Explainable Self-Attention Modeling for Financial Vulnerability Prediction

Veronica Xu[*], Weixian Waylon Li[*], Joshua Ong Jun Leang[†], Tiejun Ma[*]
Luis Felipe Costa Sperb[‡], Fengxiang He[*]

February 2026

Abstract

This study presents an explainable self-attention framework for predicting individual financial vulnerability using large-scale longitudinal banking data. While achieving strong discriminative performance (AUC-ROC ≈ 0.90), the model prioritises transparency through outcome-aware feature attribution analysis across all prediction classes. Our results reveal that high-risk classifications are driven by asymmetric "risk activation" triggered by financial and employment volatility, specifically income variability and payment irregularity, rather than static demographic factors. Analysis of misclassifications indicates that errors stem from structural limitations in evidence aggregation rather than random noise. By linking predictive outcomes to underlying feature-level logic, this work provides a scalable approach to high-stakes automated decision-making in financial risk assessment.

1 Introduction

Integrity prediction plays an important role in a wide range of real-world applications, including financial risk assessment [4], recruitment screening [6], and fraud detection [3, 9]. In settings, automated decision-making systems are increasingly relied upon to process large-scale behavioural and financial data. Although modern machine learning models are capable of achieving strong predictive performance, they often operate as black boxes, providing limited insight into how individual predictions are generated [7]. This lack of transparency poses significant challenges, as model outputs in integrity-related applications can directly affect access to financial resources, welfare benefits, or regulatory scrutiny. Consequently, strong aggregate performance metrics, such as accuracy or area under the ROC curve (AUC-ROC), alone are insufficient. Stakeholders, including regulators, policymakers, and domain experts, require explanations that clarify why a particular prediction is made and which factors contribute most strongly to the decision.

In this project, we develop an explainable self-attention model [5, 8] for individual-level financial vulnerability prediction using large-scale longitudinal financial data provided by NatWest and Smart Data Foundry [1]. The task is formulated as a binary classification problem, where the model predicts, for each individual and weekly observation, whether the individual exhibits vulnerability-related risk based on observed financial and behavioural indicators. The proposed architecture demonstrates strong discriminative capability on held-out test data, achieving high AUC-ROC [2], and precision. Rather than treating predictive performance as the primary objective, this study systematically characterises the model's internal decision structure through outcome-aware feature attribution analysis.

By comparing attribution patterns across true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), we directly examine both correct decisions and systematic errors. The analysis reveals stable structural properties of the model's behaviour: high-risk predictions are driven by

[*]University of Edinburgh
[†]Imperial College London
[‡]Kingston University London

salient instability signals, while low-risk decisions arise when such signals fail to exceed implicit activation thresholds. Financial and employment volatility emerge as dominant drivers of vulnerability representation, and misclassifications reflect predictable limitations in evidence aggregation and thresholding rather than random noise. Together, these findings enhance the interpretability and transparency of high-stakes financial risk prediction models by systematically linking predictive outcomes to their underlying feature-level decision structures.

2 Data Description

This study uses a large-scale longitudinal financial dataset provided by NatWest and Smart Data Foundry [1], consisting of anonymised individual-level financial records derived from real-world banking data. All records were processed using privacy-preserving anonymisation and de-identification procedures before analysis.

The dataset spans the period from 4 August 2019 to 10 March 2024, and is aggregated into weekly snapshots, with each observation representing a seven-day time window. In total, the dataset comprises 240 weekly snapshots and 1,437,900 individual-week records, stored in Apache Parquet format.

Each observation corresponds to a unique anonymised individual identifier (`cid`) and is represented by a fixed-dimensional vector of 24 numerical features. These features capture multiple aspects of individual financial behaviour and are grouped into four categories: demographic attributes, income composition, expenditure structure, and financial stability indicators. A summary of the feature groupings is provided in Table 1. All numerical features are standardised prior to model training to ensure scale consistency and stable optimisation. Each weekly snapshot is treated as an independent input, and no explicit temporal dependencies are modelled.

The prediction target is a binary variable, `benefits`, indicating whether an individual receives social benefits within the corresponding weekly time window. This variable is not interpreted as a direct measure of individual integrity; rather, it serves as an observable proxy for financial vulnerability, enabling scalable analysis of income instability and economic risk patterns.

Table 1: Summary of feature categories in the dataset.

Feature categories	**Description**
Demographic features	Age band, sex, postal district
Income composition	Total income; salary, benefits, pension, investment, interest, and other income
Expenditure structure	Total, essential, committed, discretionary, quality-of-life, and uncategorised expenditure
Financial stability	Salary events, salary coefficient of variation, income variability, payment interval, unemployment status, income source
Prediction target	benefits (binary)

3 Methodology

3.1 Model Architecture

This study employs a self-attention neural network architecture [5, 8] to model individuals' financial and employment-related characteristics and to perform binary classification. Each input sample consists of a fixed-dimensional vector of 24 standardized numerical features, covering demographic attributes, income composition, expenditure structure, and financial stability indicators. Each weekly aggregated observation is treated as an independent sample, and no explicit temporal dependency is modeled.

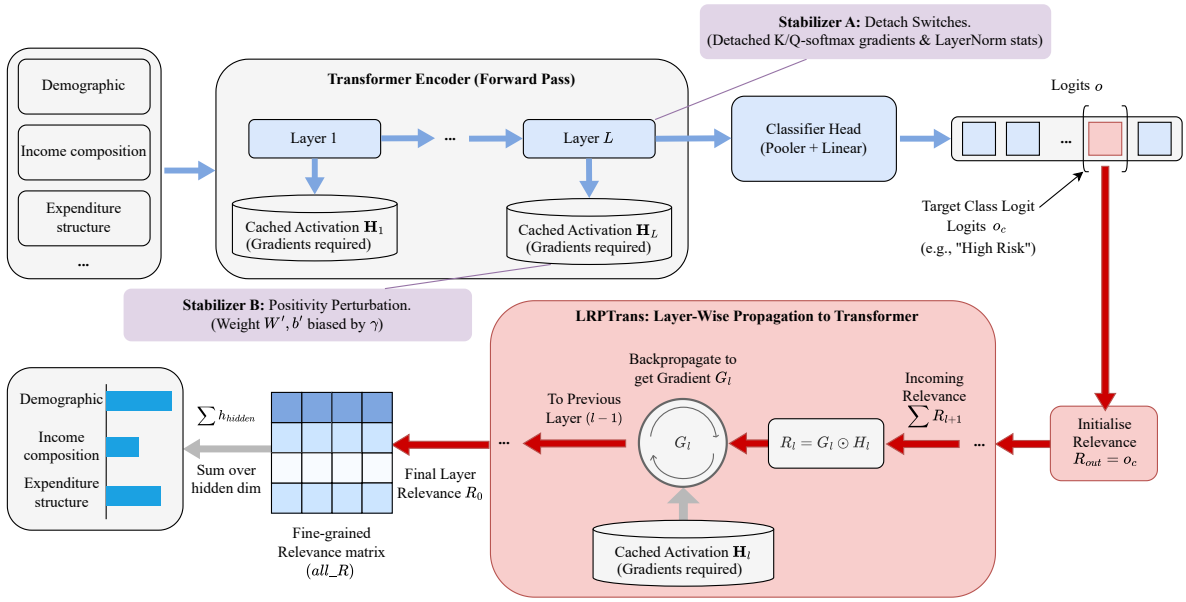


Figure 1: The XAI framework for explainable financial risk assessment.

To capture interactions among heterogeneous financial features, the model applies a self-attention mechanism along the feature dimension. Unlike conventional sequential transformers, the attention mechanism operates across feature dimensions, enabling the model to capture interactions among heterogeneous financial attributes. By projecting input features into a shared latent space and modeling their dependencies through multi-head attention, the architecture can learn complex, non-linear relationships that are difficult to represent using linear models or shallow neural networks. Stacking multiple self-attention layers further enhances the model's capacity to capture higher-order feature interactions.

The output representations from the self-attention layers are aggregated and passed to a linear classification head to generate binary predictions. Model parameters are optimized by minimizing the cross-entropy loss function. All input features are standardized prior to training, and dropout is disabled in the final configuration to ensure stable internal representations and consistent post-hoc explainability.

3.2 Explainability Methods

To enhance transparency of the model's decision-making process, we employ a feature-level explainability method aligned with the self-attention architecture [5] (Figure 1). For each individual prediction, the model computes relevance scores that quantify the contribution of each input feature to the final classification output. For systematic evaluation, samples are grouped according to predicted labels and ground-truth outcomes into true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Within each group, relevance scores are aggregated across samples to obtain average feature-level attribution values. Features are then ranked according to the magnitude of their aggregated relevance scores, and the eight most influential features for each outcome group are identified and visualized.

This outcome-aware aggregation enables direct comparison of dominant feature contributions across correct and incorrect predictions. By focusing on high-magnitude attribution signals, the analysis highlights the structural drivers of model decisions and reveals systematic patterns associated with misclassification.

4 Results

4.1 Predictive Performance

We first evaluate the predictive performance of the proposed self-attention-based integrity prediction model on a held-out test set. As shown in Figure 2, the model achieves strong overall performance, with an AUC-ROC close to 0.90 and an accuracy exceeding 0.84. These results indicate substantial aggregate discriminative ability on large-scale longitudinal financial data, suggesting that the model effectively separates positive and negative classes at the population level.

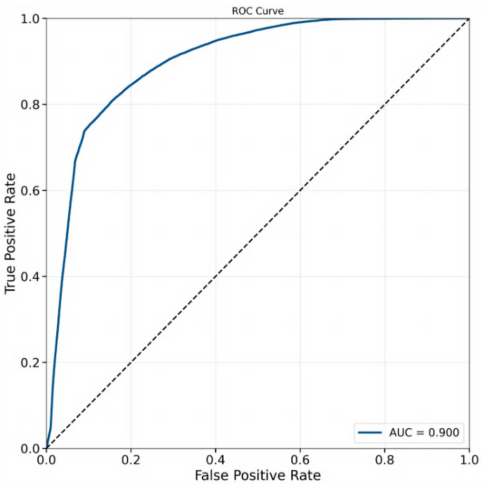


Figure 2: ROC Curve (AUC = 0.90)

However, a notable gap emerges between AUC-ROC and AUC-PR, with the latter reaching 0.47. This difference indicates that, although overall ranking performance is strong, the model's ability to precisely identify positive cases is comparatively more limited. This discrepancy is consistent with class imbalance and the inherent difficulty of identifying relatively rare positive outcomes. In particular, performance near the decision threshold suggests an increased presence of false positives and false negatives. As a result, aggregate performance metrics alone are insufficient to fully characterise model behaviour, motivating a more detailed examination of how individual features contribute to both correct and incorrect predictions across different outcome categories.

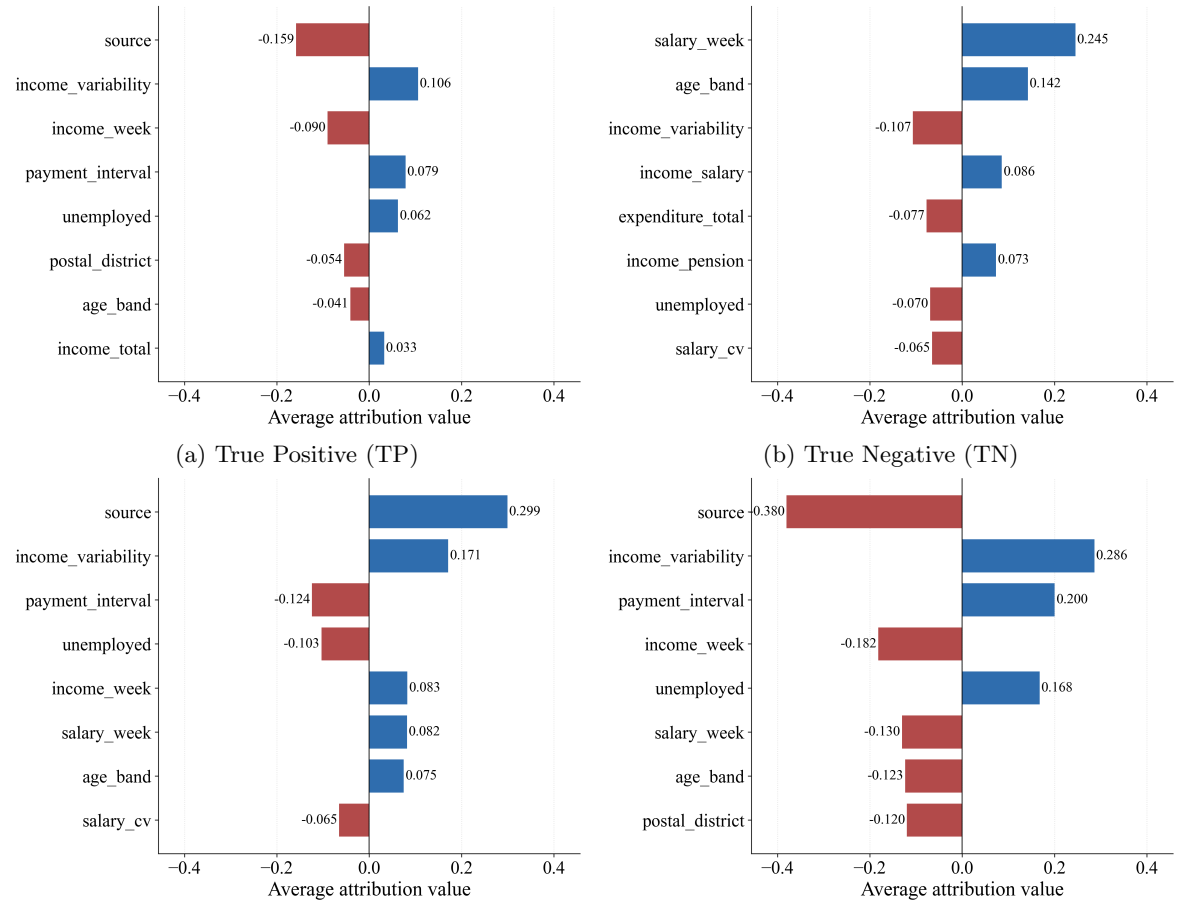
4.2 Explainability Analysis Across Prediction Outcomes

Motivated by the limitations of aggregate performance metrics discussed above, this section investigates the internal decision logic of the proposed model through outcome-aware feature attribution analysis grounded in individual-level weekly financial data. Using aggregated relevance scores computed over more than one million weekly observations, we examine how different categories of input features contribute to predictions across true positive, true negative, false positive, and false negative outcomes. For clarity and interpretability, Figure 3 visualises the top eight features with the largest absolute average attribution values for each outcome group and Table 2 summarises the key findings, which are discussed in the following paragraphs. This visualisation strategy highlights the most influential drivers of the model's decisions while suppressing low-impact features that contribute minimally at the aggregate level.

Table 2: Summary of Outcome-Aware Attribution Analysis

Decision Mechanism	**Observed Attribution Pattern**	**Structural Implication**
Asymmetric Risk Activation	**TP:** High concentration on specific features (e.g., income variability). **TN:** Uniformly weak attribution across all top features.	The model functions as a risk detector, not a balanced evaluator. Low risk is defined by the *absence* of instability signals, not the presence of protective factors.
Volatility Dominance	**High Impact:** Temporal instability (income variability, payment intervals, unemployment). **Low Impact:** Static demographics (age, sex) and absolute income levels.	Vulnerability is represented as *dynamic instability* rather than static socioeconomic status. Consumption is only relevant when aligned with volatility.
Error Aggregation	**FP:** Driven by narrow, isolated spikes in volatility features. **FN:** Characterised by diffuse, low-intensity signals across multiple features.	Misclassification arises from signal integration limits: the model over-reacts to isolated noise (FP) and fails to aggregate distributed weak evidence (FN).

Rather than interpreting outcome-specific attribution profiles in isolation, we synthesise evidence across outcomes to identify stable structural patterns in how the model activates, suppresses, and aggregates



(a) True Positive (TP)

(b) True Negative (TN)

(c) False Positive (FP)

(d) False Negative (FN)

Figure 3: Top feature attribution profiles across prediction outcomes.

risk-related information. The resulting attribution patterns are consistent with a predominantly risk-activation-based decision structure, in which high-risk predictions are associated with concentrated instability signals, whereas low-risk predictions typically occur when such signals remain below implicit activation thresholds. Importantly, these observations should be interpreted as structural regularities in attribution behaviour rather than definitive evidence of a single underlying mechanism. In addition, misclassification patterns appear to reflect systematic limitations in evidence aggregation and thresholding under sparse positive labels, rather than purely random variation.

Finding A: High-risk predictions are driven by asymmetric risk activation rather than balanced evidence evaluation.

The attribution patterns indicate that the model does not symmetrically evaluate evidence for positive and negative classes at the level of weekly financial snapshots. In true positive cases, Figure 3(a) shows that attribution mass is highly concentrated in a small subset of features, with income variability, payment interval irregularity, unemployment status, and income source dominating the relevance rankings. In contrast, true negative predictions in Figure 3(b) display uniformly weak attribution magnitudes across the top-ranked features, despite the visualisation focusing on the eight most influential variables. This contrast indicates that correct low-risk decisions do not rely on strong counteracting or protective signals, but instead emerge when no weekly instability indicator exceeds the model's implicit activation threshold. A closely related attribution structure appears in false negative cases shown in Figure 3(d), where benefit receipt occurs within the week but risk-related signals remain fragmented and individually weak. Together, these patterns are consistent with a risk-activation-style decision structure rather than a balanced accumulation of evidence across classes.

Finding B: Financial and employment volatility emerge as the dominant dimensions in the model's representation of vulnerability.

Across all outcomes involving high-risk predictions, features capturing short- and medium-term financial volatility consistently dominate the attribution rankings. As shown in Figure 3(a) and Figure 3(c), variables such as income variability, salary coefficient of variation, payment interval, unemployment status, and income source receive the largest absolute attribution values whenever the model assigns a high-risk label, regardless of prediction correctness. These features are explicitly constructed to capture temporal instability within weekly aggregation windows rather than absolute income levels. In contrast, static demographic attributes including age band, sex, and postal district, as well as level-based income measures, contribute minimally across all outcome groups and rarely appear among the top-ranked features. Expenditure-related variables exhibit moderate influence only when aligned with volatility signals, indicating that consumption behavior is interpreted conditionally rather than as an independent driver of vulnerability. This consistent dominance of volatility features suggests that the model's learned representation of vulnerability is more closely aligned with dynamic instability than static socioeconomic status.

Finding C: Model errors are consistent with limitations in evidence aggregation under sparse positive labels.

The attribution analysis further shows that model misclassifications are consistent with predictable limitations in evidence aggregation shaped by the structure of the data and the rarity of positive labels. False positive predictions in Figure 3(c) are characterized by strong attribution concentrated on a narrow subset of volatility-related features, particularly income variability and payment interval irregularity, while other financial dimensions contribute weakly or remain neutral. This pattern suggests that isolated weekly fluctuations may disproportionately influence predictions even when broader financial stability cues are present. Conversely, false negative cases in Figure 3(d) display diffuse attribution patterns across multiple features, none of which individually reach the magnitude typically observed in true positive cases. This implies that certain benefit-receiving individuals exhibit vulnerability profiles expressed through multiple low-intensity signals rather than a single dominant instability dimension. In both error types, the model appears more responsive to concentrated signals than to distributed weak evidence. These complementary patterns highlight potential limitations in how heterogeneous and weakly expressed signals are integrated under a fixed decision threshold in large-scale weekly financial data.

5 Limitations

Despite the strengths of the proposed approach, several limitations should be acknowledged. Each weekly observation is treated as an independent sample, and the model does not explicitly capture temporal dependencies across time windows, which may limit its ability to represent long-term behavioural trends or delayed effects. In addition, the prediction target used in this study serves as an observable proxy for financial vulnerability rather than a direct measure of individual integrity, thereby constraining the scope of interpretation of the results. Furthermore, while the feature-level explainability analysis provides insight into the model's decision logic under different prediction outcomes, it describes how the model utilises input features rather than establishing causal relationships between features and outcomes. As a result, the attribution results should be interpreted as descriptive rather than causal. Future work may address these limitations by incorporating temporal modelling and complementary causal analysis methods to improve the characterisation of complex and evolving risk patterns.

6 Conclusions

This study presents an explainable integrity prediction framework based on a self-attention neural architecture applied to large-scale longitudinal financial data. The proposed approach achieves strong predictive performance while using outcome-aware feature attribution to systematically characterise model behaviour across true positive, true negative, false positive, and false negative predictions, moving beyond evaluations based solely on aggregate performance metrics. The results show that model decisions are primarily driven by dynamic indicators of financial and employment instability, while static demographic attributes play a comparatively limited role; moreover, analysis of misclassified cases indicates that prediction errors arise from structural limitations in evidence aggregation and thresholding rather

than random variation. Overall, this work demonstrates the value of outcome-aware explainability for transparent evaluation of high-stakes integrity prediction systems and provides a foundation for future research on robustness, fairness, and policy-oriented model auditing.

Acknowledgments

This research was funded through a University of Edinburgh Major Initiative Fund grant held by the Edinburgh Centre for Financial Innovations. The data and computing facilities were provided by NatWest, Smart Data Foundry, and Edinburgh Parallel Computing Centre.

References

[1] Financial Data Service — sdruk.ukri.org. https://www.sdruk.ukri.org/data/financial-data-service/. [Accessed 17-02-2026].

[2] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 233–240, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143874. URL https://doi.org/10.1145/1143844.1143874.

[3] Shimin Lei, Ke Xu, YiZhe Huang, and Xinye Sha. An xgboost based system for financial fraud detection. *E3S Web of Conferences*, 214:02042, 01 2020.

[4] Weixian Waylon Li and Tiejun Ma. Learn to rank risky investors: A case study of predicting retail traders' behaviour and profitability. *ACM Trans. Inf. Syst.*, 44(1), November 2025. ISSN 1046-8188. doi: 10.1145/3768623. URL https://doi.org/10.1145/3768623.

[5] Xingqiao Li, Jindong Gu, Zhiyong Wang, Yancheng Yuan, Fengxiang He, and Bo Du. XAI for In-Hospital Mortality Prediction via Multimodal ICU Data . In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3798–3801, Los Alamitos, CA, USA, December 2025. IEEE Computer Society. doi: 10.1109/BIBM66473.2025.11356415. URL https://doi.ieeecomputersociety.org/10.1109/BIBM66473.2025.11356415.

[6] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 469–481, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372828. URL https://doi.org/10.1145/3351095.3372828.

[7] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[9] Tianjiao Zhang, Weidong Zhu, Yong Wu, Zihao Wu, Chao Zhang, and Xue Hu. An explainable financial risk early warning model based on the ds-xgboost model. *Finance Research Letters*, 56: 104045, 2023. ISSN 1544-6123.