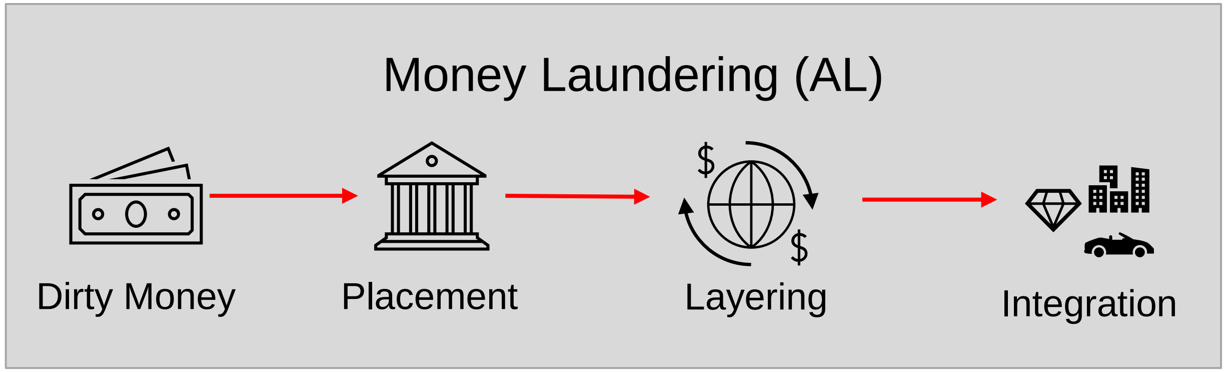
2023 IMI *BIGDataAIHub* Case Competition

Anti-Money Laundering

Team 35 (William Kwok, Juandiego Morzan, Anny Huang)

Agenda

Task 1	Name Screening
Task 2	Supervised Learning
2A	Customer Risk Rating
2B	Bad Actors
Task 3	Graph Analytics



Money Laundering (AL)

Dirty Money → Placement → Layering → Integration

Task 1
Name screening

Task 1: Name Screening

2-step screening solution to identify 50 Bad Actors

Data sources for name screening

1m customers

Scotiabank

430k sanctioned names
- 260k persons
- 170k previous names and alias

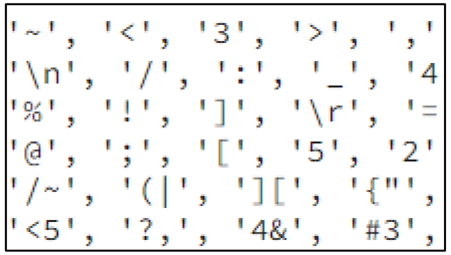
open sanctions

(nested json of 56 datasets)

430bn possible combinations

Need for fuzzy name matching

- Punctuation
- Delimiter (space, hyphen, underscore)
- Extra letter and/or words
- Missing letter and/or words
- Word ordering



```
'~', '<', '3', '>', ','
'\n', '/', ':', '_', '4
'%', '!', ']', '\r', '=
'@', ';', '[', '5', '2'
'/~', '(|', '][', '{"',
'<5', '?,', '4&', '#3',
```

2-step screening solution

1 **Large-scale fuzzy name matching**

3-gram cosine similarity

Sparse matrix multiplication

2 **Validate additional information**

Date of birth

Gender

Politically Exposed Person (PEP)

Task 1: Name Screening

Step 1: Large-scale fuzzy name matching with 3-gram cosine similarity

Sanctioned person = **Young, Marie Mildren** Text processing → **youngmariemildren** for 3-gram extraction

oun mil

you ung rie ild

Possible matches | **Vector space model: 3-gram for flexibility + binary occurrence (1 or 0) for stability** | **Filter >= 0.5**

Variations	Examples	aar	amr	are	ari	arr	ary	dre	eim	emi	eny	gam	gma	gmi	gmm	…	rei	rem	ren	rie	rre	rri	rym	ung	ymi	you	yun	Cosine Similarity
Exact match	**Young, Marie Mildren**				1			1		1			1						1	1				1		1		1.0000
No space	Young MarieMildren				1			1		1			1						1	1				1		1		1.0000
Word order	Marie Mildren Young				1			1		1	1								1	1				1		1		0.8667
Extra letter	Young, Maarrie Mildren	1				1		1		1			1			…			1	1		1		1		1		0.8141
Extra word	Young, Mildren							1						1					1					1		1		0.7348
Abbreviation	Young M Mildren							1							1				1					1		1		0.7006
Phonetic	Yung Mary Mildren						1	1					1						1				1	1	1		1	0.6445
Typo	Young, aMrei Mildren		1					1	1			1					1		1					1		1		0.5333
Wrong person	**Arei mr Remi**			1					1	1						…	1	1			1							0.0913

430bn possible combinations reduced to **5.4mio** with CSR sparse matrix multiplication + top-n result selection

Task 1: Name Screening

Step 2: Validate additional information to identify 50 Bad Actors

Scotiabank

open sanctions

Scotiabank customers	GENDER1	DOB1	OpenSanctions targets	GENDER2	DOB2	Cosine Similarity
Paul Franklin Watson	Male	1950-12-02	PAUL FRANKLIN WATSON	Male	1950-12-02	1.0000
Alexey Alexeyevich Gromov	Male	1960-05-31	Alexey Alexeyevich GROMOV	Male	1960-05-31	1.0000
Emilie Samra Konig	Female	1984-12-09	Emilie Samra Konig	Female	1984-12-09	1.0000
Tetiana Viktorivna Pereverzeva	Female	1964-06-20	Tetiana Viktorivna Pereverzeva	Female	1964-06-20	1.0000
Basova, Lidiya Oleksandrivna	Female	1972-01-01	Lidiya Oleksandrivna Basova	Female	1972	0.9130
Bezrukov, Sergey Vitalyevich	Male	1973-10-18	Sergey Vitalyevich BEZRUKOV	Male	1973-10-18	0.9130
Zheynova, Marina Nikolaevna	Female	1985-02-15	Marina Nikolaevna ZHEYNOVA	Female	1985-02-15	0.9091
Rakhim Azizboevich Azimov	Male	1964-08-16	AZIMOV Rakhim Azizboevich	Male	1964-08-16	0.9000
Oleksin, Alexei Ivanovich	Male	1966-10-29	OLEKSIN Aleksei Ivanovich	Male	1966-10-29	0.8721
Herlinto Chamorro Acosta	Male	1956-01-10	ELIECER HERLINTO CHAMORRO ACOSTA	Male	1956-01-10	0.8607
Jose Benito Cabrera Cuevas	Male	1963-07-06	Jose Benito Cabrera	Male	1963-07-06	0.8452
Poklonskaya, Natalija Vladimirovna	Female	1980-03-18	Natalia Vladimirovna POKLONSKAYA	Female	1980-03-18	0.8422
O Jong Gil	Male	1962-08-30	Jong Gil O	Male	1962-08-30	0.8333
Hlaing, Min Aung	Male	1956-07-03	Min Aung Hlaing	Male	1956-07-03	0.8182

50 Bad Actors

Same **gender**

DOB <= 2 years

Same **PEP** status

High **cosine similarity**

High **risk rating**

Other considerations include

DOB difference, country, target / non-target on sanction list, length of name in database

Name screening practices

Reference

MAS Monetary Authority of Singapore

Reference: Monetary Authority of Singapore Strengthening AML / CFT Name Screening Practices Information Paper April 2022

Task 2A
Risk rating model

Task 2A: Supervised Learning of Customer Risk Rating

Using KYC and transaction statistics to assign each customer a risk rating

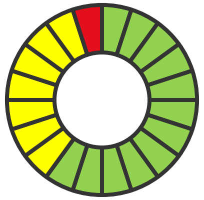
FINTRAC Indicators of a high-risk customer include:

- Anonymity → Multiple transactions below the reporting threshold amount
- Speed over cost-effectiveness → High volume of wire transfers instead of one single large transfer

Type of data	Provided features	Created features
Customers (KYC)	Name, Customer ID Gender PEP Occupation risk Birth date Onboarding date Country of residence Country of income	Time since onboarding Age

LTM transactions	Type = CASH or WIRE?
	Direction = IN or OUT?
	Sum of transaction amount
	Count of transactions
		Avg of transaction amount
		Net balance in LTM
		Ratio of CASH vs WIRES
		Ratio of IN vs OUT

Target variable = Risk Rating



60% Low
35% Medium
5% High

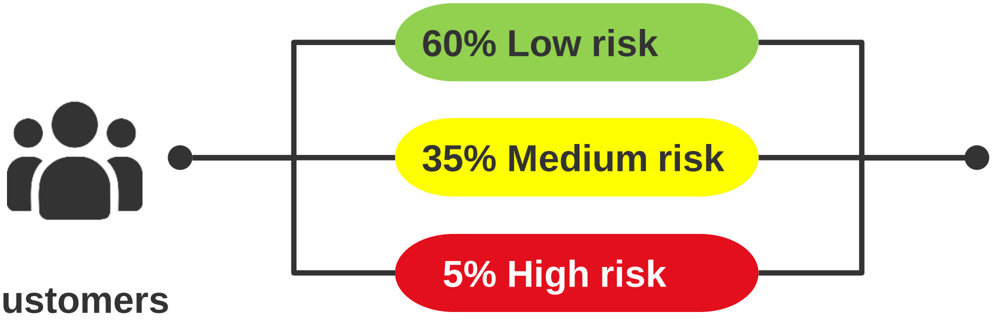
Train / validation / test set

Train / test split = 80% / 20%
5-fold cross validation
Stratify on Risk Rating
Shuffle = True

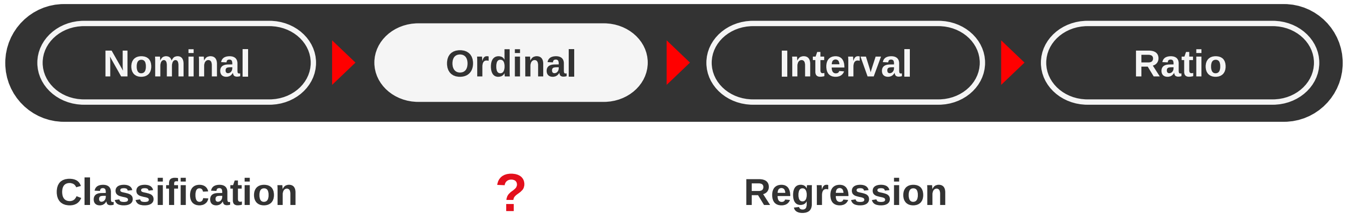
Reference: FINTRAC Money laundering and terrorist financing indicators—Financial entities

Task 2A: Supervised Learning of Customer Risk Rating

Evaluation metric for ordinal classification to assign customers into 3 risk buckets



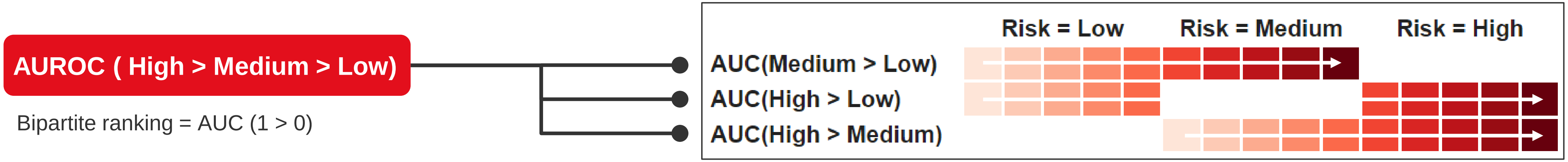
4 Levels of Measurement



Extension of AUROC from bipartite ranking to multipartite ranking (Furnkranz, Hullermeier and Vanderlooy, 2009)

Multipartite ranking

OvO decomposition into **3** bipartite ranking problems



Bipartite ranking = AUC (1 > 0)

```
def multipartite_AUC(y_true, y_score, average = 'macro'):
```

Macro average to address class imbalance

Reference: Binary Decomposition Methods for Multipartite Ranking (J. Furnkranz, E. Hullermeier and S. Vanderlooy, 2009)

Task 2A: Supervised Learning of Customer Risk Rating

Transparent modelling alternative with binary classifiers instead of multiclass classification

Available ordinal decompositions

Option 1: Ordered Partitions

	Low	Medium	High
Split 1	0	1	1
Split 2	0	0	1

OR

Option 2: One Vs Followers

	Low	Medium	High
Split 1	0	1	1
Split 2	■	0	1

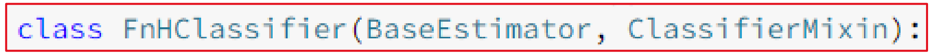
OR

Option 3: One Vs Previous

	Low	Medium	High
Split 1	0	0	1
Split 2	0	1	■

Frank and Hall (2001)

2 binary classifiers **combine probabilities**

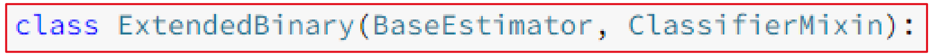


```
class FnHClassifier(BaseEstimator, ClassifierMixin):
```

OR

Data Replication Method (2007)

1 binary classifier **2x augmented data**



```
class ExtendedBinary(BaseEstimator, ClassifierMixin):
```

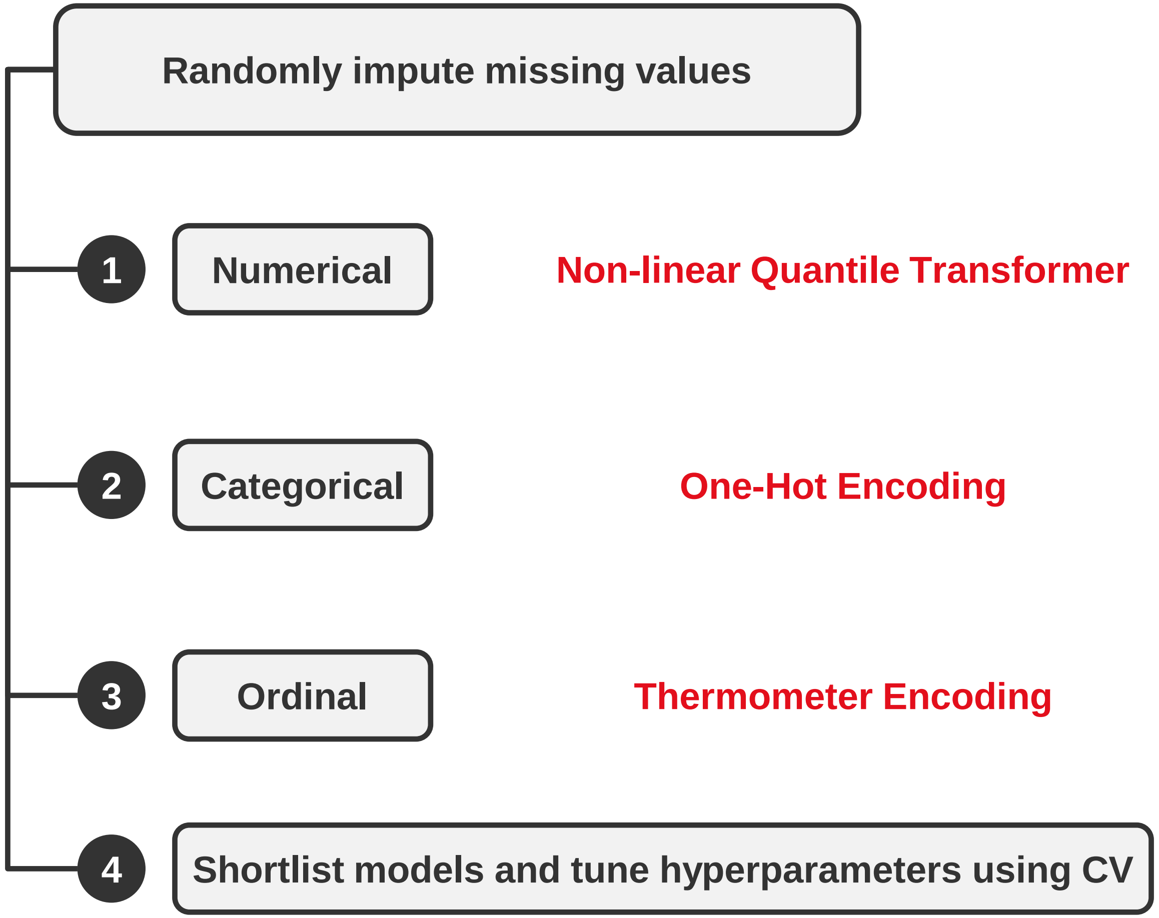
Logistic Regression	Multipartite AUROC score on 5-fold CV	
Frank and Hall OneVsFollowers	**0.9536** Mean	**0.0005** Std
Multiclass (baseline)	**0.9212** Mean	**0.0016** Std

Benefits 1 Linear classifiers 2 Flexibility in decomposition

Reference: A Simple Approach to Ordinal Classification (E. Frank and M. Hall, 2001), Learning to Classify Ordinal Data: The Data Replication Method (J. Cardoso and J. Pinto da Costa, 2007)

Task 2A: Supervised Learning of Customer Risk Rating

Data transformation pipeline and modelling



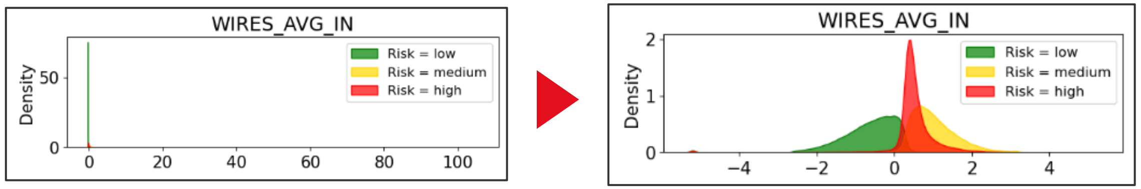
Randomly impute missing values

1 Numerical — Non-linear Quantile Transformer

2 Categorical — One-Hot Encoding

3 Ordinal — Thermometer Encoding

4 Shortlist models and tune hyperparameters using CV



Gender

Female **Male**

1	0
0	1

Drop 1

Occupation risk

Low **Med** **High**

1	0	0
1	1	0
1	1	1

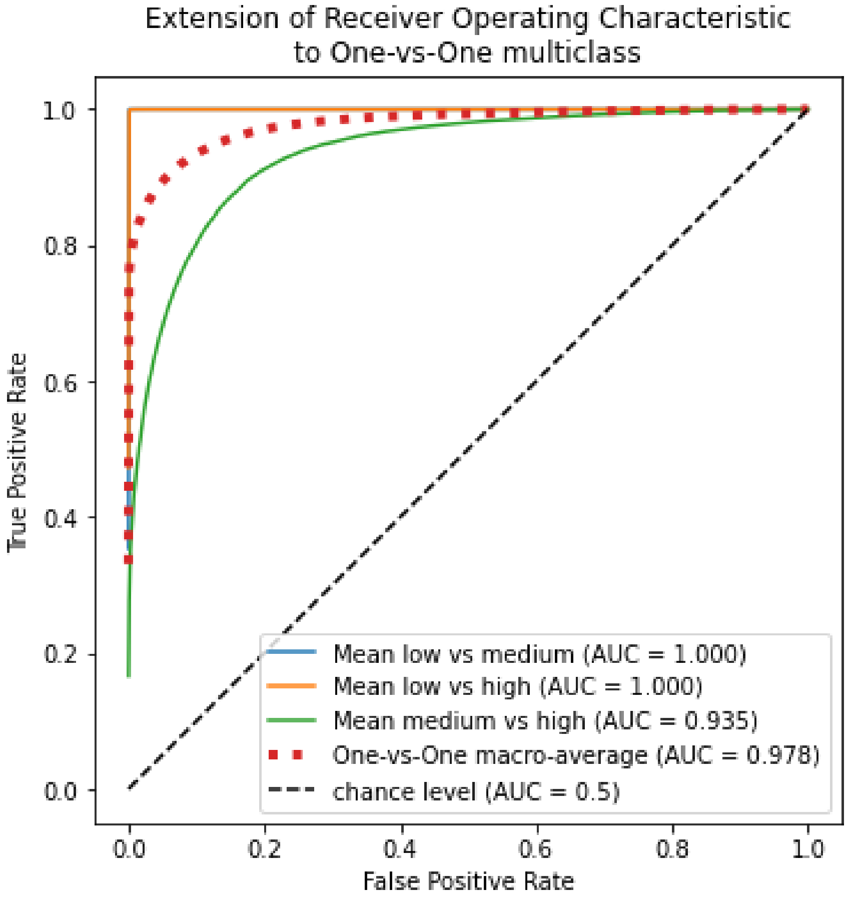
Drop 1

Frank and Hall (LightGBM) + One Vs Followers

Reference: Thermometer Encoding: Evaluating the Impact of Categorical Data Encoding and Scaling on Neural Network Classification Performance (E. Norris, S. Vahid and C. Hand, 2012)

Task 2A: Supervised Learning of Customer Risk Rating

Model performance evaluation on test set: multipartite AUROC



Low vs Medium

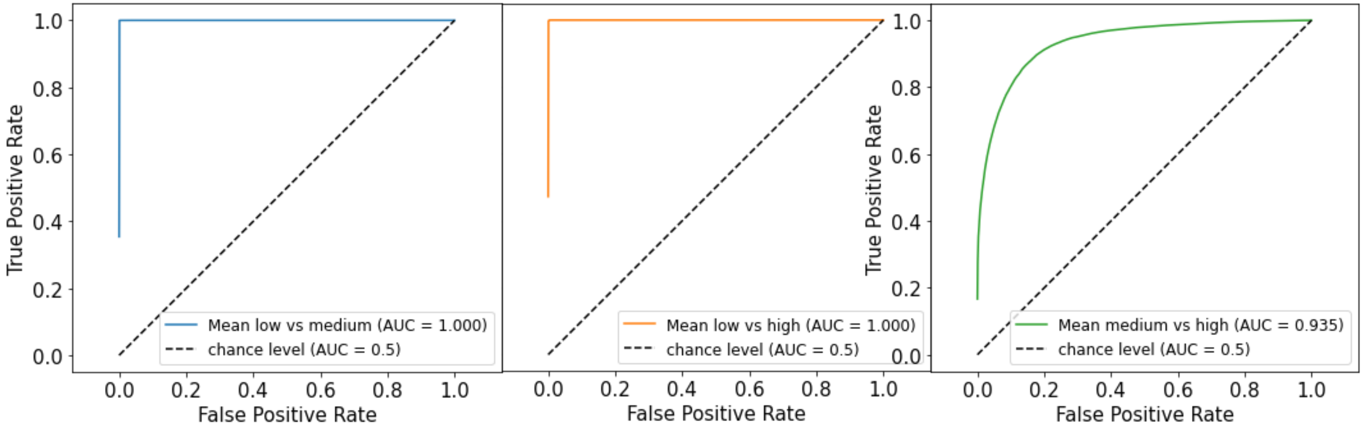
mean AUC: 1.0

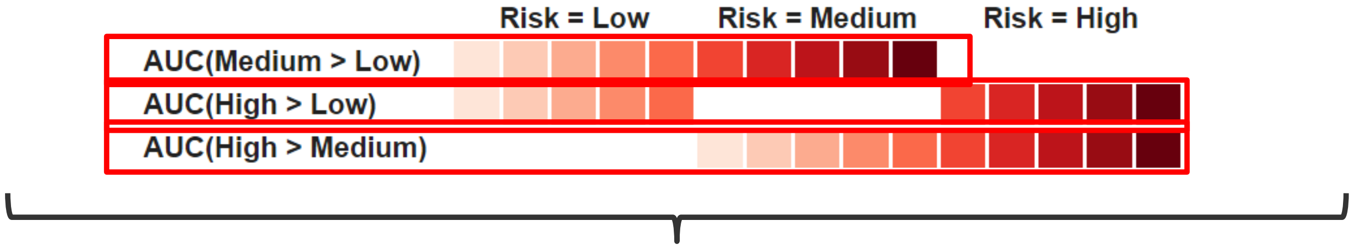
Low vs High

mean AUC: 1.0

Medium vs High

mean AUC: 0.935





Risk = Low Risk = Medium Risk = High

AUC(Medium > Low)

AUC(High > Low)

AUC(High > Medium)

Macro-averaged AUC: 0.978

Task 2A: Supervised Learning of Customer Risk Rating

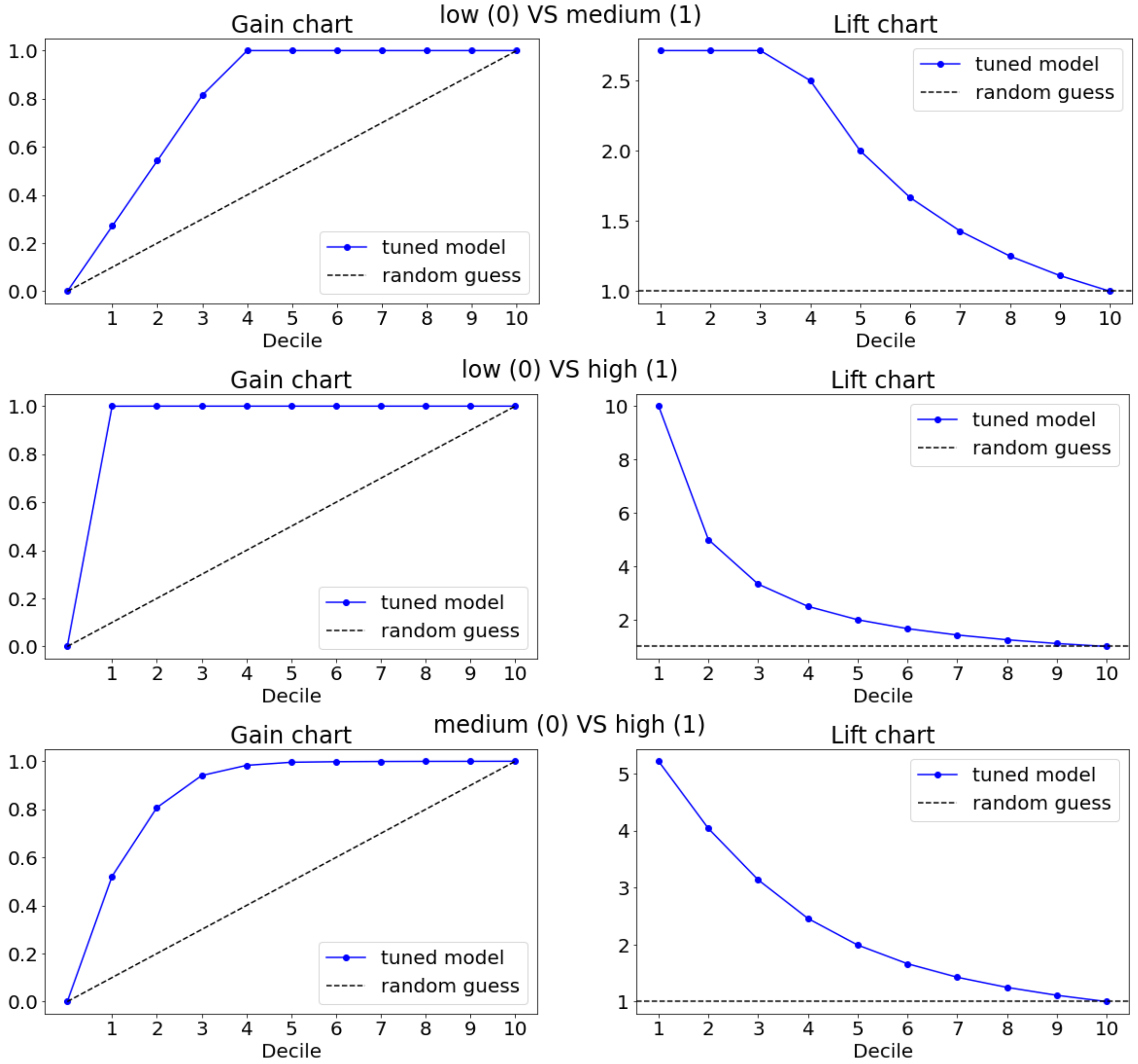
Model performance: gain and lift on test data

Gain @ 1st Decile

Low vs Medium:	27%
Low vs High:	100%
Medium vs High:	50%

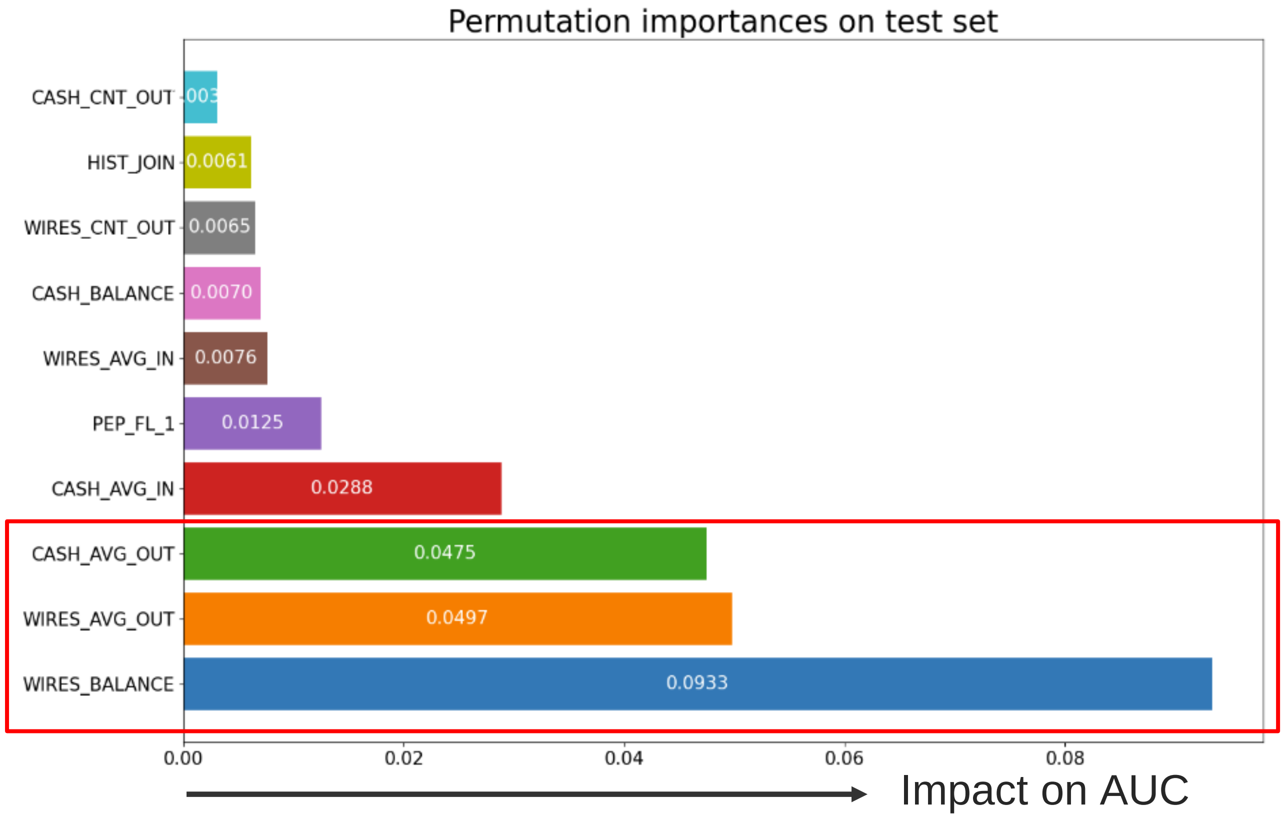
Lift @ 1st Decile

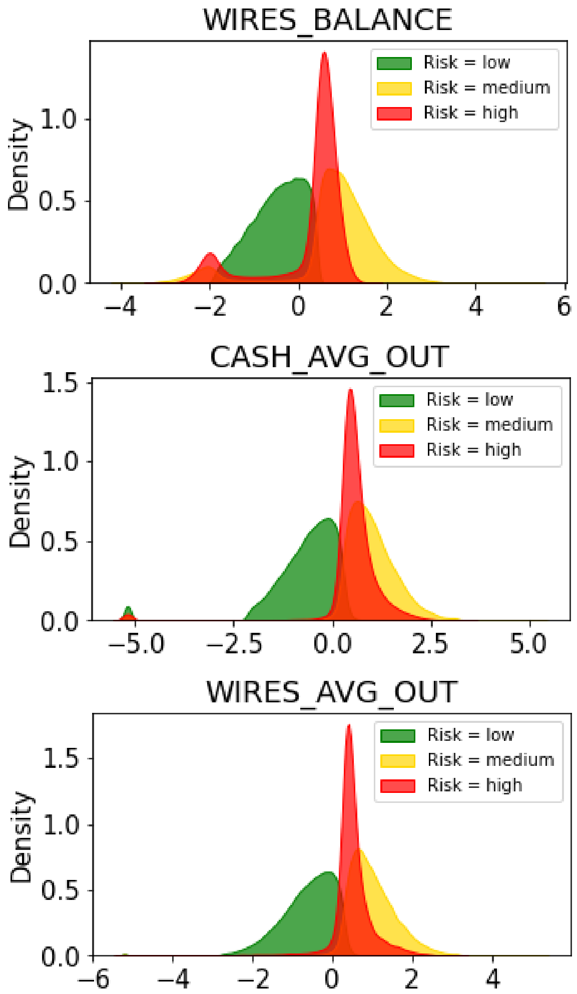
Low vs Medium:	2.7x (max possible lift)
Low vs High:	10x
Medium vs High:	5x



Task 2A: Supervised Learning of Customer Risk Rating

Model performance: analysis insights on test data





* These are scaled values

FINTRAC ML Indicators

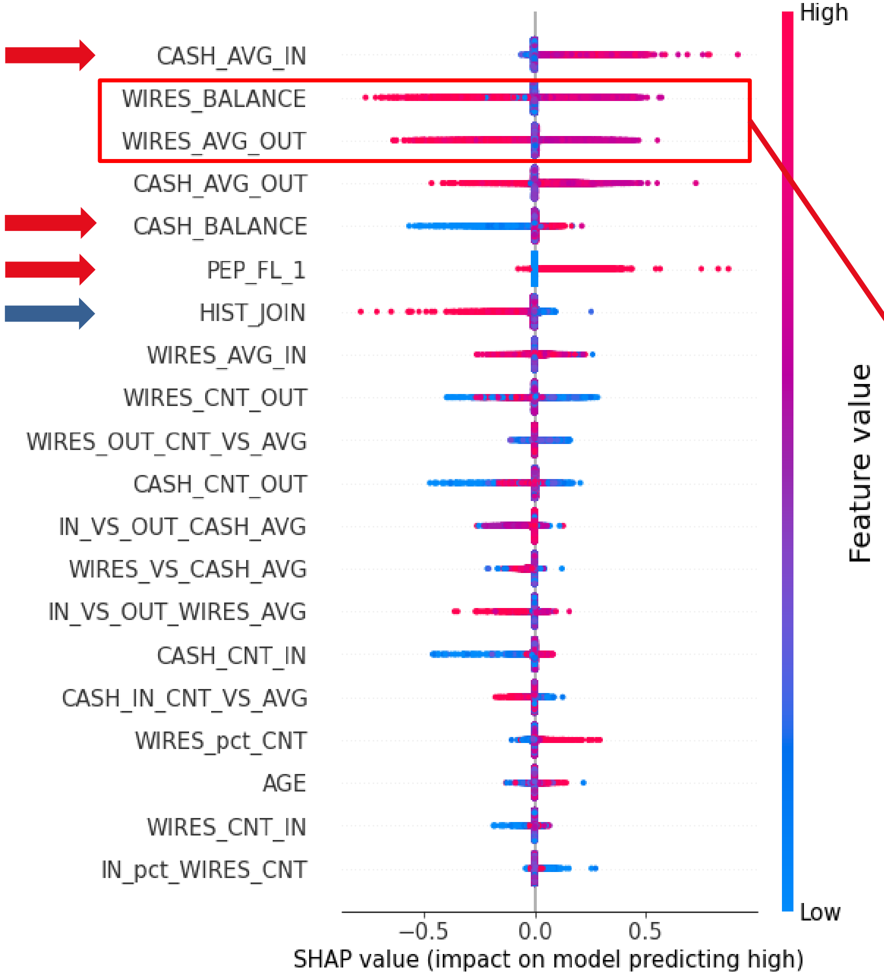
“…transfers on an in and out basis…”

“…structuring amounts to avoid client identification or reporting thresholds…”

Reference: FINTRAC Money laundering and terrorist financing indicators—Financial entities

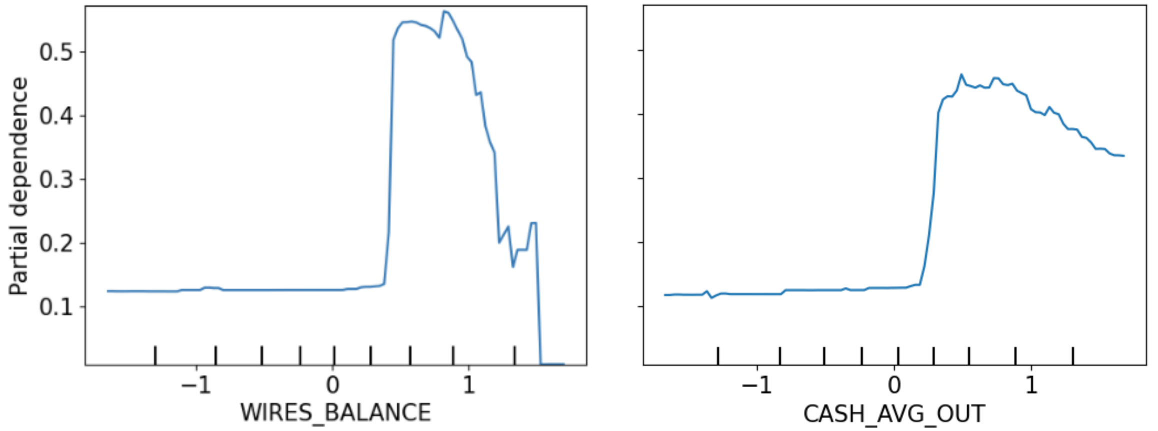
Task 2A: Supervised Learning of Customer Risk Rating

How does the model predicts high risk customers?



Explanation Model Using SHAP

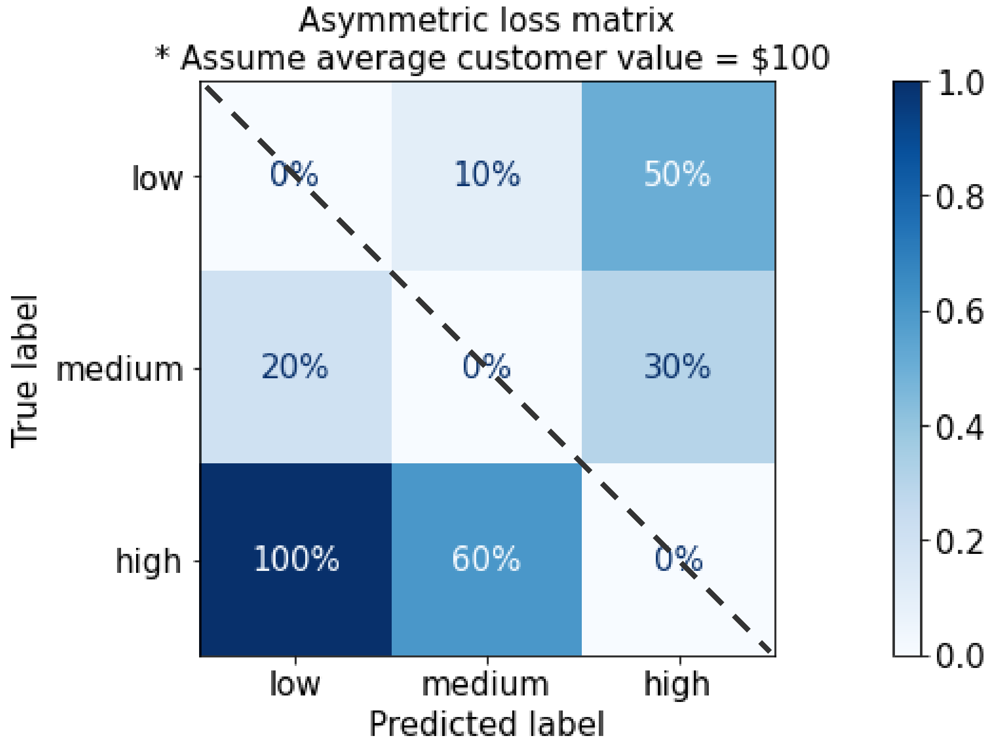
- + Average Cash Deposit
- + PEP
- + Cash balance
- - Time with Bank
- …



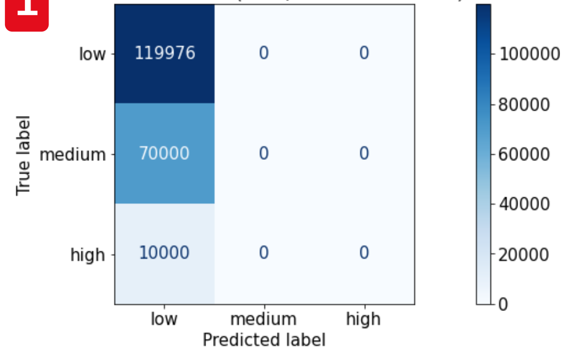
** These are scaled values*

Task 2A: Supervised Learning of Customer Risk Rating

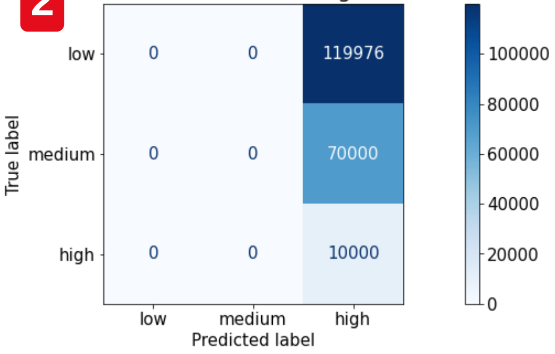
Prescriptive Analytics: applying cost-sensitive structure to improve financial inclusion



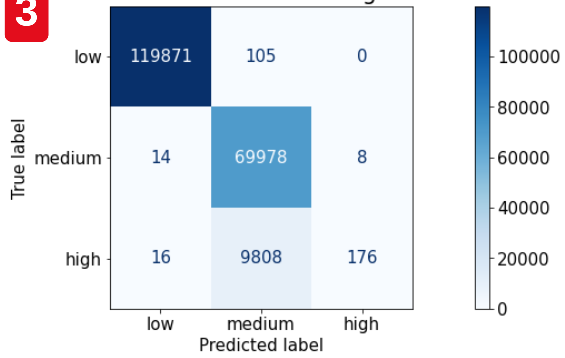
1 Baseline Naive (199,976 customers)



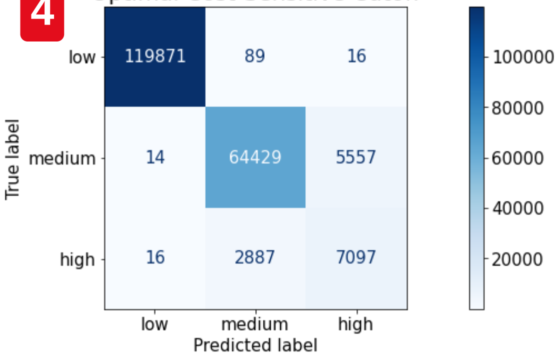
2 Maximum Recall for High Risk



3 Maximum Precision for High Risk



4 Optimal Cost-Sensitive Cutoff



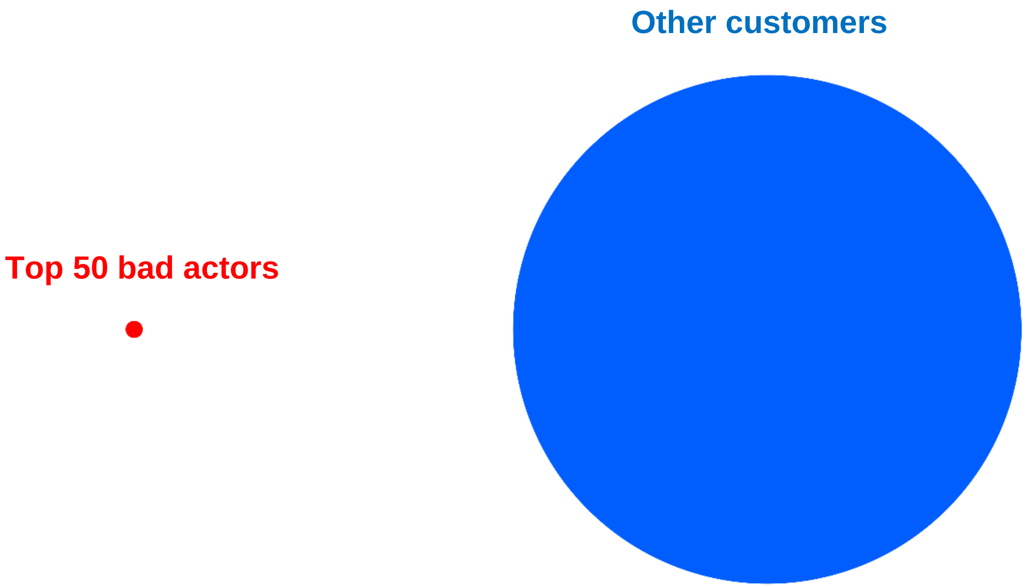
	Scenario	Total Misclassification Cost	Cost Per Misclassified Customer	% of Baseline
1	Naïve Classification	$ 24,000	$ 30	50%
2	Maximum Recall	$ 70,988	$ 37	63%
3	Maximum Precision	$ 5,917	$ 59 (baseline)	100%
4	Optimized Cutoff	$ 3,435	$ 22	37%

Task 2B
50 bad actors

Task 2B: Supervised Learning of Bad Actors

Binary Classification Approach

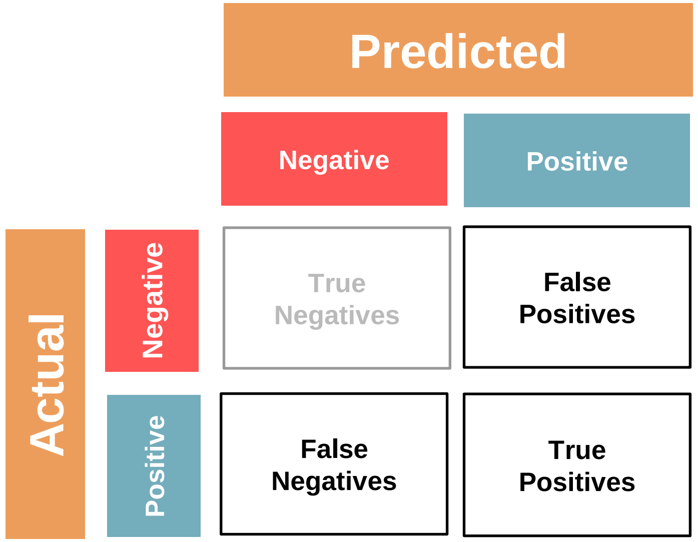
Highly imbalanced dataset



Top 50 bad actors represent just **0.005%** of all customers. (*)

(*) Balanced class weights during training to deal with class imbalance

Average precision as performance metric



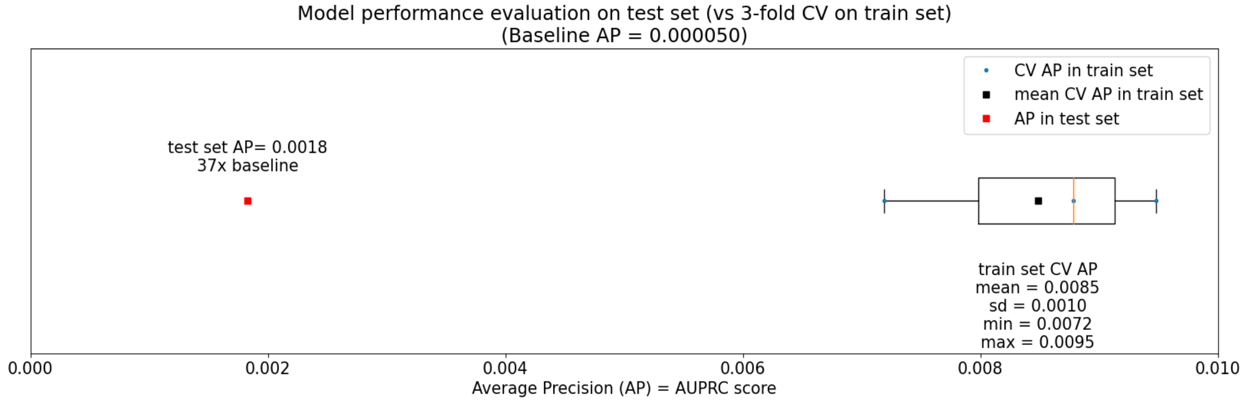
		Predicted	
		Negative	Positive
Actual	Negative	True Negatives	**False Positives**
	Positive	**False Negatives**	**True Positives**

- Measures area under **Precision-Recall curve**
- Useful when the **positive class** is **rare**
- **Emphasizes** high **TPR** in **top-ranked positive** samples
- **Less sensitive** to class **imbalance**

Task 2B: Supervised Learning of Bad Actors

Performance Evaluation and Insights

Low AP = 0.0018 (37x baseline)

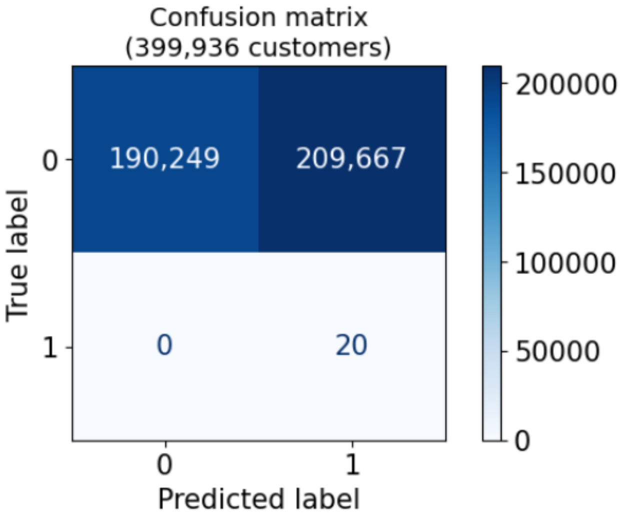


Classification Tradeoff

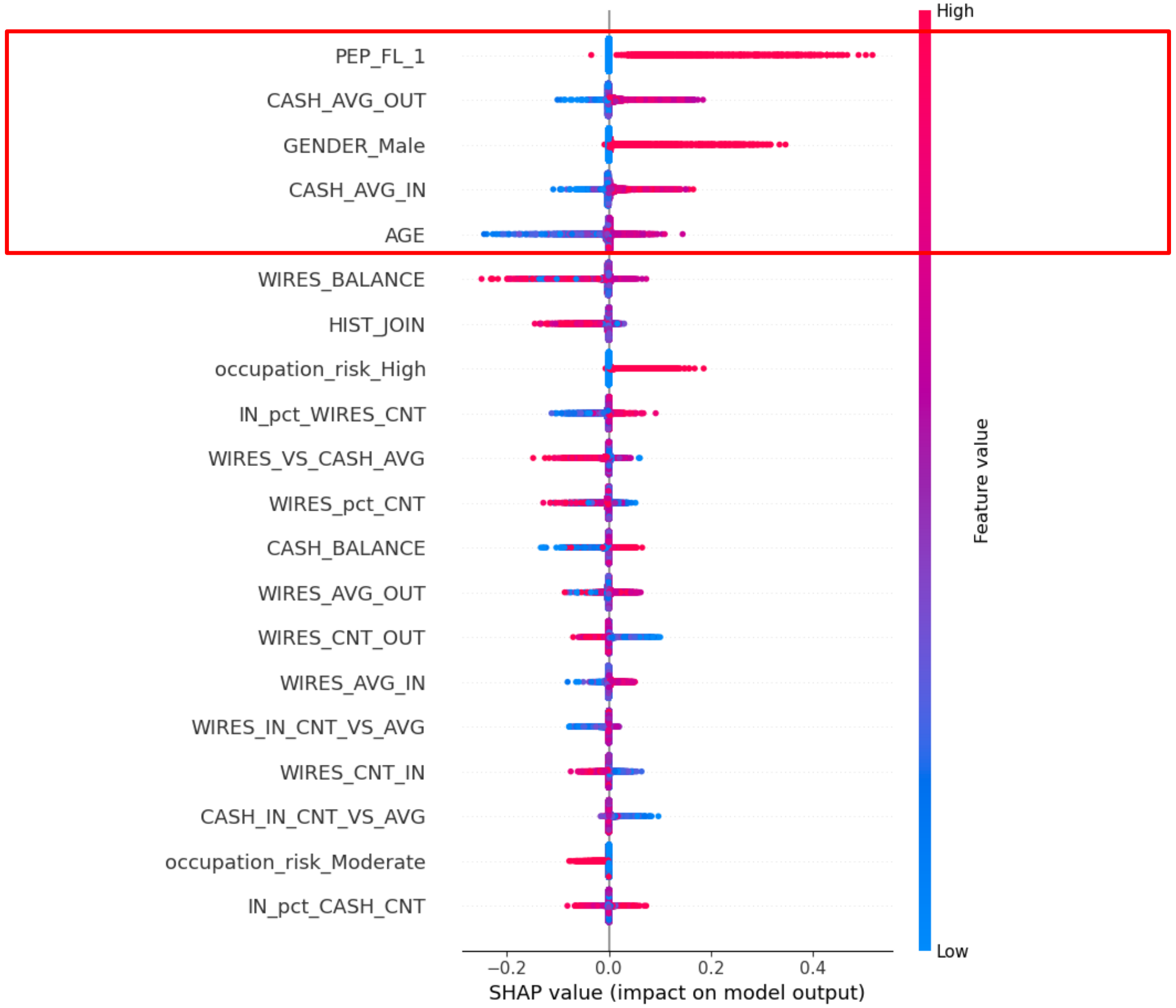
Prob. Threshold = 0.0044

100% **recall**

52% **FPR**



Important features



Task 3
Graph data

Task 3 Graph Analytics

Customer connections: feature engineering with self-supervised learning to enhance risk models

Aggregated Features

OR

Embeddings

Manual feature engineering

neighbour transactions statistics
(max, min, std dev, correlation coefficients)

Automated feature extraction

node2vec

One-hop neighbourhood

one hop forward/backward

Flexible walk

Breadth and depth search strategies

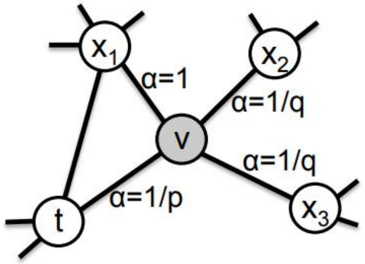
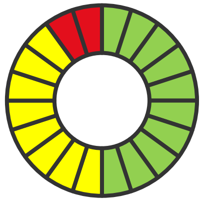


Figure 2: Illustration of the random walk procedure in *node2vec*. The walk just transitioned from t to v and is now evaluating its next step out of node v. Edge labels indicate search biases α.

48% Low
40% Medium
12% High



Customers Connections

Provided features

- Customer ID
- EMT (over 12 months)

Created features

- Node2vec embeddings

Network Statistics

361k customers (*)

466k directed payments

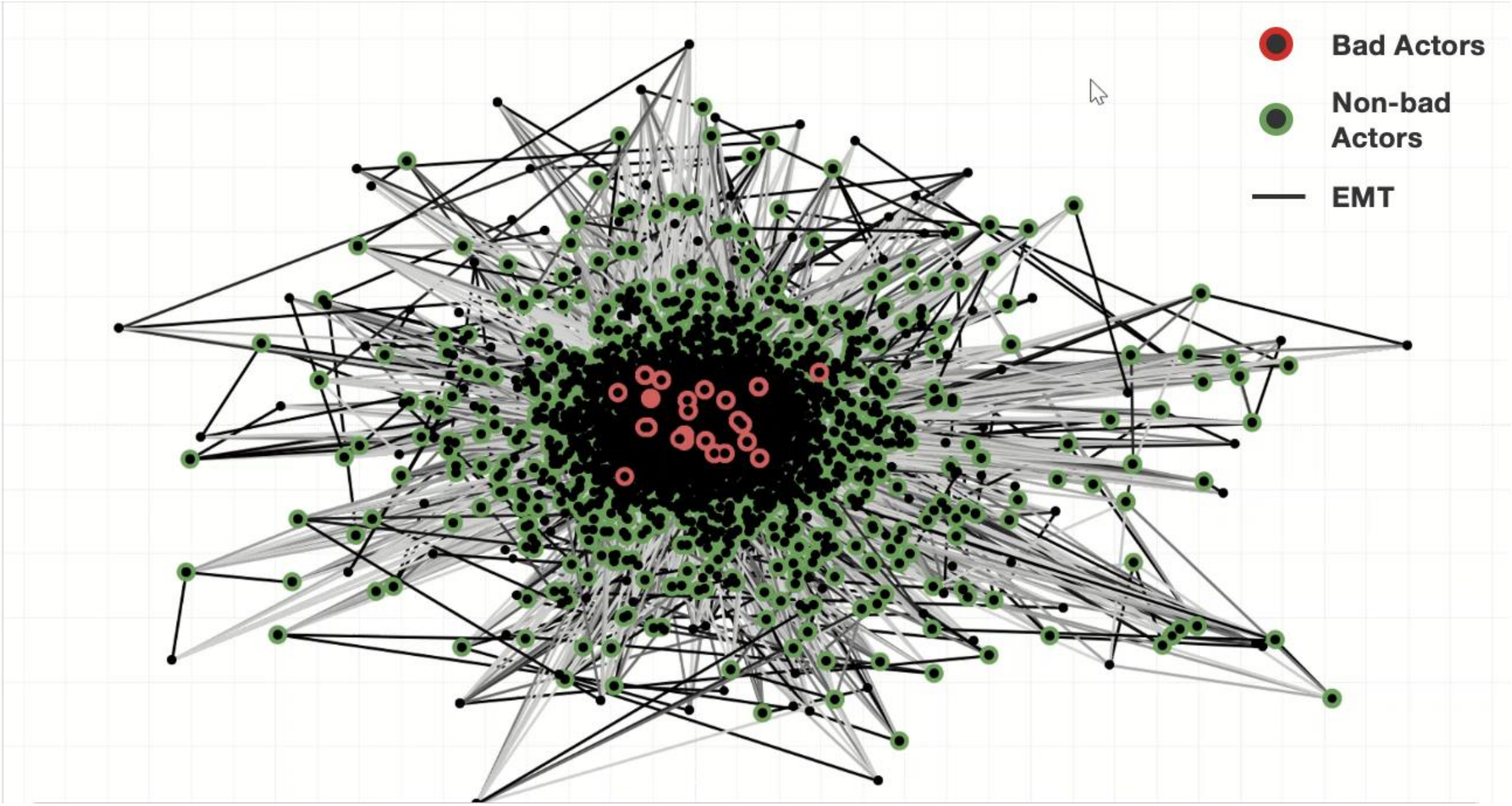
Edge weights as probability

Reference: node2vec: Scalable Feature Learning for Networks: http://arxiv.org/abs/1607.00653

(*) Followed a random imputation within class for customers not present in the network.

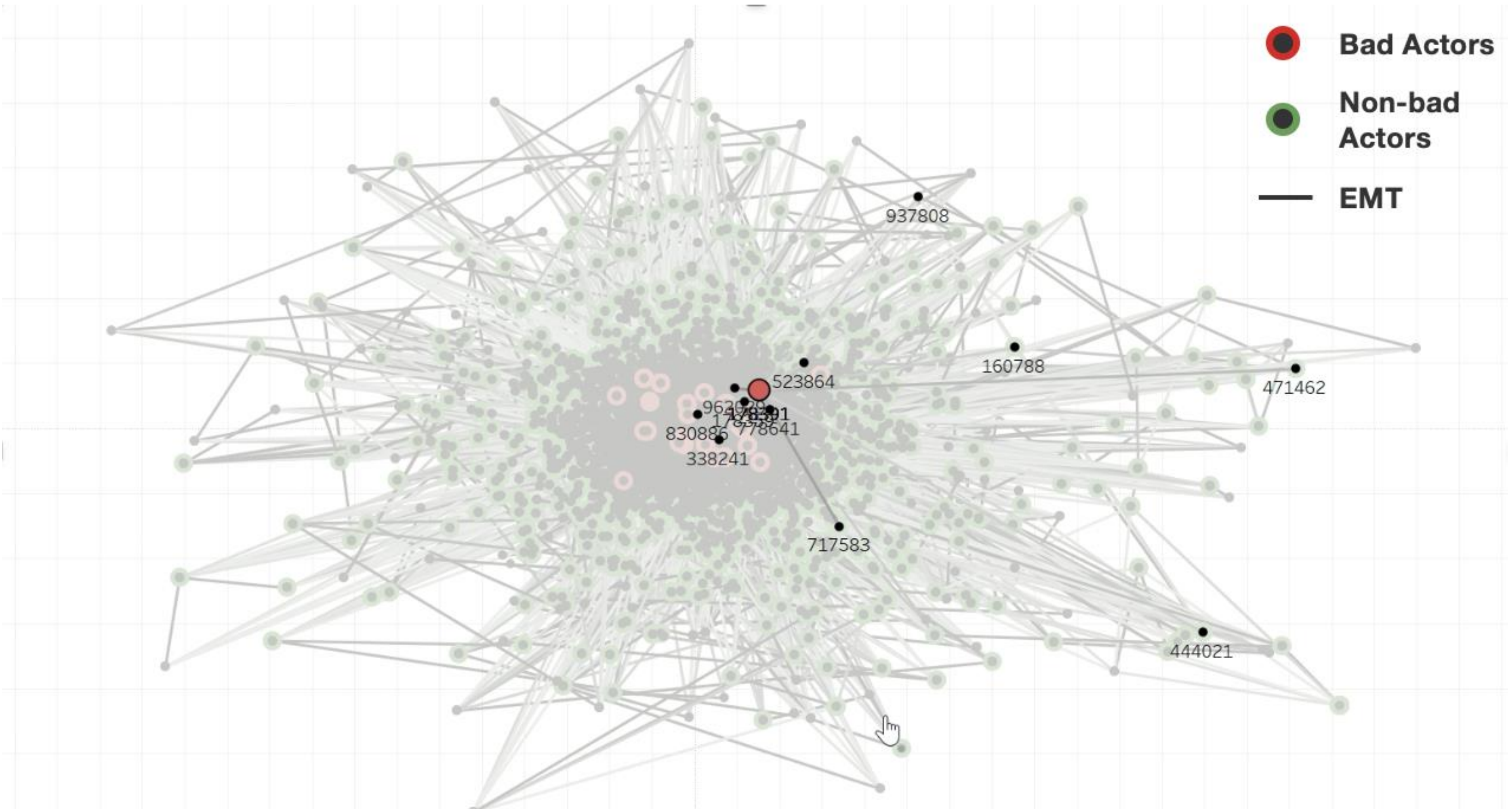
Node2Vec directed graph embedding visualization

Bad actors as middle man for layering



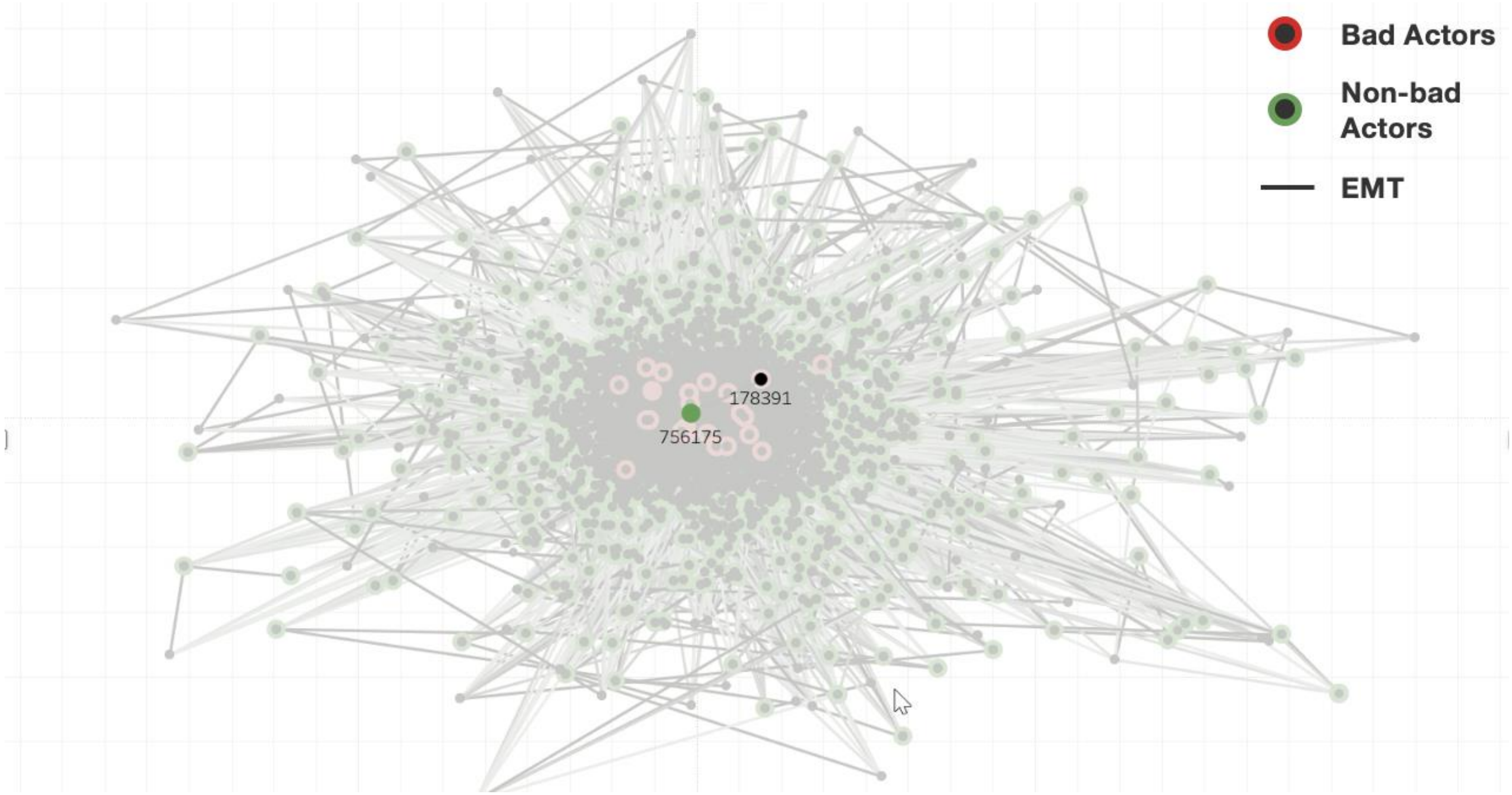
Node2Vec directed graph embedding visualization

Bad actors as middle man for layering - Out Transactions



Node2Vec directed graph embedding visualization

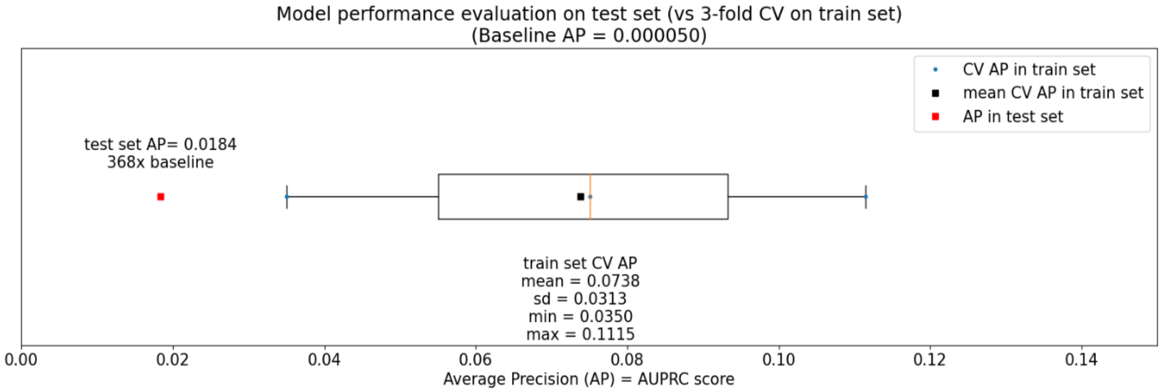
Bad actors as middle man for layering – In Transactions



Task 3 Graph Analytics

Performance improvements and importance of graph embeddings

High AP = 0.0184 (368x baseline)



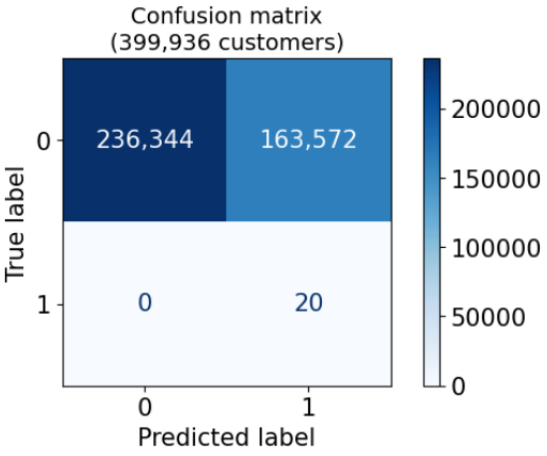
Model performance evaluation on test set (vs 3-fold CV on train set)
(Baseline AP = 0.000050)

Classification Tradeoff

Prob. Threshold = 0.0011

100% **recall**

41% **FPR**



Confusion matrix
(399,936 customers)

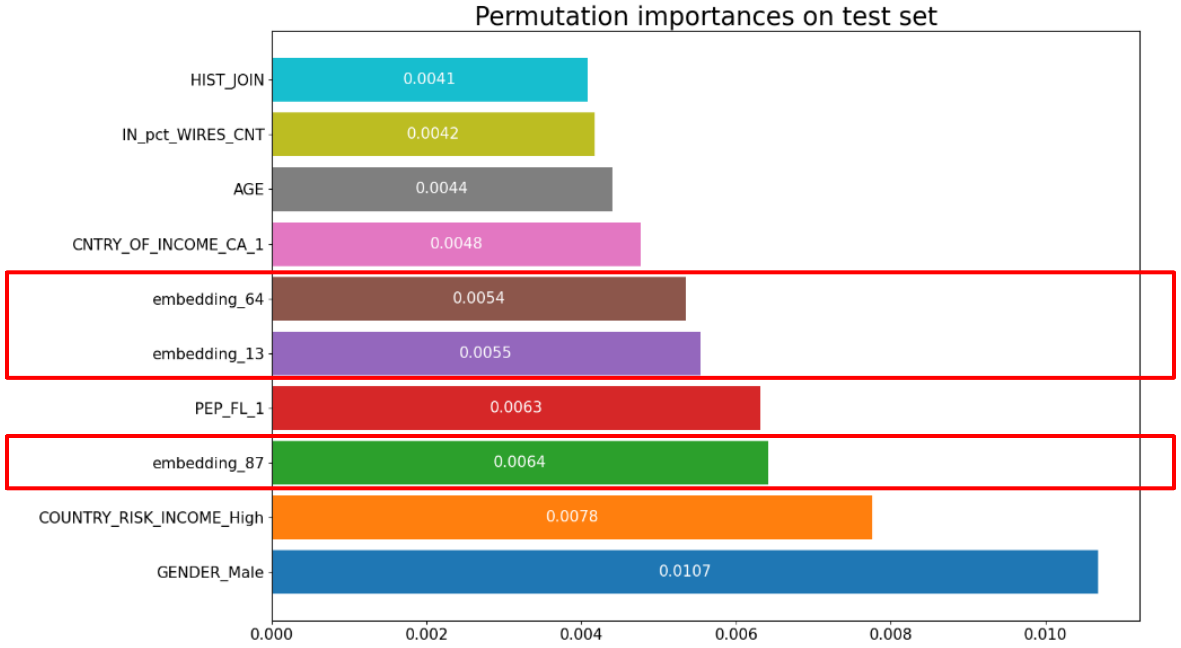
Improvement on Task 2B

Average Precision

10x improvement

FPR with 100% Recall

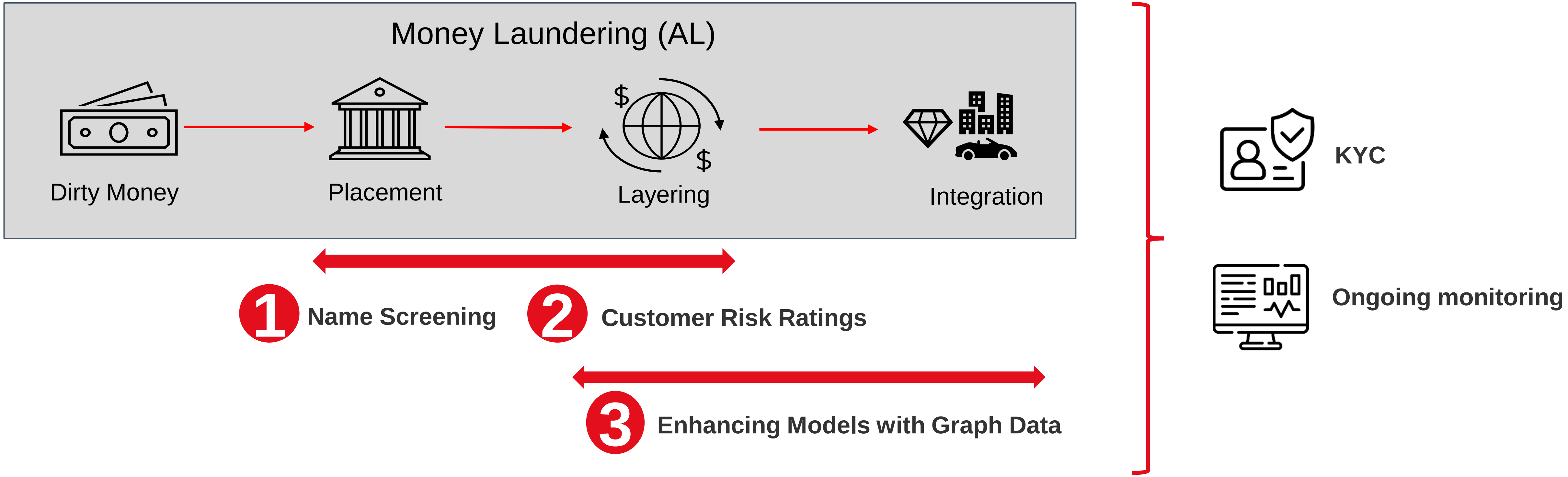
Reduced FPR by **11%**



Permutation importances on test set

Conclusions and recommendations

Conclusions and Recommendations



Money Laundering (AL)

Dirty Money

Placement

Layering

Integration

1 Name Screening

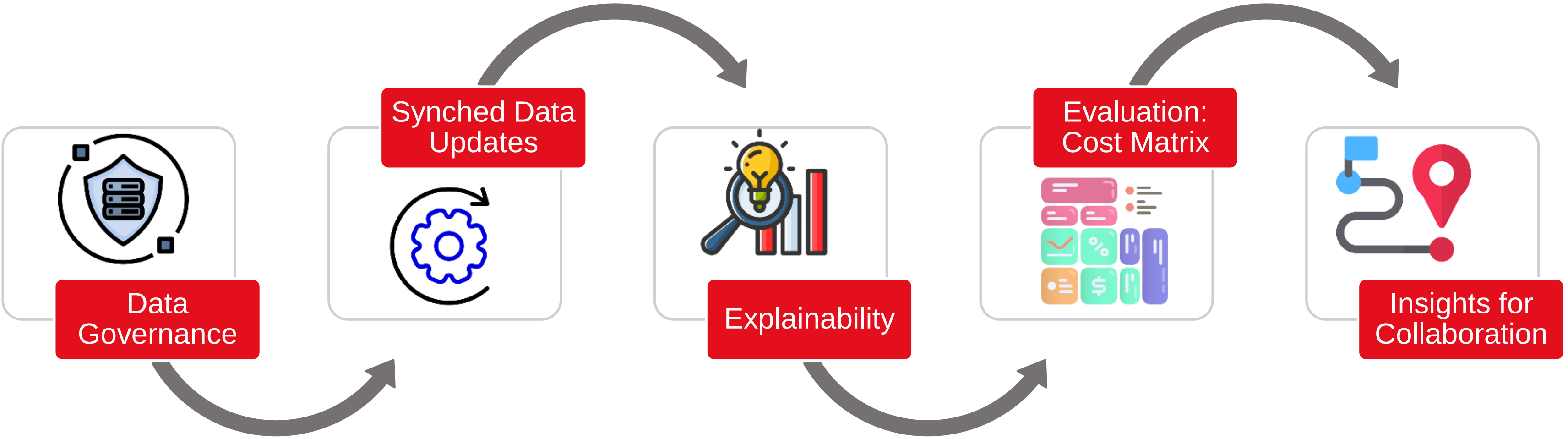
2 Customer Risk Ratings

3 Enhancing Models with Graph Data

KYC

Ongoing monitoring

Enablers



Data Governance

Synched Data Updates

Explainability

Evaluation: Cost Matrix

Insights for Collaboration

Thank you

March 25th, 2023