УДК 004.85:336.717 **DOI:** 10.31673/2412-9070.2025.050346
G. GAINA, Candidate of Technical Sciences, Prof.;
ORCID: 0000-0003-0260-0950
D. MASIUK, PhD student,
ORCID: 0009-0002-0840-5558
Taras Shevchenko National University of Kyiv

EVALUATING RULE-BASED VS. MACHINE LEARNING APPROACHES FOR FRAUDULENT TRANSACTION DETECTION

Financial institutions nowadays rely heavily on rule engines, such as thresholds, white/black lists, velocity checks to flag suspicious transactions. Machine learning (ML) models, on the other hand, while promising higher accuracy and adaptability, are being dependent on data characteristics, class imbalance, latency constraints, and interpretability requirements. In this paper, I present a controlled evaluation of a configurable rule-based baseline and several supervised ML models (logistic regression, random forest, gradient boosting) on an imbalanced transaction dataset. I measure detection performance (ROC-AUC, PR-AUC, precision/recall at operating points), operational costs (false-positive rate, alerts per 1k transactions), and engineering trade-offs (inference latency, feature complexity, interpretability). Results show that while rules remain competitive at high-precision, low-recall regimes, ML approaches achieve substantially better recall at comparable precision, especially when coupled with calibrated thresholds and class-imbalance handling. I discuss deployment-oriented considerations and outline a hybrid strategy that combines rules for policy compliance with ML for generalization.

Keywords*:* fraud detection; financial risk; anomaly detection; rule engine; machine learning; class imbalance; interpretability.

Introduction

Card-present and card-not-present payments, peer-to-peer(p2p) transfers, and mobile money platforms constantly face a serious risk of fraud. Rule engines remain attractive due to simplicity and auditability, but they struggle with evolving attack patterns and high volumes. One has to manually change them in time, in order to catch the latest attack strategy. ML models can adapt to complex interactions in features (amount, merchant, device, velocity signals), however they are more demanding. They want labeled data, model governance, and latency budgets. Expanding on my earlier article on computational intelligence in eCommerce fraud detection, this paper moves from a conceptual overview toward a hands-on comparison. Previously, I highlighted the layered structure of antifraud systems—blocklists, rules, and scoring engines—and mentioned the growing need for hybrid methods that blend domain knowledge with machine learning. Here, I want to extend that discussion by experimentally comparing a pure rule-based baseline with machine learning classifiers. I make three contributions in this paper. First and foremost, I formalize a transparent rule-based baseline, representative of industry practice (thresholds, velocity checks, lists). Secondly, I compare it against standard supervised models under severe class imbalance, tracking metrics all the way. Thirdly, I provide deployment guidance, such as when to favor rules, when ML wins and how to combine them.

Related work

Fraud detection in financial systems has been extensively studied, which might pose a question of scientific novelty. Yet no single method has emerged as universally optimal. It might be that finding such a method is impossible, and systems that combat fraud must constantly evolve and be handheld by their creators. However, I would like to either find alternatives or prove that theory once and for all. Early surveys [1] provide an overview of detection techniques, from rule-based systems and

statistical methods to machine learning and deep learning. Their respective trade-offs are noted, in the form of cost, accuracy and implementation speed. Other authors [2] extend this approach by classifying a number of academic studies into categories by model type, such as classification, clustering, regression, and outlier detection. Thus, highlighting the dominance of supervised models in credit card and insurance fraud. Nevertheless, there are still areas that remain underexplored, for example money laundering and its countermeasures. More up-to-date work makes stronger emphasis on hybrid and ensemble approaches [3]. It demonstrates that combining unsupervised scoring models with supervised classifiers can significantly improve precision, especially under concept drift. There are also similar comparisons [4], of supervised models such as different boosting algorithms and random forests, to unsupervised ones, such as generative adversarial networks. Those show that while supervised models win overwhelmingly in terms of Area Under the Receiver Operating Characteristic Curve (AUROC) metric, they perform significantly worse on datasets with scarce data. There are also studies [5-7], which propose complex hybrid ensembles that incorporate both decision trees, neural networks and transformers, claiming they show great results in terms of detective power. However, they seem to be prone to overfitting and due to such complex structure – severe lack of interpretability.

In this paper I am expanding on my earlier report [9], where I previously outlined the layered antifraud architecture commonly used by payment providers. Those usually consist of some combination of block/allow lists, rule engines, and scoring models, which operate either independently or in combination. There I emphasize the benefits of hybrid systems, where rules are being weighted dynamically by machine learning models. That approach combines "the best of both worlds" – flexibility of models, and interpretability of rule-based engine. I also stressed the importance of interpretability to meet regulatory and ethical requirements. This paper builds on this, by empirically comparing a rule-based scoring baseline with supervised machine learning models under severe class imbalance.

Experiments

Baseline comparison

Our first experiment would be to compare baseline rule-based approach to baseline machine learning models. For the dataset I chose Kaggle Credit Card fraud dataset [11], which satisfies our requirements of being anonymized and containing high class imbalance. This dataset contains approximately 285000 transactions, of which only 492 (0.17%) are labeled as fraudulent. This is actually a good illustration of a real fraud problem, where you are looking at a really small subset of transactions. Each transaction here is described by 30 anonymized numerical variables (V1-V28, Time and Amount).

Firstly, we split the data into training (60%), validation (20%) and test (20%) using stratified sampling, in order to preserve class imbalance in each subset. We then pick three models, in the form of Logistic regression (LR), Random Forest (RF) with 400 tress and maximum depth of 12, and Gradient Boosting (GB) with 300 estimators and learning rate of 0.1. For a rule-based system we construct a simple additive risk score for each transaction.

$$S(x) = \sum_{i=1}^{k} I(x_i < \tau_i) \quad (1)$$

Here *S* is a risk score, *x* is a value of a specific feature, *t* is a threshold for a respective feature and *I* is an indicator function. We used two modes for a rule-based engine going forward, Precision favored – where *S >=3*, and Balanced, where *S >=2*. We also evaluated both approaches using the following metrics.

$$P = \frac{tp}{Tp+fp} \quad (2)$$

$$R = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = \frac{P*R}{P+R} \quad (4)$$

$$Alerts = 1000 * \frac{Flagged\ transaction}{Total\ transactions} \quad (5)$$

Here (2) is Precision, (3) is Recall and (4) is F1 score. We also used Area Under Precision Recall Curve (AUROC). We then looked at all possible threshold for each feature, to see which would be more suitable to stop there. At each step we calculated precision, recall and f1 score. Finally, we have chosen the threshold with best precision-recall balance, which equals to the best F1 score. After training the models and comparing results, we have discovered that results of machine learning models where overwhelmingly better (fig.1).

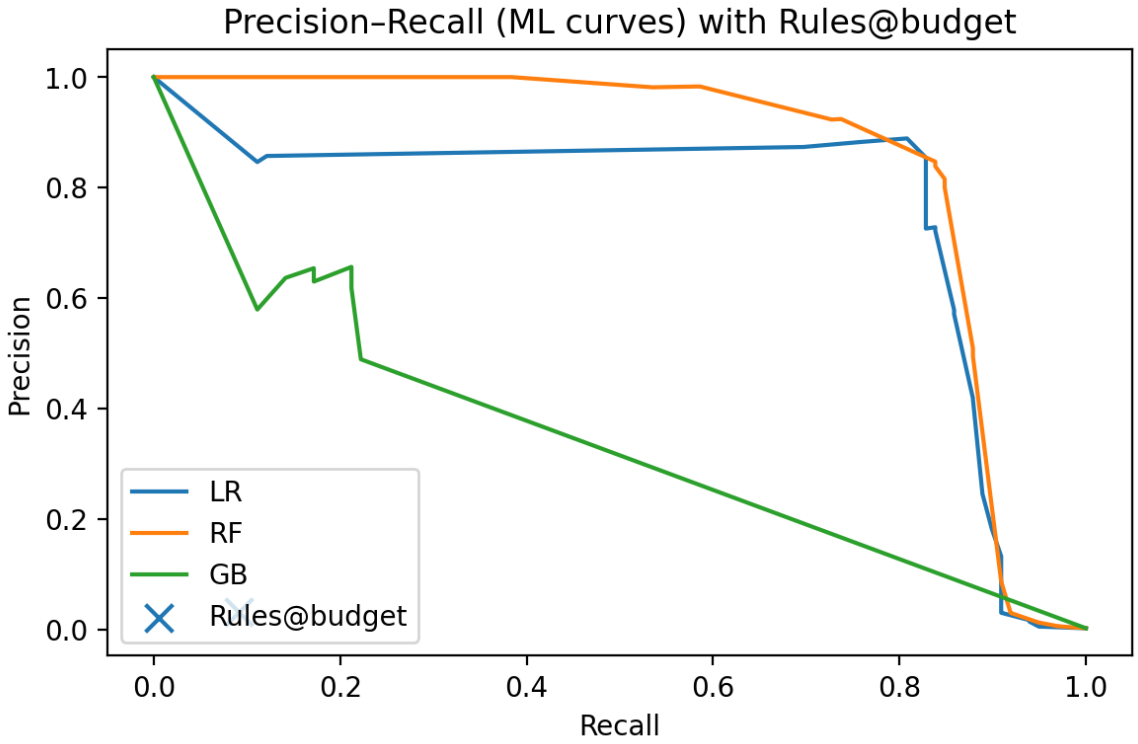

Fig. 1. Model results significantly overperform rule-based approach

Overall, it seems that machine learning models have a significant lead here, however it is important to note that training and testing of those models took up approximately 16 minutes, rule-based approach was almost instant. With this in mind, we move to the second experiment.

Advanced ruleset

In this experiment, we would like to see if we can push the rule-based approach to be on par with machine learning models results, while maintaining efficiency and interpretability. In order to improve from baseline, we are going to use Information Value to understand which of our features hold the most predictive power. We are going to calculate this metric the following way.

$$IV = \sum_{i=1}^{n}(p_i - q_i) * \ln(\frac{p_i+\varepsilon}{q_i+\varepsilon}) \quad (6)$$

Here, p is a proportion of fraud cases in a respective bin i, while q is a proportion of non-fraud cases in the same bin. After calculating the metric for all features and looking at the results, we have created a few groups via predictive power, in order to see which metrics are the most important (table).

IV value thresholds I have chosen to use in our approach

IV value	**Predictive strength**
<0.02	Not useful
0.02-0.1	Weak

0.1-0.3	Medium
0.3-0.5	Strong
>0.5	Very strong

With this grouping in mind, we have selected the best features for our approach V4, V14, V12, V17, and V10. We also combined individual rules in two ways – precision favored, which needs three and more rules to trigger, and balanced which need two.

After testing we have discovered that this more advanced rule-based approach can perform as decent, as baseline machine learning models (fig. 2). And once again, it is important to note, that IV calculations have taken up approximately 2 minutes, against 16 minutes of training machine learning models. In the face of constantly evolving fraud threat, we would have to train such models or make and tweak such rules on almost a day-to-day basis, and this x8 difference in time would certainly be noticed. The point here is that, while machine learning models seem superior at baseline and I am sure would be superior in case we dedicate time to tune their hyperparameters and better arrange our data, rule-based approach is simple, fast and still quite effective. In order for the model ensemble to work properly and be maintained – you need a qualified specialist, while with a rule-based option and argument could be made that with enough explanation almost anyone can make the tuning. Moreover, rules are easily interpretable because they can be read and understood by humans, which leads to easier implementation into the workspace and much easier maintenance.

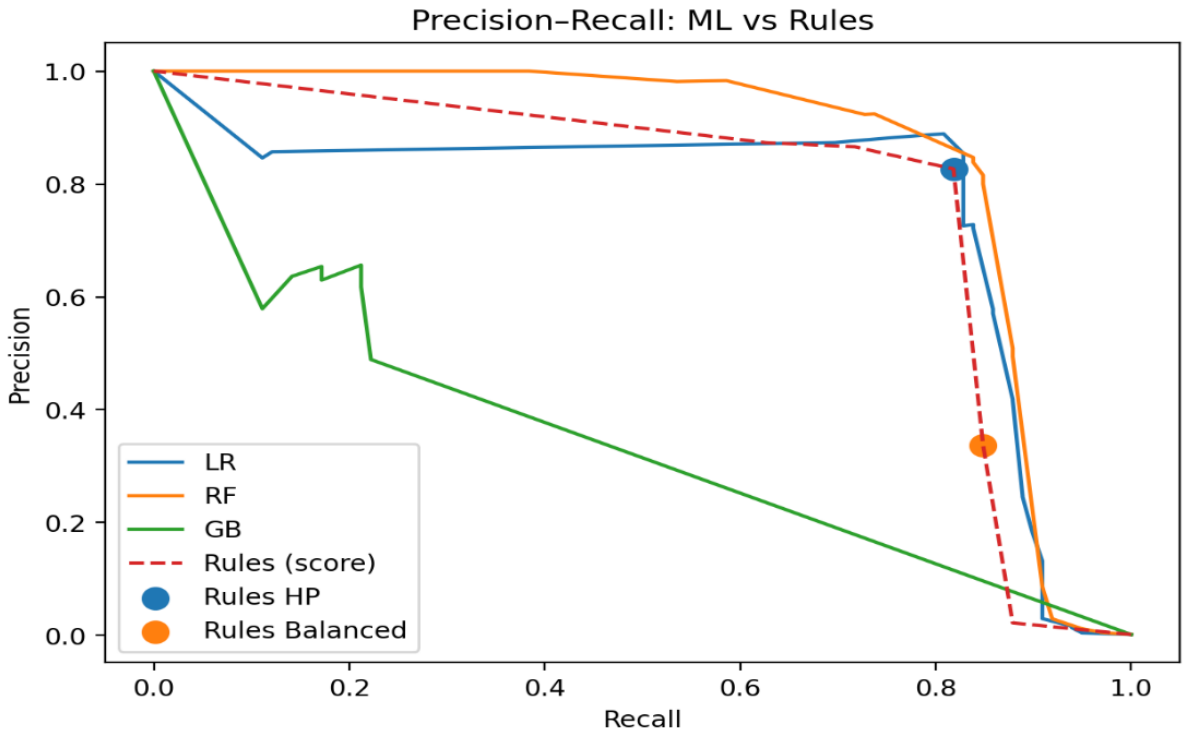

Fig. 2. Precision favored approach performs on par with our top models

Advanced models

In order to further push our theory, we now want to test our improved rules against something better than just baseline models. Main concern for our models here is that fraud is too rare. Baseline learners minimize average error and can optimize and win by predicting everything as non-fraud. As mentioned by other authors[8], the simplest way to combat that is undersampling. We will reduce the amount of non-fraud data, in order to better train for fraud detection. The data is going to be split into train, validation and test as usual. Then, on train dataset only we are going to drop examples of

a majority class until minority to majority ratio reaches the desired values, say ½. Then we are going to fit the model on the undersampled dataset, and evaluated on the untouched datasets as usual.

Alternatively, we will try oversampling in the form of Synthetic minority oversampling technique (SMOTE). What it does is takes an example of a minority class, say *x*, within that minority class find its k nearest neighbors. We will look for 5. Randomly pick one of the neighbors, and generate a synthetic sample along the line between the original example and the random neighbor.

$$x_{new} = x + \delta * (x_{nn} - x) \quad (7)$$

We will try to train LF and GB on undersampled datasets, and RF on SMOTE dataset. The results were not surprising (fig. 3). While the models did outperform even our improved rule-based approach, it took almost an hour on my hardware to compute and train all three models. Of course, an enterprise would have a much more competent hardware at hand, however the fact of needing such a hardware can be interpreted as a cost. Thus, we come to a conclusion – that while this was definitely an improve-ment, it was indeed costly, five times more costly than our previous experiment. I am interested in comparing such approaches in terms of cost in my future works.

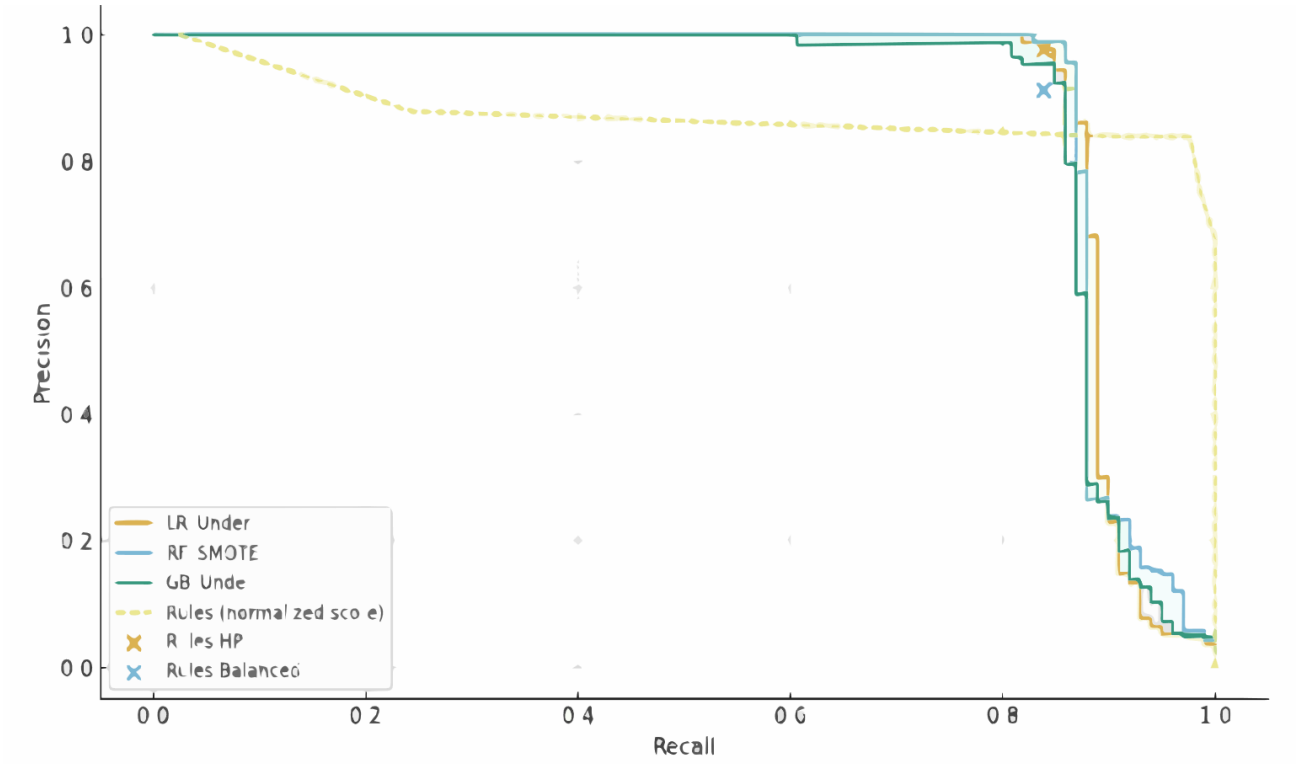

Fig. 3. Advanced models against advanced rules

Conclusion

Our experiments show that there is promise in rule-based approach still, and that there are pros and cons to both methods. Machine learning models are better out of the box and probably in the long run too, but require a significant resource investment in the form of time and skill. Moreover, they are hard to explain to a bystander which might pose challenges during workspace implementation and explaining result to potential clients. Additional work has to be done in order to research ethical and legal challenges that such system might pose.

Rule-based approach on the other hand is much faster, interpretable and can be tuned to match baseline models without much effort. However, additional research is needed to understand what might happen if one needed to scale such a system.

I would also be quite interested in comparing different approaches in terms of a cost metric, and that is something I will focus on in future.

References

1. Delamaire, L (2009). Credit card fraud and detection techniques: A review.

2. Ngai, E. W. T. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review.

3. Carcillo, F.(2019). Combining unsupervised and supervised learning in credit card fraud detection.

4. Xuetong Niu (2024). A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised.

5. Yvan Lucas (2019). Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs.

6. Alamin Talukder (2024). Securing Transactions: A Hybrid Dependable Ensemble Machine Learning Model using IHT-LR and Grid Search.

7. Diego Vallarino (2025). Detecting Financial Fraud with Hybrid Deep Learning: A Mix-of-Experts Approach to Sequential and Anomalous Patterns.

8. Dal Pozzolo, A.(2015). Calibrating Probability with Undersampling for Unbalanced Classification.

9. Dmytro Masiuk (2025). Analytical overview of computational intelligence application in e-Commerce fraud detection.

Г. А. Гайна, Д. В. Масюк

ПОРІВНЯЛЬНИЙ АНАЛІЗ ПІДХОДІВ, ЩО БАЗУЮТЬСЯ НА ПРАВИЛАХ ТА МАШИННОГО НАВЧАННЯ ДЛЯ ВИЯВЛЕННЯ ШАХРАЙСЬКИХ ТРАНЗАКЦІЙ

Фінансові установи сьогодні значною мірою покладаються на системи, побудовані на логічних правилах, що включають порогові значення, «білі»/«чорні» списки, перевірки кількості, щоб знаходити підозрілі транзакції. Такі методи використовуються широко насамперед через високу пояснюваність, так як для фінансових установ важливо мати можливість показати кінцевому користувачу, чому саме внутрішні системи та/або співробітники прийняли певне рішення. Також, важливим є те, що такі методи просто побудувати і підтримувати, бо через відсутність складних алгоритмів за допомогою їх легше навчити співробітників. Моделі машинного навчання, з іншого боку, хоча й обіцяють вищу точність і адаптивність, залежать від характеристик даних, дисбалансу класів, часових рамок та вимог до інтерпретованості. Для побудови ансамблів подібних моделей необхідно мати багато кваліфікованого людського ресурсу. При масштабуванні такого ансамблю на компанію, витрати часу і людських ресурсів будуть значно вищими. У цій статті описується контрольована оцінка конфігурованої системи правил і кількох моделей контрольованого ML (логістична регресія, випадковий ліс, градієнтний бустинг) на незбалансованому наборі транзакційних даних. Вимірюється якість виявлення шахрайства за допомогою метрик ROC-AUC, PR-AUC, а також точності і повноти. Вимірюються операційні витрати (рівень хибнопозитивних результатів, кількість сповіщень на 1 тис. транзакцій) та інженерні компроміси, наприклад, інтерпретованість. Результати показують, що хоча правила залишаються конкурентними в режимах високої точності при низькій повноті, ML-підходи забезпечують суттєво кращу повноту за подібної точності, особливо у поєднанні з методами обробки дисбалансу класів. Однак, вказується, що простіші алгоритми суттєво виграють в часі, особливо на середньому обладнанні, та у вартості, що може зробити їх привабливішими для менших установ з обмеженими ресурсами. Обговорюються аспекти застосування методів у реальних робочих просторах та окреслюється гібридна стратегія, яка поєднує правила та машинне навчання. Закладається фундамент для поглибленого дослідження таких підходів саме в рамках метрик «вартості» і «швидкості».

Ключові слова: виявлення шахрайства; фінансовий ризик; виявлення аномалій; логічні правила; машинне навчання; дисбаланс класів; інтерпретованість.