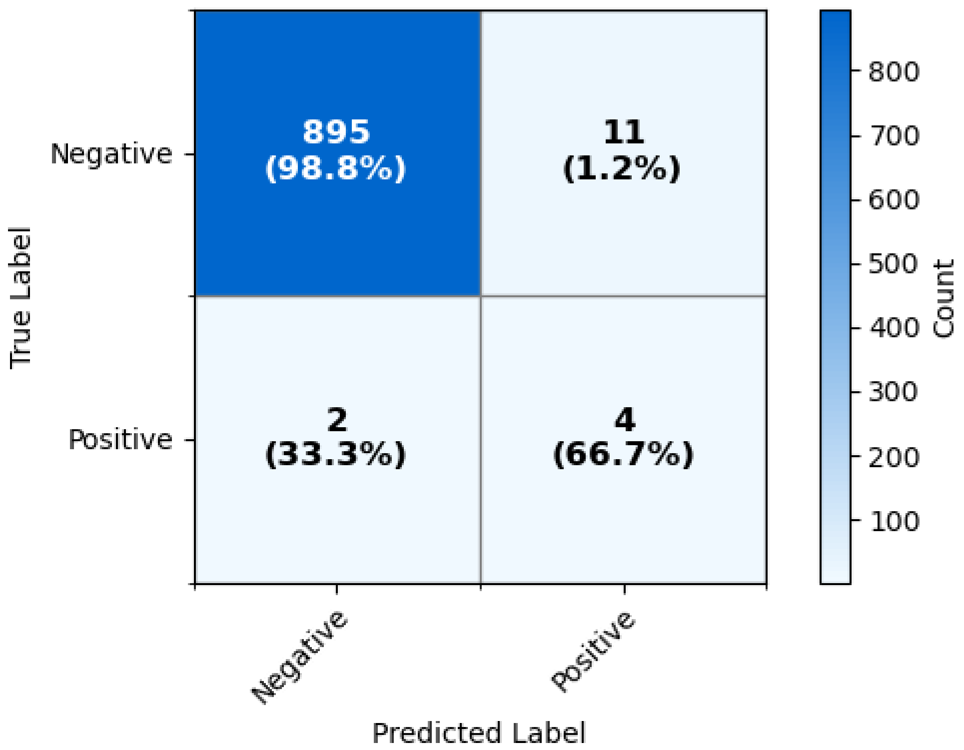


Individual Model Performance

0.27 Precision

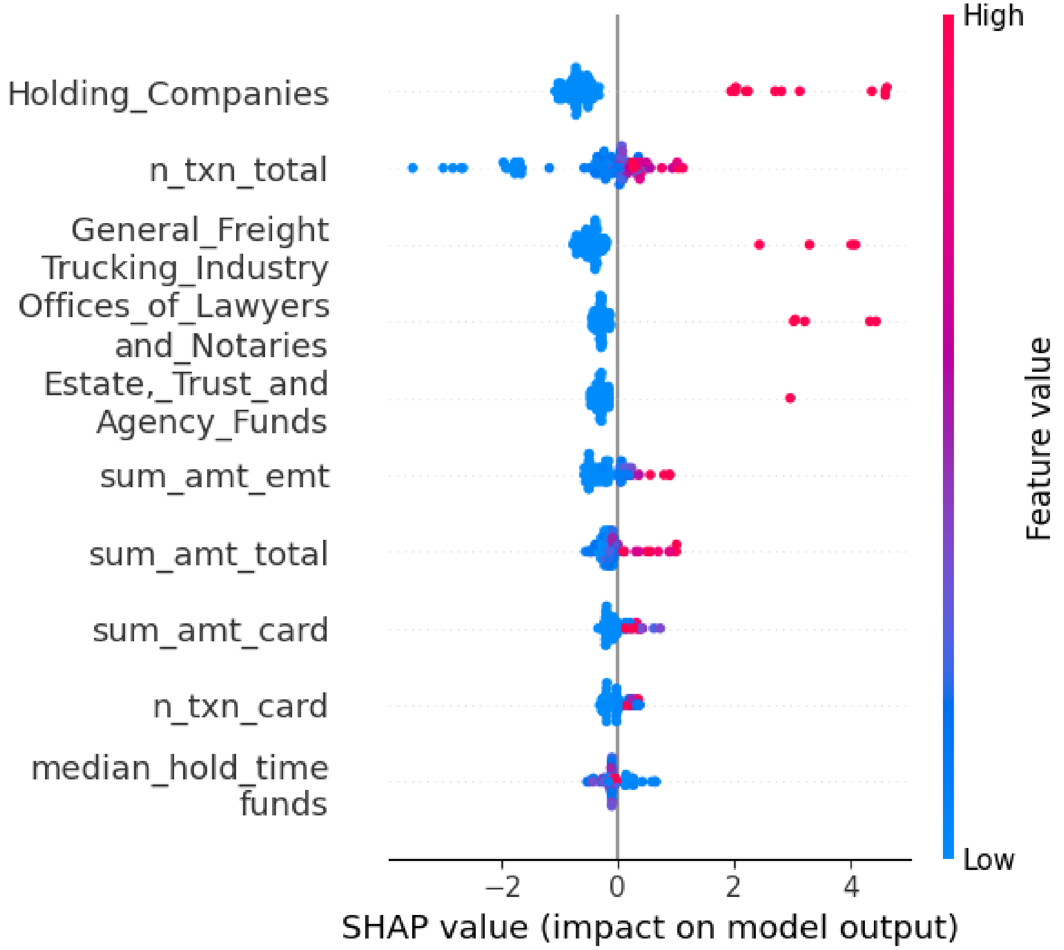
0.67 Recall

0.42 MCC

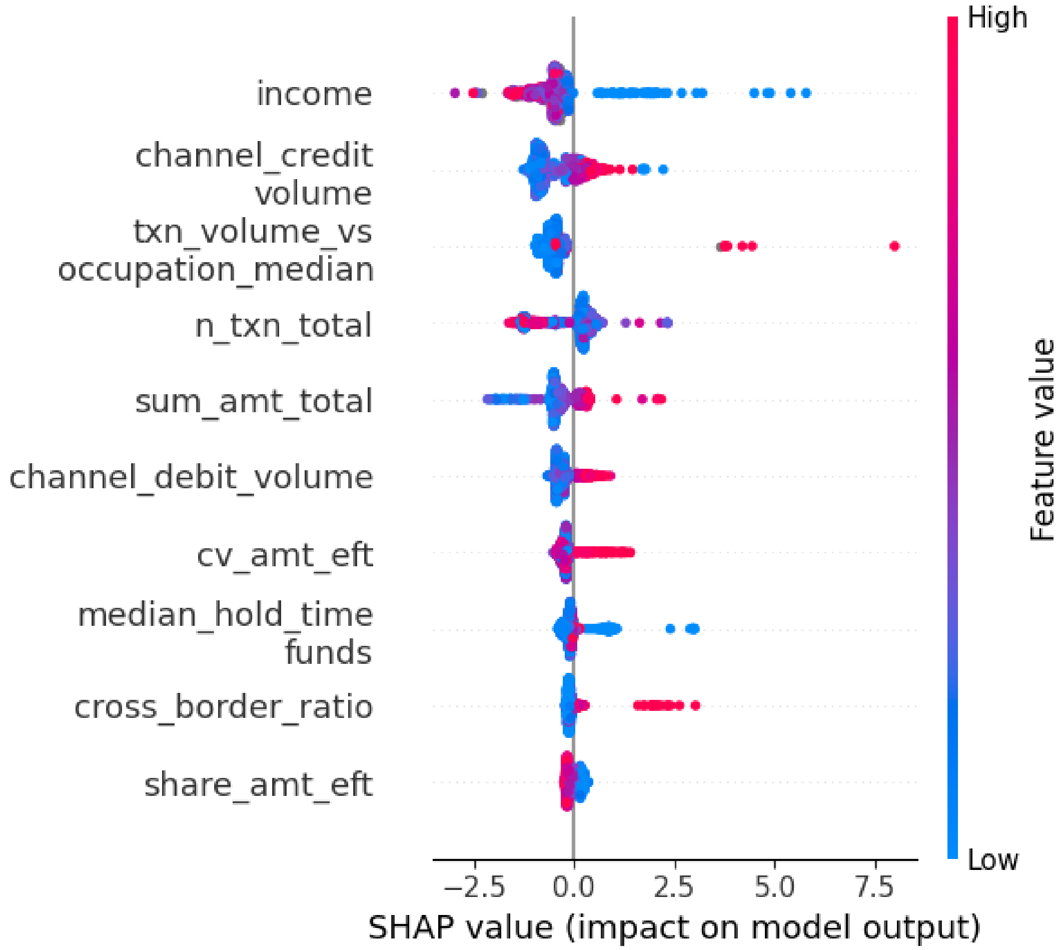


Feature Importance

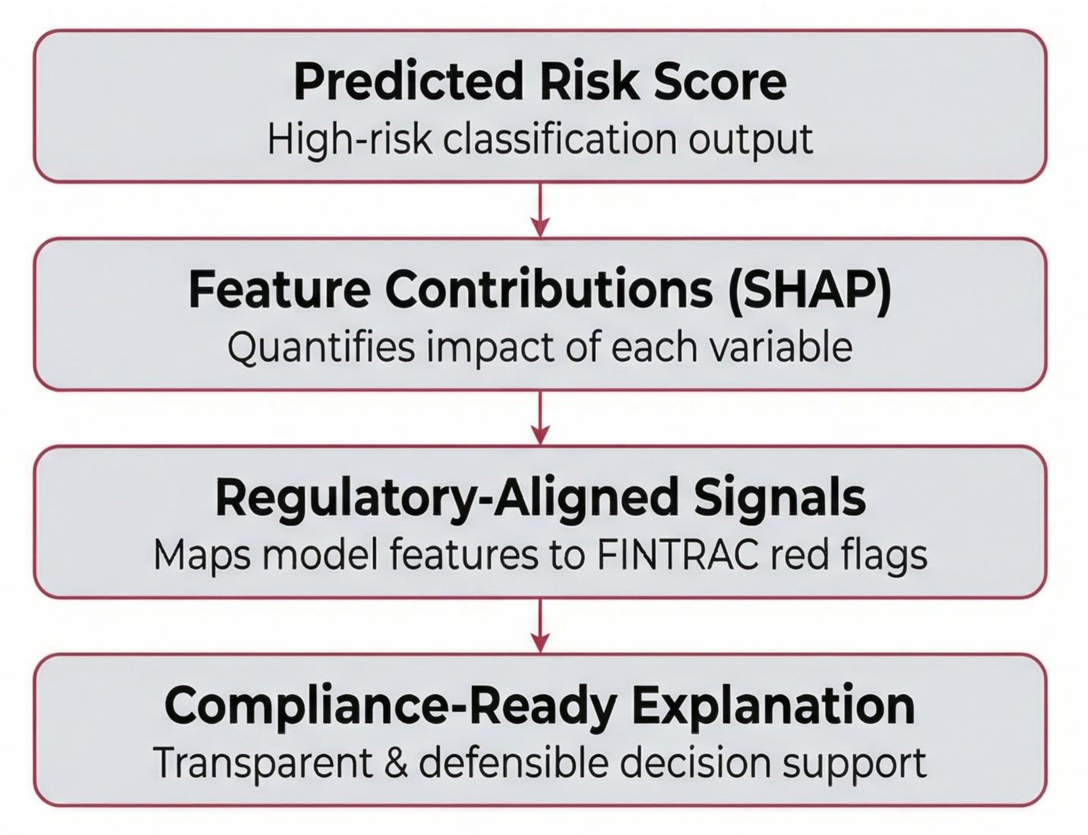
Top Features for Businesses



Top Features for Individuals



Task 3: Explaining Model Results in the Context of Red Flags and Indicators



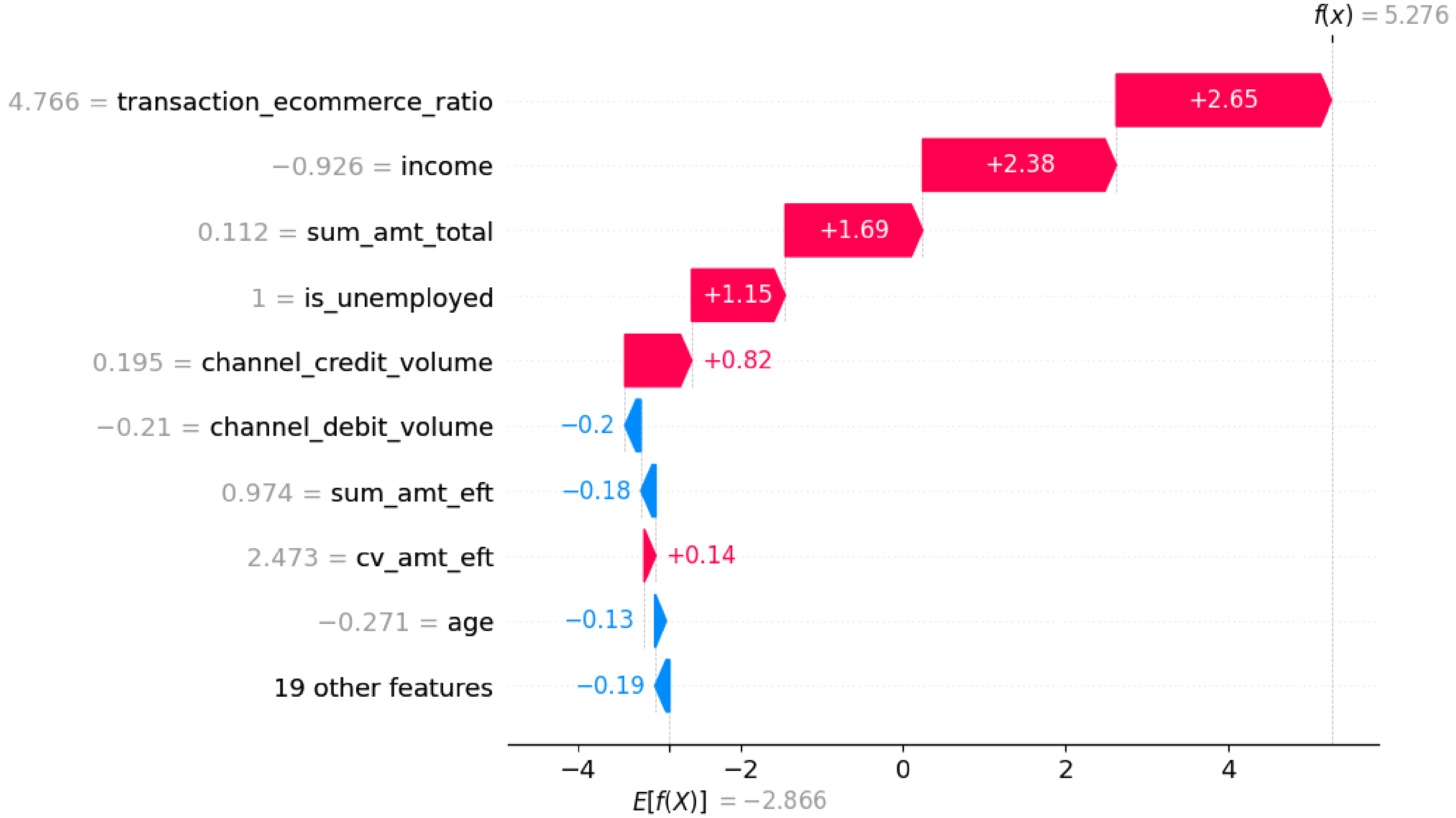
Every risk score is backed by measurable financial behaviour

No black boxes, only observable financial behaviour

Example

Predicted high risk individual

Customer ID: SYNID0103998138
Predicted risk score: 0.99
Explanation: - High transaction to income ratio - Ecommerce occupation mismatch Top features: - Low income ($16) - High sum_amt_total ($45,070) - High transaction_ecommerce_ratio (1.0) - High channel_credit_volume ($38,967) - Satisfied is_unemployed

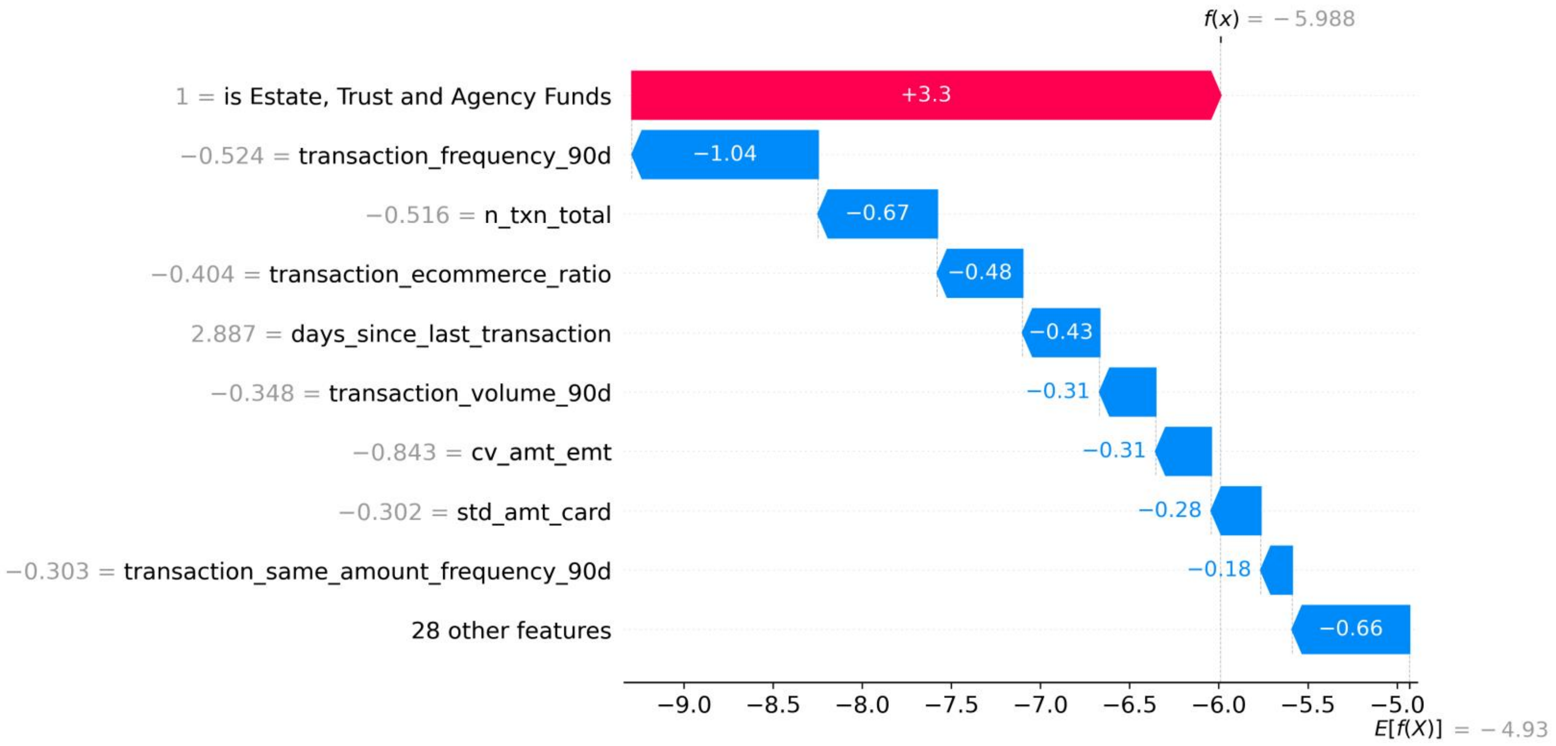


The average fraud probability in the individual dataset is 5.4%

Example

Predicted Low risk business

Customer ID: SYNID0200037978
Predicted risk score: 0.0031
Explanation: - Short hold time Top features: - Low transaction_frequency_90d (3) - Low n_txn_total (4) - High days_since_last transaction (43)



The average fraud probability in the business dataset is 0.7%

Strengths, Weaknesses & Recommendations

Strengths

Process adapts quickly to distribution shifts by updating or adding labelling factors to capture emerging fraud behaviours and adversarial tactics without costly manual relabeling

Weaknesses

Binarizing weak labels can reduce signal richness. However, this loss is mitigated by incorporating high-precision labelling factors that provide strong, reliable supervision and help stabilize model training.

Recommendations

Increase the number of labelling factors, then repeat the experiment using probabilistic outputs instead of binarized labels and compare performance across both approaches.

Thank you!

Team 16 - SignalWeave

Scotiabank BIGDataAIHUB

Institute for Management & Innovation
UNIVERSITY OF TORONTO
MISSISSAUGA