Financial Data Science: The Accuracy vs. Explainability Paradox in Balancing Predictive Power and Transparency

An Analysis Report

By

Megang Nkamga Junile Staures

Department of Mathematical Modelling, Kaunas University of Technology (KUT)

Foundation of Data Science (FDS)

Date of submission

April 20, 2025

Table of Contents

1 Introduction

The accuracy-explainability paradox in machine learning presents significant challenges in finance, where high-stakes decisions require both predictive power and regulatory compliance. This tension arises because complex models like neural networks often achieve superior accuracy but lack transparency, while interpretable models like logistic regression provide clearer decision-making logic at the potential cost of performance (Bell et al., 2022).

2 Theoretical Analysis

2.1 Key Implications in Financial Contexts

2.1.1 Regulatory Compliance and Ethical Accountability

Financial institutions must adhere to strict regulations, such as GDPR and the Equal Credit Opportunity Act, which require clear explanations for credit denials and risk assessments. For instance, banks using opaque models risk penalties if they cannot justify loan rejections, and systems detecting suspicious activities must provide justifications to avoid false accusations (Bailey, 2025; Zhang, 2024).

2.1.2 Model Trust and Adoption Barriers

Research indicates that explanations from SHAP (Shapley Additive Explanations) significantly enhance understanding of black-box models among non-experts, yet users often over-trust these explanations without proper validation. This underscores the need for hybrid models that combine accurate predictions with understandable outputs, as well as training programs to improve critical evaluation of AI-generated results (Bracke et al., 2019).

2.1.3 Performance Trade-offs in Practice

Empirical research reveals nuanced outcomes. In public policy contexts, interpretable models often match black-box accuracy (Bell et al., 2022). In financial applications, gradient-boosted models outperform linear regression by 8-12% AUC in credit scoring but require SHAP to bridge the explainability gap (Bracke et al., 2019; Bailey, 2025).

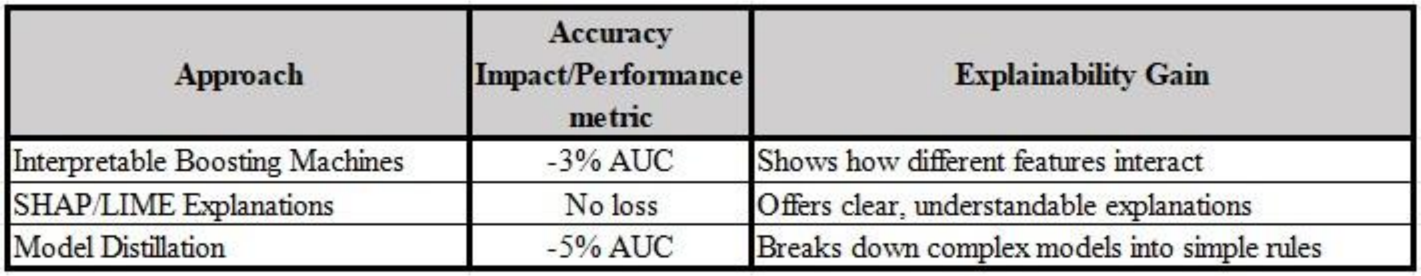
2.2 Real-World Case: European Central Bank’s (ECB’s) AnaCredit project

The ECB’s granular credit dataset requirement for banks to use machine learning for default predictions and to provide feature importance rankings has compelled institutions to either adopt interpretable models or invest in explanation tools. This has led to a 15-20% increase in development costs, while enhancing auditability (Zhang, 2024).

2.3 Emerging Solutions

Financial institutions are trying to solve this tricky problem: they need to create models that are both accurate and easy to understand. Here’s how they are addressing this:

Table 1 Source: Adapted from (Zhang, 2024; Bailey, 2025)



Approach	Accuracy Impact/Performance metric	Explainability Gain
Interpretable Boosting Machines	-3% AUC	Shows how different features interact
SHAP/LIME Explanations	No loss	Offers clear, understandable explanations
Model Distillation	-5% AUC	Breaks down complex models into simple rules

Some studies suggest that the trade-off between accuracy and explainability isn’t as big of a deal when dealing with simpler financial data (Bell et al., 2022). However, regulations require that decisions be explained, even if this might lower accuracy. This need for explainability is pushing the development of new AI techniques designed for finance, though there are still challenges in creating consistent ways to measure how transparent these models are (Zhang, 2024; Wouters & Hammann, 2025). The finance industry shows that finding the right balance isn’t a one-size-fits-all solution. Choosing the best model depends on factors like legal risks, data complexity, and the costs of making mistakes (Milvus, 2025; Wouters & Hammann, 2025).

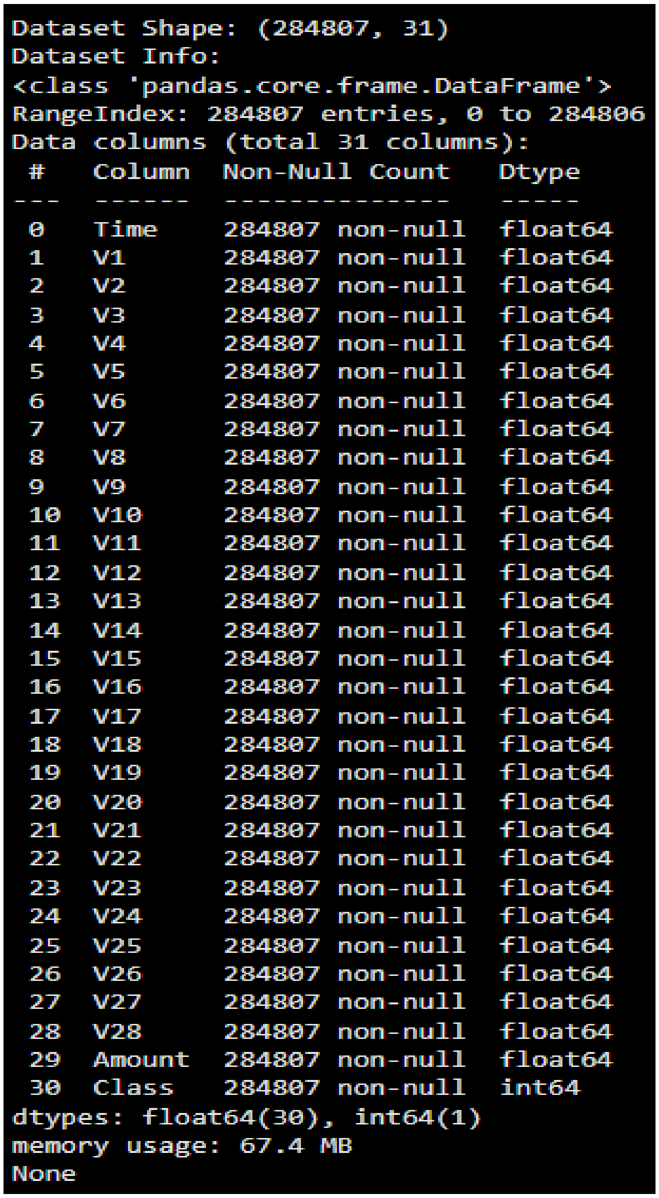
.

3 Empirical Experiment

3.1 Objective

This study demonstrates the trade-off between accuracy and explainability by comparing a simple, interpretable model (Logistic Regression, LR) with a more complex, less interpretable model (Random Forest, RF) on a financial classification task

3.2 Dataset



```
Dataset Shape: (284807, 31)
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 284807 entries, 0 to 284806
Data columns (total 31 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   Time     284807 non-null  float64
 1   V1       284807 non-null  float64
 2   V2       284807 non-null  float64
 3   V3       284807 non-null  float64
 4   V4       284807 non-null  float64
 5   V5       284807 non-null  float64
 6   V6       284807 non-null  float64
 7   V7       284807 non-null  float64
 8   V8       284807 non-null  float64
 9   V9       284807 non-null  float64
 10  V10      284807 non-null  float64
 11  V11      284807 non-null  float64
 12  V12      284807 non-null  float64
 13  V13      284807 non-null  float64
 14  V14      284807 non-null  float64
 15  V15      284807 non-null  float64
 16  V16      284807 non-null  float64
 17  V17      284807 non-null  float64
 18  V18      284807 non-null  float64
 19  V19      284807 non-null  float64
 20  V20      284807 non-null  float64
 21  V21      284807 non-null  float64
 22  V22      284807 non-null  float64
 23  V23      284807 non-null  float64
 24  V24      284807 non-null  float64
 25  V25      284807 non-null  float64
 26  V26      284807 non-null  float64
 27  V27      284807 non-null  float64
 28  V28      284807 non-null  float64
 29  Amount   284807 non-null  float64
 30  Class    284807 non-null  int64
dtypes: float64(30), int64(1)
memory usage: 67.4 MB
None
```

Figure 1 Overview of the credit card fraud detection dataset used in this study, obtained from Kaggle (Source:https://www.kaggle.com/code/gpreda/credit-card-fraud-detection-predictive-models/input)

Figure 1 presents an overview of the credit card fraud detection dataset, which we downloaded from Kaggle. This dataset includes:

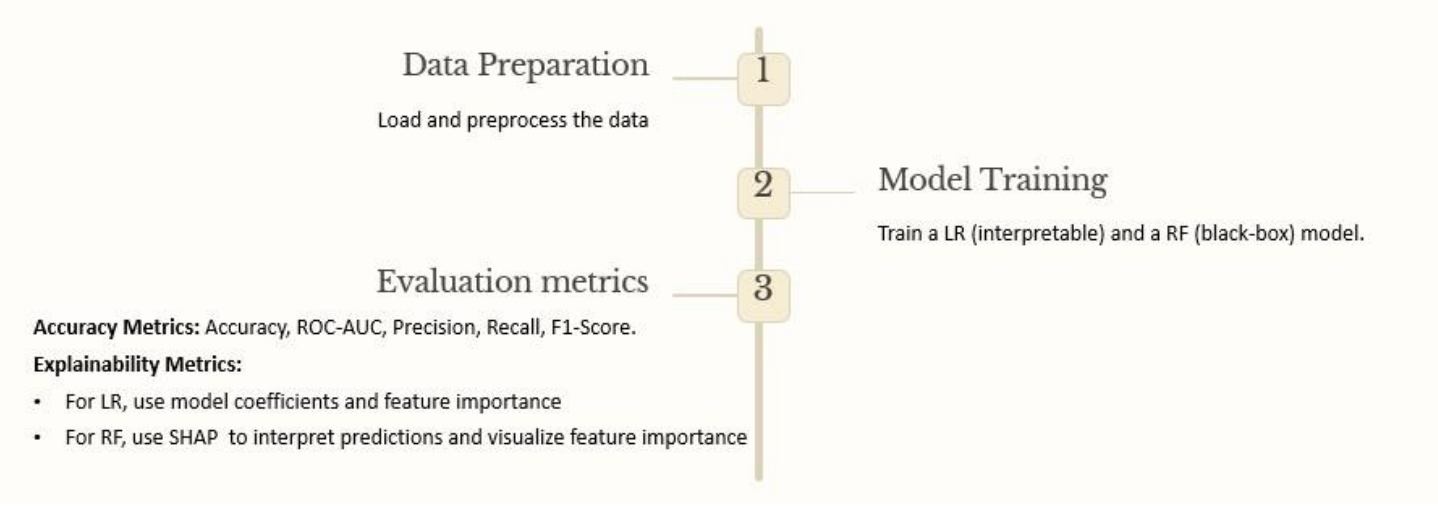
✧ The Time variable which indicates the elapsed time since the first transaction

✧ V1 to V28 are the results of PCA transformations, which help protect sensitive information

✧ Amount represents the transaction amount

✧ Class (target) which indicates whether the transaction is fraudulent (1) or not (0).

3.3 Methodology



Data Preparation

Load and preprocess the data

Model Training

Train a LR (interpretable) and a RF (black-box) model.

Evaluation metrics

Accuracy Metrics: Accuracy, ROC-AUC, Precision, Recall, F1-Score.

Explainability Metrics:

- For LR, use model coefficients and feature importance
- For RF, use SHAP to interpret predictions and visualize feature importance

Figure 2 Method utilized

3.3.1 Data Preparation

Duplicate transactions were removed, and the dataset was checked for and found to have no missing values. The 'Amount' and 'Time' features were analyzed for outliers. Both features were then standardized using the StandardScaler to ensure they have a mean of zero and a standard deviation of one.

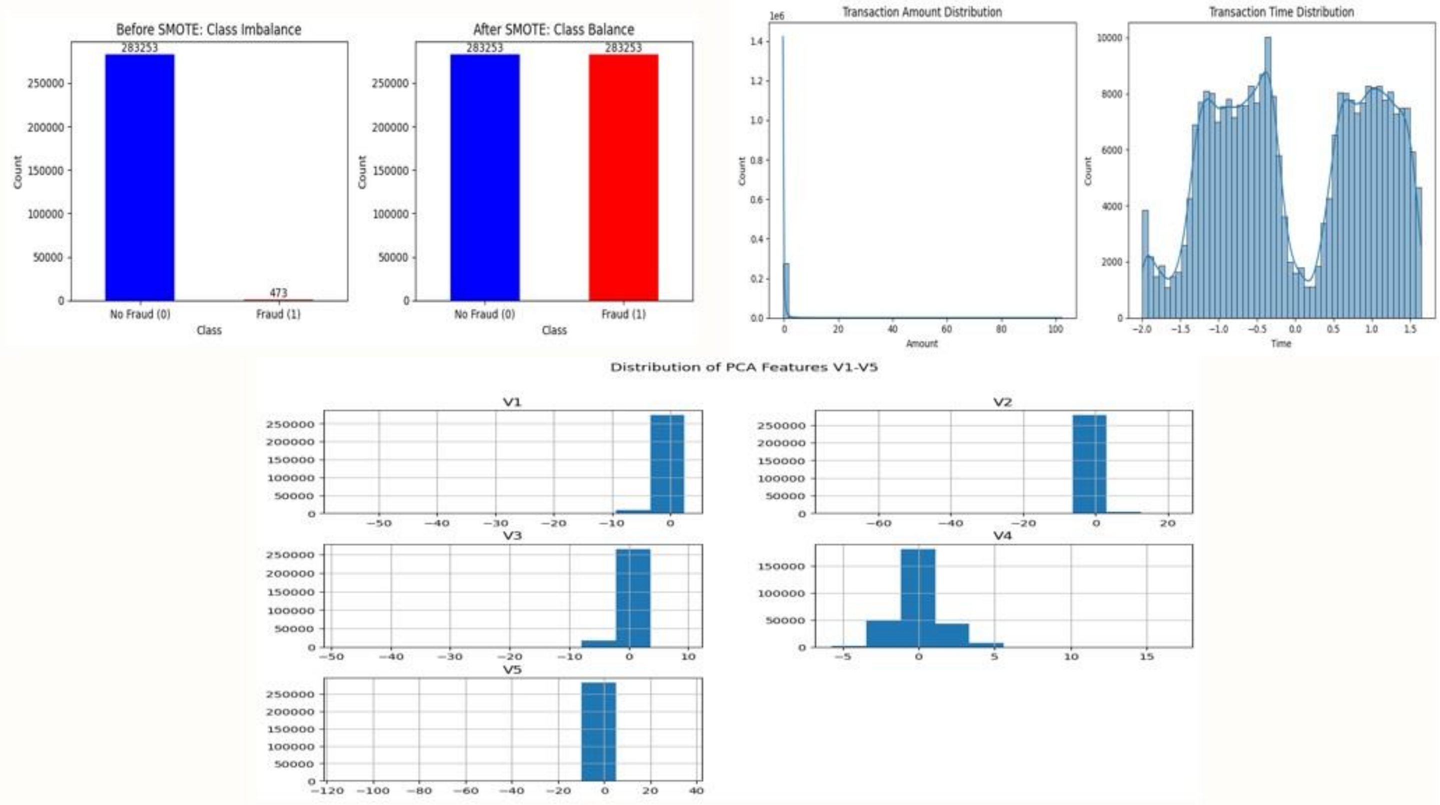
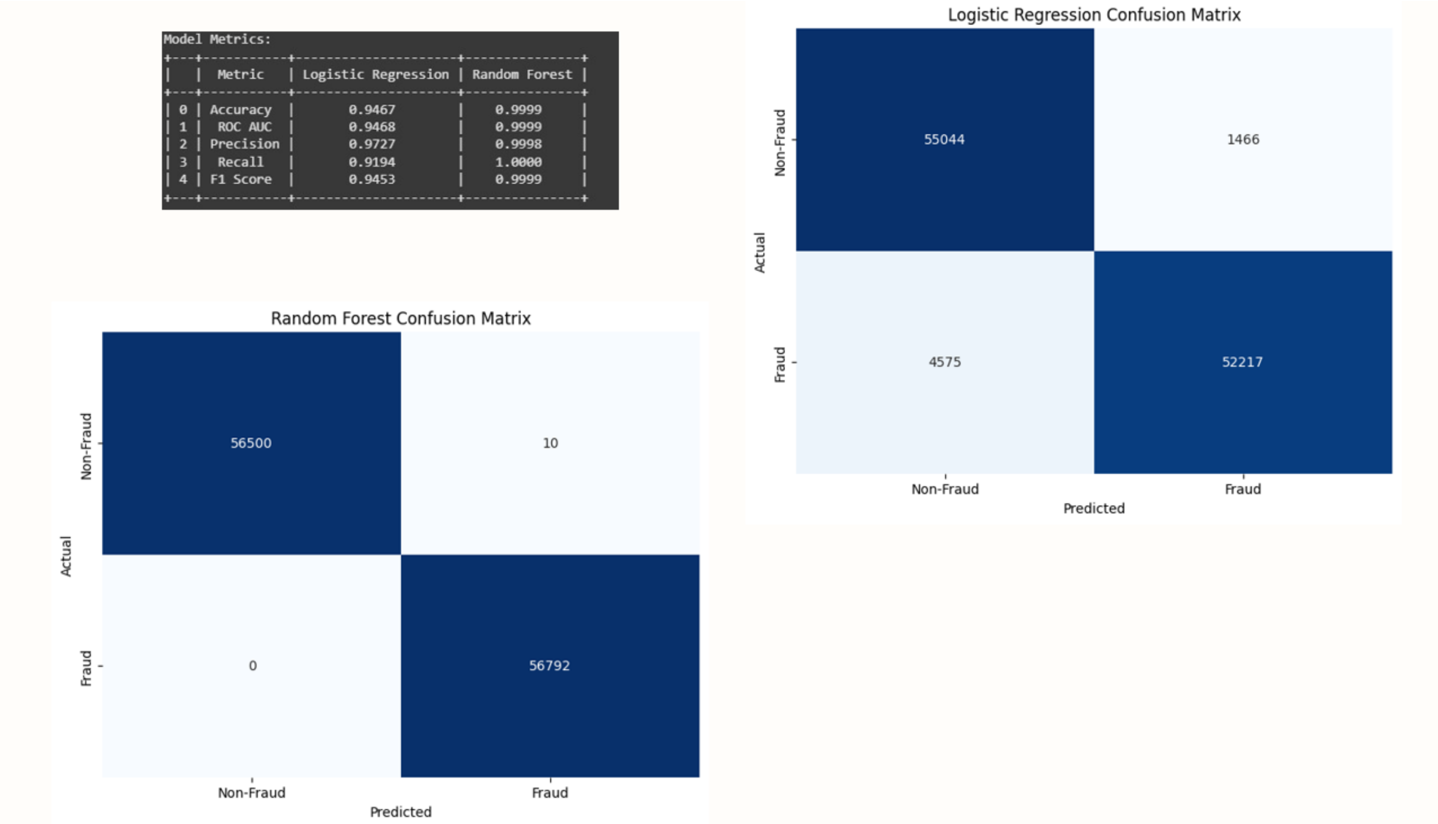


Figure 3 Overview of data preparation steps used in this study

The provided PCA-transformed features (V1–V28) were utilized for dimensionality reduction. The dataset was split into training and testing sets using an 80/20 split to enable unbiased model evaluation. To address class imbalance, SMOTE was applied to balance the number of fraud and non-fraud cases. (See appendices for the complete code.)

3.3.2 Model Training and Metrics Evaluation



```
Model Metrics:
+---+-----------+---------------------+---------------+
|   |  Metric   | Logistic Regression | Random Forest |
+---+-----------+---------------------+---------------+
| 0 | Accuracy  |       0.9467        |    0.9999     |
| 1 |  ROC AUC  |       0.9468        |    0.9999     |
| 2 | Precision |       0.9727        |    0.9998     |
| 3 |  Recall   |       0.9194        |    1.0000     |
| 4 | F1 Score  |       0.9453        |    0.9999     |
+---+-----------+---------------------+---------------+
```

Figure 4 Performance Metrics for Fraud Detection Models

Figure 4 demonstrate that the Random Forest model outperformed Logistic Regression across all metrics, achieving nearly perfect accuracy (99.99% vs. 94.67%) and ROC AUC (0.9999 vs. 0.9468). It correctly identified all fraud cases with minimal false positives or negatives, while Logistic Regression missed several frauds and misclassified more transactions. Overall, Random Forest proved to be a far more reliable and effective model for credit card fraud detection in this dataset.

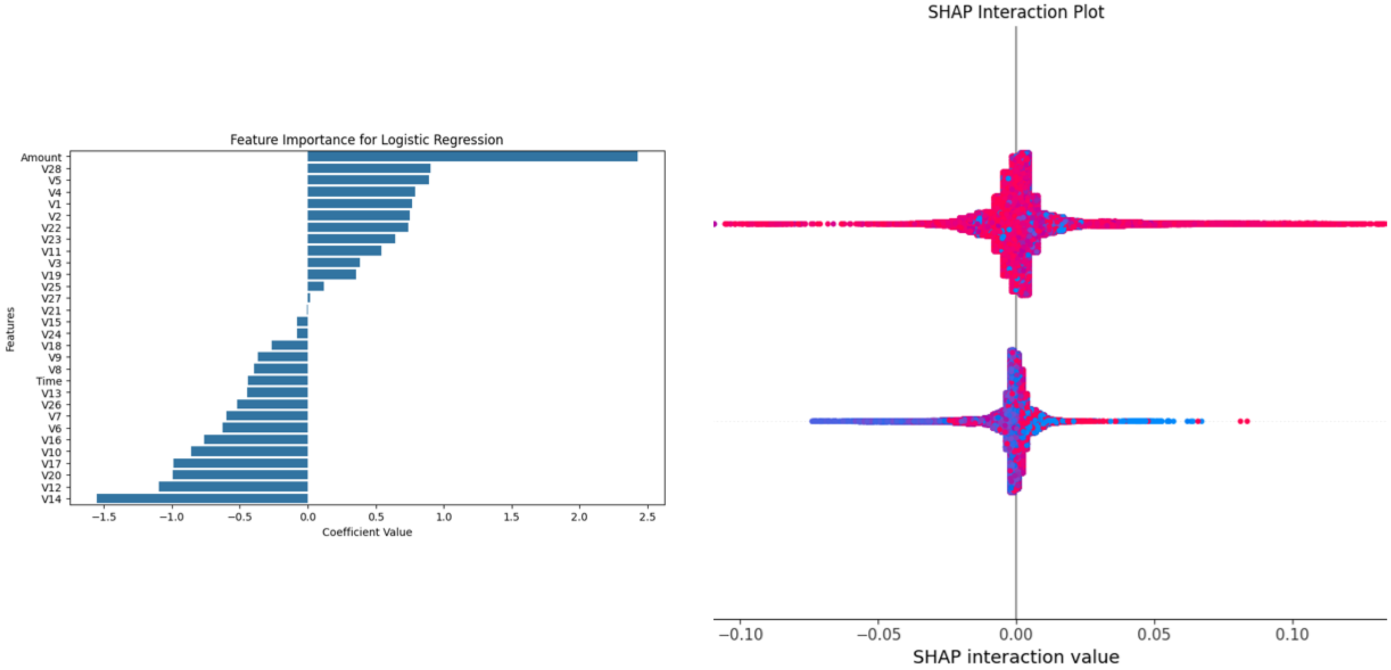


Figure 5 Feature importance and interactions

The SHAP interaction plot (Figure 5) shows the distribution of feature contributions to the model predictions. The symmetrical pattern around zero indicates both positive and negative contributions from features, with most interactions clustered around the centre. The distribution suggests that while many features have minimal impact individually, their combined interactions help distinguish between fraud and legitimate transactions.

The feature importance plot for Logistic Regression (Figure 5) reveals that Amount has the strongest positive correlation with fraud prediction, followed by V28. Conversely, V14 shows the strongest negative association with fraud, followed by V12. These opposing influences help create decision boundaries for classification.

.

4 Discussion

The experiment highlights the accuracy-explainability paradox in financial fraud detection: while the Random Forest model achieves much higher accuracy than Logistic Regression, it is inherently less interpretable. SHAP plots help to visualize and partially bridge this explainability gap by illustrating how different features contribute to the predictions made by complex models. The confusion matrices further reveal the practical consequences of model selection, as Random Forest misses almost no fraud cases, whereas Logistic Regression fails to detect many, underscoring the significant financial implications of these choices. Feature analysis identifies specific variables that are strongly associated with either fraudulent or legitimate transactions, which align well with established domain knowledge and existing literature. Overall, the findings reinforce that as data complexity increases, the trade-off between accuracy and explainability becomes more pronounced, particularly in high-stakes applications like fraud detection. Although advanced tools such as SHAP can enhance model interpretation, they introduce additional complexity and require specialized expertise to be used effectively.

5 Conclusion

This study empirically demonstrates the accuracy-explainability paradox in financial data science by comparing Random Forest and Logistic Regression models in a fraud detection context. The results show that while Random Forest significantly outperforms Logistic Regression in classification accuracy, achieving 99.99% versus 94.67%, it requires additional interpretability tools such as SHAP analysis to provide the level of transparency necessary for regulatory compliance and stakeholder trust. These findings align with the theoretical framework and existing literature, confirming that financial institutions face real trade-offs when selecting machine learning approaches, as superior performance in minimizing missed frauds can lead to substantial cost savings but introduces challenges for explainability. For practical applications, a hybrid approach is recommended: leveraging complex models like Random Forest for their predictive power, while investing in advanced explainability techniques to meet regulatory and stakeholder needs. Future research should focus on developing improved explanation methods and examining the accuracy-explainability trade-off across different financial contexts, while the adoption of standardized explainability metrics would help institutions systematically balance accuracy and transparency in high-stakes model selection.

6 References

Bailey, K. (2025, March 28). *Why Explainable AI is Critical for Financial Decision Making*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/artificial-intelligence-ai/why-explainable-ai-matters-finance/

Bell, A. J., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). *It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy*. https://doi.org/10.1145/3531146.3533090

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019, August 9). *Machine Learning Explainability in Finance: An Application to Default Risk Analysis*. Papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3435104

Milvus, Z. (2025). *What are the trade-offs between explainability and accuracy in AI models?* Milvus.io. https://milvus.io/ai-quick-reference/what-are-the-tradeoffs-between-explainability-and-accuracy-in-ai-models

Wouters, M. J. F., & Hammann, D. (2025). Explainability versus Accuracy of Machine Learning Models: The Role of Task Uncertainty and Need for Interaction with the Machine Learning Model. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5125772

Zhang, H. (2024, November 10). *The Application of Machine Learning in Finance: Situation and Challenges*. Ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5120025

7 Appendices

Link to code
(https://colab.research.google.com/drive/1ymGOCAUgUYj3qwpBLlCm4uZkGuofCyA7#scrollTo=H4IBTHRIvWVo)