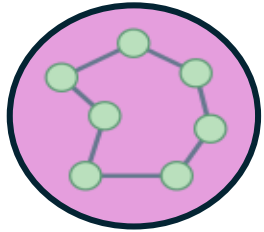


# **A Scalable, Explainable Machine Learning Approach for Granular-Level Credit Dataset's Quality Assurance**

**Anak Yodpinyanee, Peranut Nimitsurachat,  
Nontawit Cheewaruangroj, Supachai Saengthong**

# Regulatory Data Transformation (RDT)

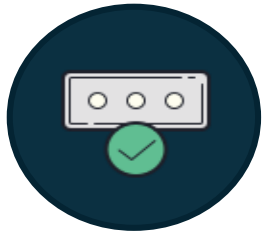
A transformative redesign from static Data Management System to **granular-level** regulatory reporting standard for **credit data**



Include variables across many dimensions of credit data such as credit lines, interest rate plan, and outstanding amount



Multi-faceted data architecture with specialized cubes for diverse and thorough credit data analysis



Traditional validation framework ensuring quality through format/range check, referential integrity check, and data consistency check

This project aims to enhance RDT validation framework with **scalable** model that will **accurately** detect **subtle anomalies** and **interpret results** effectively

# Model Selection

## Three requirements

### Accuracy



**Accurately identify  
anomalies of any types  
among all existing anomalies  
(high recall)**

### Scalability



**Scalable on our production  
stack (Spark) without  
consuming excessive  
computational resources**

### Explanability



**Accurately identify and  
rank the top fields that  
cause anomalies**

# Model Selection

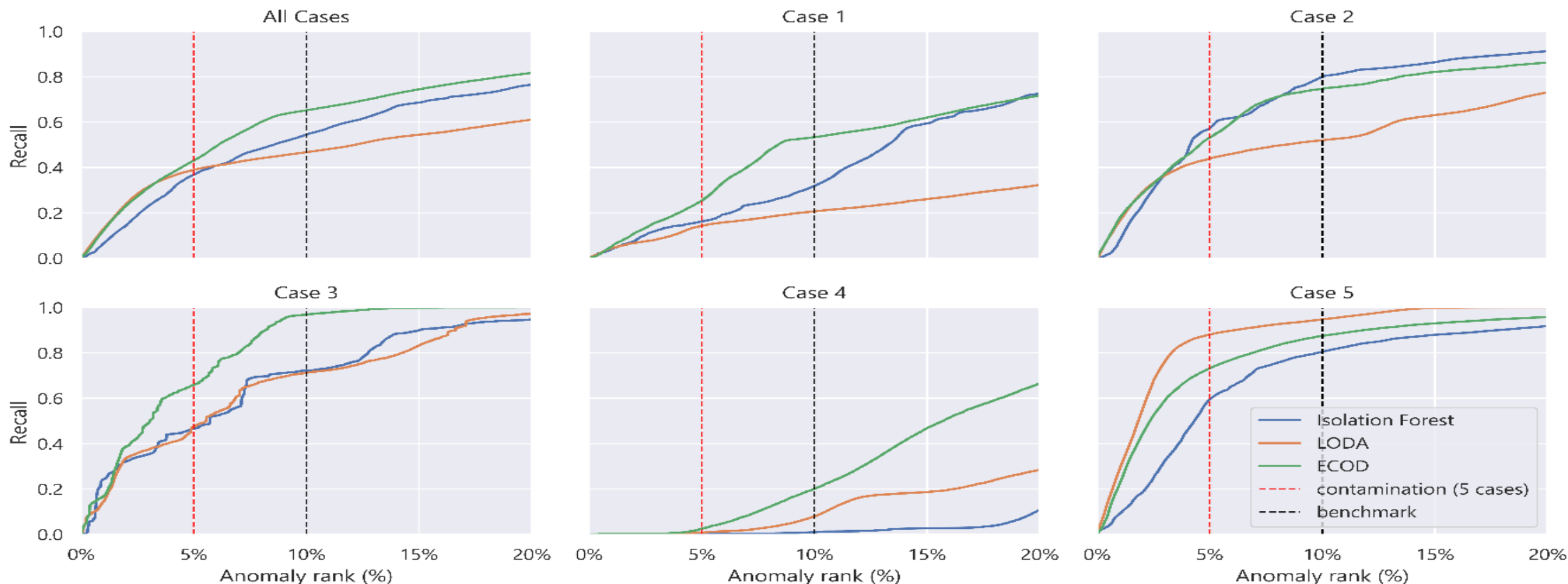
	Model Summary	Strength	Weakness
<b>ECOD</b>	<ul style="list-style-type: none"><li>• Estimate tail probabilities for each dimension using empirical cumulative distribution</li><li>• Compute anomaly score by aggregating probabilities from all dimensions</li></ul>	<ul style="list-style-type: none"><li>• Interpretability</li><li>• Simplicity of the algorithm</li><li>• Fast computation time</li><li>• High accuracy compared to other SOTA models</li></ul>	<ul style="list-style-type: none"><li>• Designed for linear relationship among variables</li><li>• Feature independence assumption</li></ul>
<b>LODA</b>	<ul style="list-style-type: none"><li>• Use random projection and histogramming to detect anomalies</li><li>• Anomaly score is computed from the bin's sparsity across multiple projections</li></ul>	<ul style="list-style-type: none"><li>• Interpretability</li><li>• Simplicity of the algorithm</li><li>• Fast computation time</li></ul>	<ul style="list-style-type: none"><li>• Designed for linear relationship among variables</li></ul>
<b>Isolation Forest</b>	<ul style="list-style-type: none"><li>• Detect anomalies by randomly selecting features and splitting values and construct binary trees</li><li>• Compute scores by considering average path length from root to anomalies</li></ul>	<ul style="list-style-type: none"><li>• More robust to non-linear relationships among variables</li><li>• Relatively fast compared to other ML anomaly detection models</li></ul>	<ul style="list-style-type: none"><li>• Scalability issue and slower runtime than ECOD and LODA</li><li>• Interpretability</li></ul>

# Model Evaluation

Case	Scope Condition	Anomaly Condition	Normality Condition	Fraction
1	Stage = Non-Performing	$PD \leq 50\%$	$PD \geq 80\%$	1.770%
2(a)	Stage = Performing	$DPD \geq 40$	$DPD \leq 30$	0.967%
2(b)	Stage = Under-Performing	$DPD \geq 100$	$DPD \leq 90$	0.411%
3	Firm Size = Small or Micro	Credit Line $\geq 300M$ THB	Credit Line $\leq 100M$ THB	0.208%
4	Revolving Flag = False	Utilization Ratio $\geq 1.5$	Utilization Ratio $\leq 1.2$	0.520%
5	Any	Pledge / Valuation $\geq 2$	Pledge / Valuation $\leq 1$	1.136%



# Recall Rates for Anomaly Detection



*This chart shows the recall rates of anomaly detection algorithms, for all five cases and for each individual case. The x-axis represents the fraction of total data points reported, and the y-axis represents the fraction of anomalous samples uncovered by the reported samples.*

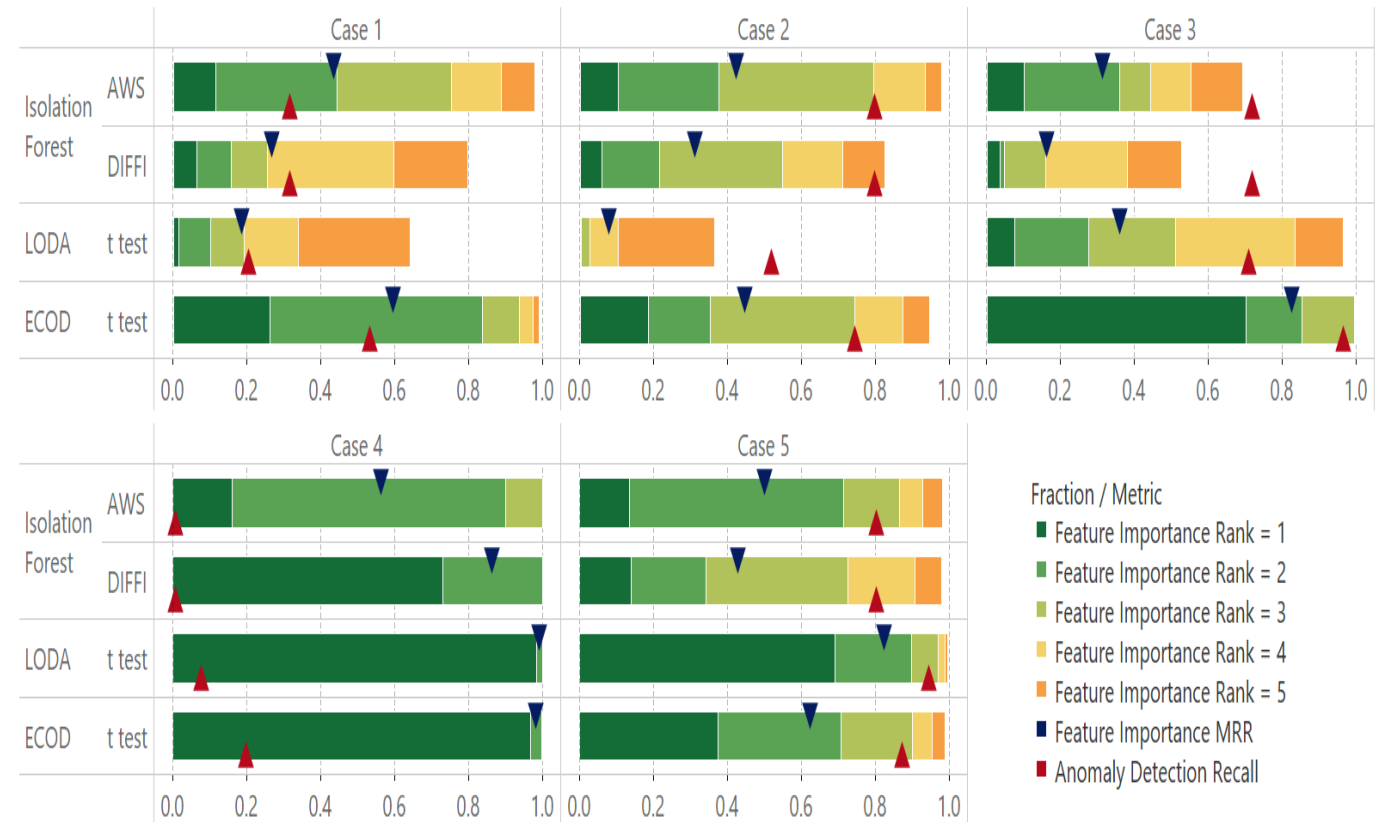
# Interpretability & Feature Importance

Mean Reciprocal Ranks

Case	Isolation Forest		LODA	ECOD
	AWS	DIFFI	t test	
1	0.436	0.271	0.188	<b>0.597</b>
2	0.425	0.313	0.082	<b>0.448</b>
3	0.315	0.166	0.363	<b>0.827</b>
4	0.564	0.865	<b>0.993</b>	0.984
5	0.502	0.430	<b>0.825</b>	0.626

For each cell, MRR (truncated at k = 5 features) is reported.

Occurrences of Anomaly Features by Feature Importance Rank



This stacked bar chart displays the rank of the first correct anomaly detected by each method and case. Bar width indicates the fraction of correctly detected anomalies at each rank; x-axis shows the cumulative percentage of correctly explained anomalies. Metrics MRR and recall are marked by triangles.

# Discussion & Conclusion

## Model Performance

- LODA excels at recognizing linear patterns in numerical data but has difficulty with discrete variables
- Isolation Forest excels at dealing with conjunctions and categorical variables, but offers mediocre performance otherwise.
- ECOD performs well across diverse anomaly types and mixed variable types.

## Model Interpretability

- ECOD's t-test offers competitive model interpretability, aligning with its strong anomaly detection performance.
- Top-3 features explain over 70% of anomalies in all cases. For Isolation Forest, AWS outperforms DIFFI in feature scoring accuracy.
- All methods are more effective for numerical features than categorical ones, due to simpler encoding.