Feature Stability Index (FSI): A Multi-Axis Metric for Assessing Robustness of Features in Imbalanced Fraud Detection

ID I-7

Dalia Breskuviene and Gintautas Dzemyda,
Data Science and Digital Technologies Institute, Vilnius University

Motivation

- Evolving data distributions introduce temporal instability.
- Feature importance varies strongly across model architectures.
- Random initialization and sampling amplify instability.
- High-dimensional engineered feature spaces mask the true signal.

Purpose

In high-dimensional fraud detection, feature importance often changes drastically when the model architecture varies, when randomness affects training, or when the data is drawn from a different time period. This instability makes it difficult to identify features that genuinely contribute to fraud detection versus those that appear important only under specific experimental conditions.

Contribution

We propose the Feature Stability Index (FSI) - a multi-axis stability metric that measures feature robustness across:

- time windows (concept drift),
- model architectures,
- random seeds.

FSI integrates continuous feature importances, uses variance and entropy-based metrics, and provides a reliable stability score for real-world fraud detection systems.

Proposed Feature Stability Index

This framework evaluates how consistently each feature behaves across time windows, model types, and random seeds. Feature importance is computed for every model/seed/window combination, aggregated along each dimension, and converted into stability scores using coefficient-of-variation and entropy measures.

Input Dataset

↓

Split into time windows (w)

↓

For each model type m
For each random seed s
Tain model (m, s) on window w

↓

Compute feature importance for each tuple (f, w, m, s)

$$I_{f,w,m,s} = \begin{cases} |B_f|, & m = logreg \\ MDI\ importance, & m = rf \\ XGBoost\ Gain\ importance, & m = xgb \\ PredictionValueChange, & m = catboost \\ PermutationImportance, & otherwise \end{cases}$$

↓

Aggregate importance across:

- Windows: $I_{f,w}$

For each feature f and each window w:

$$\bar{I}_{f,w} = \frac{1}{N_{f,w}} \sum_{(m,s)\epsilon M\times S} I_{f,w,m,s}$$

where:
- $N_{f,w}$ =number of runs in window w for feature

Resulting aggregated values across **all windows**: $\{\bar{I}_{f,w} : w = 1, \ldots, W\}$

- Models: $I_{f,m}$

For each feature f and each model m:

$$\bar{I}_{f,m} = \frac{1}{N_{f,m}} \sum_{(w,s)\epsilon W\times S} I_{f,w,m,s}$$

where:
- $N_{f,m}$ =number of runs for model m

Resulting aggregated values across **all model types**: $\{\bar{I}_{f,m} : m\ \epsilon\ M\}$

- Seeds: $I_{f,s}$

For each feature f and each seed s:

$$\bar{I}_{f,s} = \frac{1}{N_{f,s}} \sum_{(w,m)\epsilon W\times M} I_{f,w,m,s}$$

where:
- $N_{f,s}$ =number of runs for seed s

Resulting aggregated values across **all seeds**: $\{\bar{I}_{f,s} : s\ \epsilon\ S\}$

↓

Calculate stability metrics:

CV-based: $S\_var\_CV, S\_model\_CV, S\_seed\_CV$

$$S = 1 - \frac{\sigma}{|\mu| + \varepsilon}$$

IE-based: $S\_var\_IE, S\_model\_IE, S\_seed\_IE$

$$S = 1 - H(p)$$

$$H(p) = -[p \log_2 p + (1-p) \log_2 (1-p)]$$

↓

Compute final FSI scores:

$$\mathrm{FSI_{CV}} = \alpha Sv + \beta Sm + \gamma Ss$$

$$FS\_IE = \alpha Sv + \beta Sm + \gamma Ss$$

Method	Continuous Importances	Handles Ranks	Time Windows	Model Diversity	Random Seeds	Decomposes Variability	Suitable for FD
Kuncheva Index	✗	✗	✗	✗	limited	✗	✗
Nogueira Stability	✗	✗	✗	✗	partial	✗	✗
Rank Correlation	✓ (after ranking)	✓	✗	✗	✓	✗	partially
Importance Correlation	✓	✓	✗	✗	✓	✗	partially
FSI (Proposed)	✓	✓	✓	✓	✓	✓ **(Sv, Sm, Ss)**	✓

Data Used for experiments

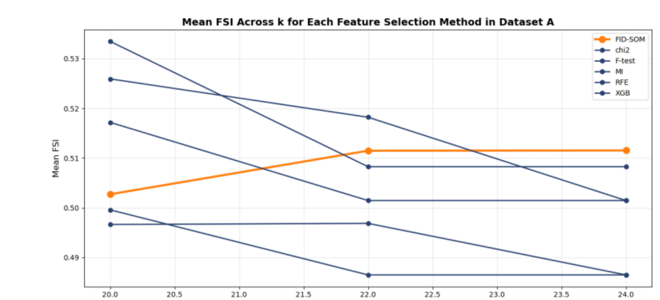
DataSet-A: a large synthetic credit card transaction dataset containing over 3.27 million records, designed to mimic highly imbalanced real-world fraud scenarios.

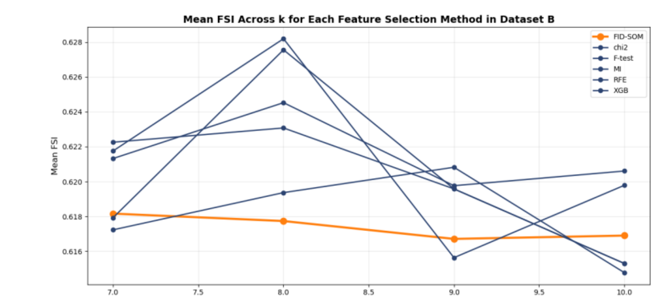
DataSet-B: A structurally different synthetic dataset with 1.85 million transactions, used to assess method robustness across varying feature sets and distributions.

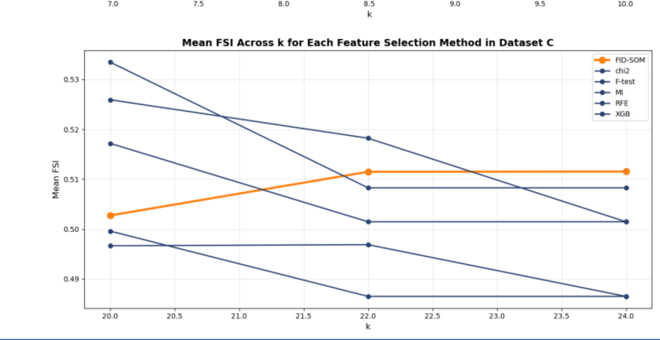
DataSet-C: A strongly anonymized real financial transaction dataset with 284,807 records, included to validate the practical effectiveness of the proposed method on realistic data.

Experiments

Although FID-SOM does not consistently achieve the top score, its key strength lies in its exceptional robustness and cross-dataset consistency. It avoids the large performance fluctuations seen in traditional selectors and maintains high, stable performance across all tested datasets.







References

1. "Breskuvienė, D., Dzemyda, G.: Categorical feature encoding techniques for improved classifier performance when dealing with imbalanced data of fraudulent transactions. International Journal of Computers Communications & Control 18(3) (2023)"
2. Breskuvienė, D., Dzemyda, G. (2024). Enhancing credit card fraud detection: highly imbalanced data case. Journal of Big Data, 11(1), 182.