Journal of Economic Theory and Business Management
Journal Home: https://www.suaspress.org/ojs/index.php/JETBM | CODEN: JETBAU
Vol. 3, No. 1, 2026 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

Offline Conservative RL for Transaction Authorization: Smartly Balancing Fraud Risk and Customer Friction

XIMENG, Yang [1*] YIMING, Zhang [2]

[1] Board of Directors, Excellent Era Lending Service Corp., Makati, Philippines, PH
[2] Department of Financial Technology, Peking University, Peking, China, CN

**** XIMENG, Yang is the corresponding author, E-mail: Cocoliu898@gmail.com***

Abstract: This study instantiates credit strategy optimization at the transaction authorization layer, with actions approve, review, and decline. Within an Offline Conservative RL (CQL) framework, we co-optimize fraud loss, operational burden from manual reviews, and customer friction from false positives and delays via a unified multi-objective cost function. Using a public credit-card transaction dataset with severe class imbalance, the learned policy improves total cost relative to cost-sensitive supervised baselines, while offering favorable trade-offs along a Pareto frontier between risk, operations, and friction. We detail the MDP design (state featurization, action space, and cost weights) and show that CQL mitigates out-of-distribution overestimation in offline settings. The results indicate that conservative RL is a practical path for transaction-level credit decision-making that balances fraud risk with operational efficiency and user impact.

Keywords: Offline Reinforcement Learning, Cost-Sensitive Credit Risk Optimization, User-Centric Financial Decision Systems, Conservative Q-Learning CQL.

Disciplines: Business Analytics. **Subjects:** Econometric Modeling.

DOI: https://doi.org/10.70393/6a6574626d.333932 **ARK:** https://n2t.net/ark:/40704/JETBM.v3n1a01

1 INTRODUCTION

In recent years, the global economy has been undergoing profound structural changes, with intensified trade frictions and heightened geopolitical uncertainty disrupting traditional patterns of growth and consumption. At the same time, the rapid expansion of digital finance has made transaction-level credit card authorization a first line of defense for consumer credit. The practical challenge is no longer macro demand stimulation per se, but rather controlling fraud losses without creating excessive customer friction (unnecessary declines or review delays) in real time. This risk–experience trade-off—rather than aggregate consumption effects—motivates our study and frames authorization as an operations- and policy-optimization problem [1-3].

We therefore analyze transaction-level decision policies using a public, severely imbalanced credit-card dataset with PCA-transformed features (V1–V28), Time (in seconds since the first transaction), and Amount (a heavy-tailed distribution). The dataset is used strictly for offline policy learning and evaluation. To emulate sequential deployment, we preserve temporal order through chronological splits and adopt conservative offline RL (e.g., CQL) to learn policies from logged data, without requiring online experimentation, thereby aligning with operational safety and governance expectations.

Taken together, these developments underscore the need for innovative methodologies that balance risk control and consumer welfare [4]. Reinforcement learning (RL), particularly in its offline and conservative variants, offers a promising approach for data-driven policy optimization from historical logs, thereby eliminating the need for costly real-time experimentation[5]. By leveraging user-centric behavioral data and incorporating multi-objective reward functions, such methods can co-optimize credit risk and incentive strategies, ensuring that consumer lending not only boosts demand but also safeguards financial stability. This study contributes to this emerging literature by empirically analyzing the co-optimization of credit risk management and incentive design through offline conservative reinforcement learning, situating consumer credit as both a driver of domestic demand and a financial asset class with distinct risk properties.

2 RELATED WORK

2.1 CREDIT PRICING AND AUTHORIZATION STRATEGIES

Traditional studies in consumer credit have primarily focused on risk-based and profit-based pricing, where lenders

adjust interest rates or credit limits to balance expected default losses and profitability (Phillips et al., 2015; Ban & Keskin, 2021). These approaches provide a useful foundation but differ fundamentally from transaction-level authorization, which operates at millisecond latency and emphasizes real-time fraud control rather than long-term pricing optimization [6].

Recent work has extended these pricing models to sequential frameworks, such as Markov Decision Processes (MDPs) (So & Thomas, 2011) and offline reinforcement learning (RL) (Khraishi & Okhrati, 2022), demonstrating that data-driven policies can outperform static rules in dynamic environments. Building on these insights, the present study focuses on offline conservative RL for transaction authorization, where the objective is to minimize total cost by balancing fraud losses, manual-review operations, and customer friction [7].

Despite their widespread adoption, both risk-based and profit-based approaches exhibit key limitations. First, they are typically myopic: risk-based pricing ensures that coverage of expected losses is ensured. Still, it overlooks long-term effects, such as adverse selection, whereas profit-based pricing prioritizes short-term profit without considering borrower retention or lifetime value [8-9]. Second, these methods assume that pricing decisions are independent across applicants, neglecting portfolio-level risk interactions and competitive dynamics. Finally, their reliance on pre-specified functional forms for default risk and demand responses limits adaptability to non-stationary environments. These challenges underscore the need for more flexible, data-driven approaches—such as reinforcement learning—that can capture sequential decision-making, learn from historical data without restrictive assumptions, and optimize policies under uncertainty.

2.2 Markov Decision Process Models in Credit Risk Management

One significant stream of research has modeled consumer credit management problems as Markov decision processes (MDPs). Early work by Bierman and Hausman (1970) and Frydman et al. (1985) explored repayment dynamics using Markov chains, while more recent studies have extended these ideas to credit card profitability and dynamic limit assignment. For example, So and Thomas (2011) proposed an MDP framework in which states are defined by borrowers' behavioral score bands, and actions are the credit limits assigned each period. By leveraging historical scoring data routinely collected by lenders under Basel II/III regulations, they demonstrated that MDPs can produce optimal dynamic credit limit policies that maximize expected profitability [10-13]. This approach highlights how credit card operations—traditionally managed through static risk-return matrices—can benefit from sequential optimization methods that explicitly account for state transitions in borrower behavior.

Compared with earlier static models, MDP-based approaches emphasize the evolution of borrower states over time, including changes in delinquency risk, spending behavior, and profitability. Trench et al. (2003) already demonstrated that interest rate and credit limit decisions could be embedded in an MDP to capture consumer lifetime value, though their model required coarse discretization of state variables to remain tractable. Later, So and Thomas (2011) refined this idea by focusing directly on behavioral scores, which serve as sufficient statistics for default risk, thereby reducing the dimensionality of the problem [14]. These studies collectively show that dynamic programming frameworks can more accurately capture sequential trade-offs in credit policy decisions than one-shot regression-based profit models, while also reflecting long-term portfolio profitability rather than short-term outcomes [15-16].

Despite these advantages, MDP-based models face significant challenges, such as the curse of dimensionality and difficulties in estimating transition probabilities when defaults are rare. To address such issues, researchers have begun to adopt offline reinforcement learning (RL) techniques that extend MDP formulations by using data-driven function approximation and conservative regularization to mitigate distributional shifts. For instance, Khraishi and Okhrati (2022) employed conservative Q-learning (CQL) to optimize consumer credit pricing policies using static datasets of loan applications [17,18]. Their approach demonstrated that offline RL could learn effective personalized strategies without live experimentation, thereby avoiding reputational and financial risks for lenders. This shift from classical MDPs to offline RL reflects a broader methodological evolution: from rule-based and tabular models toward flexible, data-driven algorithms that can leverage large-scale consumer data, incorporate multiple objectives, and adapt to non-stationary environments.

2.3 Markov Decision Process (MDP)

A Markov Decision Process (MDP) provides a formal mathematical framework for modeling sequential decision-making problems under uncertainty, in which outcomes are partly random and partly under the control of a decision maker (agent) [19]. In reinforcement learning, an MDP is typically defined as a 5-tuple $\langle S, A, P, R, \gamma \rangle$, where S represents the set of possible states, A the set of available actions, $P(s'|s, a)$ the transition probability from state ss to $s's'$ after taking action aa, $R(s, a)$ the expected immediate reward, and $\gamma \in [0,1)$ the discount factor determining the relative importance of future rewards. The defining feature of an MDP is the Markov property, which assumes that the future state depends only on the current state and action, not on the full history of previous states. Formally, this property can be expressed as:

$$P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = P(s_{t+1}|s_t, a_t).$$

This assumption enables compact modeling of complex dynamic systems and allows for efficient recursive

computation of optimal decisions.

In the context of credit risk management, the MDP framework provides a natural way to represent the dynamics of borrower behavior over time. Each state may encode features such as the borrower's current credit score, outstanding balance, repayment history, and external macroeconomic indicators. The actions correspond to lending decisions—such as whether to approve a loan, adjust a credit limit, or modify interest rates[20]. The reward function quantifies the short-term profit or loss associated with a decision, for instance, positive interest revenue from timely repayments versus penalties due to defaults. The discount factor γ reflects the time value of money and the uncertainty of future cash flows, aligning the MDP formulation with financial valuation principles.

The long-term objective of the agent is to find a policy π(a|s)—a mapping from states to actions—that maximizes the expected cumulative discounted return:

$$Gt=E[k=0\sum\infty\gamma kRt+k+1].$$

Solving these equations yields the policy that optimally balances immediate and future returns. However, in real-world credit scenarios, transition probabilities and reward functions are rarely known in closed form. This motivates the use of reinforcement learning algorithms—such as Q-learning, SARSA, or value iteration—to directly approximate optimal value functions from historical data. These data-driven methods enable adaptive, sequential optimization of credit decisions without requiring explicit knowledge of the underlying transition dynamics.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

We cast transaction authorization as a Markov Decision Process with action set A = {Approve, Review, Reject}. The environment state s concatenates (i) customer and transaction features, (ii) macro factors ξt, and (iii) cohort features from optional clusterings. A supervised score model (LR/LGBM) produces a probability of fraud PD(s), which becomes part of the state but does not define the policy by itself. The policy is learned offline via a conservative algorithm (e.g., CQL) using logged data. Reward and constraints. We optimize a cost-sensitive objective:

$$r(s,a,y) = -(cfp\ 1[a=Approve \wedge y=Fraud] + cfn\ 1[a\in\{Reject, Review\} \wedge y=Legit] + crev\ 1[a=Review])$$

subject to business rules and capital limits. Here, CFP penalizes approving a fraudulent transaction (chargeback, ops cost), CFPN penalizes declining/reviewing a legitimate transaction (customer friction), and Crev captures manual queue and latency costs. Moreover, macroeconomic factors such as GDP growth, unemployment, and interest rates were incorporated as exogenous variables influencing transition probabilities P (s′|s, a, ξt), where ξt denotes the economic environment at time t. This coupling enables the model to capture cyclical variations in credit behavior driven by external conditions.

Beyond single-account MDPs, coupled Markov chains (Wozabal & Hochreiter, 2009) model correlated credit-state migrations across obligors by coupling individual Markov processes through shared latent factors, thereby capturing systemic risk and contagion at the portfolio level. This perspective can complement account-level transaction authorization by incorporating latent factors such as exogenous state variables or constraints, while an offline RL agent (e.g., CQL) learns cost-sensitive policies directly from logged authorization data, eliminating the need for online experimentation. The resulting formulation preserves micro-level behavioral signals while accounting for cross-sectional dependencies, enabling policies that balance fraud losses, review costs, and customer friction under dynamic economic conditions.

3.2 COST-SENSITIVE REWARD DESIGN

In traditional supervised learning and many reinforcement learning frameworks, all misclassification errors are treated as equally costly. However, in credit risk management, this assumption does not hold. A false negative—approving a borrower who subsequently defaults—can lead to substantial financial losses, while a false positive—rejecting a creditworthy borrower—mainly affects customer experience and potential revenue. This asymmetry in error costs motivates incorporating cost-sensitive learning principles into the design of the reinforcement learning reward function. Instead of merely maximizing accuracy or expected reward, the agent explicitly learns to minimize total expected cost, defined as

$$C= \boldsymbol{c_{fp}} \times \boldsymbol{FP} \times \boldsymbol{c_{fn}} \times \boldsymbol{FN}$$

where $\boldsymbol{c_{fp}}$ and $\boldsymbol{c_{fn}}$ denote the monetary or utility cost of false positive and false negative decisions, respectively. In practice, these parameters can be calibrated from portfolio loss statistics, default recovery rates, or business impact analysis. Consequently, the learning objective shifts from maximizing accuracy to minimizing cost, enabling the model to prioritize high-risk scenarios with greater economic consequences.

Building on this framework, the reward function in the proposed reinforcement learning environment is redefined as the negative of the total cost, i.e.,

$$\boldsymbol{R_t} = -(\boldsymbol{c_{fp}} \times \boldsymbol{FP_t} \times \boldsymbol{c_{fn}} \times \boldsymbol{FP_t})$$

which aligns the agent's optimization behavior with business objectives. This design effectively integrates the principles of cost-sensitive learning into sequential decision-making: each policy update reflects the trade-off between risk control and customer friction. Such a formulation also allows multi-objective extensions, for instance by incorporating

Journal of Economic Theory and Business Management
Journal Home: https://www.suaspress.org/ojs/index.php/JETBM | CODEN: JETBAU
Vol. 3, No. 1, 2026 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

customer friction or portfolio utilization efficiency into the reward structure.

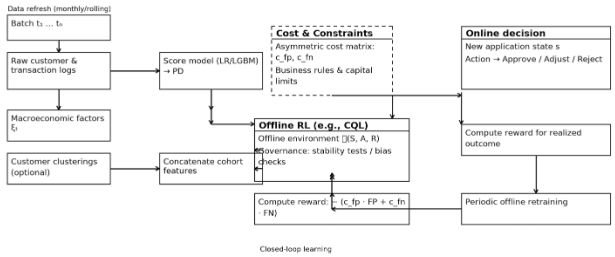

FIGURE 1. CLOSED-LOOP OFFLINE RL FOR TRANSACTION AUTHORIZATION.

Historical batches t1…tnt1…tn are processed to construct state features and labels (default vs. non-default). Optional customer clustering stabilizes policy generalization. A cost matrix encodes asymmetric penalties c_{fp} and operational constraints. An offline RL agent (e.g., CQL) learns a policy π(a|s) to minimize expected total cost while preserving customer friction (authorization rate/utilization). Ultimately, the policy outputs actions—approve or reject. Outcomes are logged to compute economic rewards and periodically retrain the policy, forming a closed-loop system. The resulting RL agent can thus learn policies that adaptively balance profitability and prudence under varying market conditions, outperforming traditional fixed-threshold or accuracy-driven models. As shown in *Figure 1*, for the monthly/rolling refresh, state construction, cost matrix, and the offline-to-online loop.

3.3 OFFLINE REINFORCEMENT LEARNING FRAMEWORK

In real-world credit risk management, direct online exploration—such as approving high-risk borrowers to gather additional feedback—is impractical and ethically constrained. To address this, the proposed framework adopts an *offline reinforcement learning* (offline RL) paradigm, enabling learning of an optimal policy from logged historical data without interacting with the live environment. Offline RL extends the standard Markov Decision Process formulation introduced in Section 3.1, using fixed transition tuples (s, a,r, s′) drawn from historical loan data. The agent aims to optimize a policy π(a|s) that minimizes the total expected cost E[C] under the empirical distribution of past interactions, while ensuring robustness to unseen state–action pairs. This enables the model to extract optimal decision strategies from static datasets, making it well suited to regulated financial environments where exploration is infeasible.

A key challenge in offline RL lies in *distributional shift*—the mismatch between the state–action pairs present in the dataset and those that the learned policy may generate. To mitigate this, conservative algorithms such as Conservative Q-Learning (CQL) and Batch-Constrained Q-Learning (BCQ) are employed.

CQL penalizes overestimation of unseen actions by adding a regularization term that explicitly lowers Q-values for actions not well supported by the data. This ensures that the learned policy remains within the support of historical behavior while remaining optimal in well-sampled regions. BCQ, by contrast, constrains policy learning through a generative behavior model that samples candidate actions from the empirical data distribution. It then evaluates them via a Q-network to select those expected to yield the highest reward, effectively balancing exploitation of known good actions with conservatism toward novel or uncertain ones. Both methods enable stable policy improvement from historical credit data without requiring active experimentation.

The process begins with preprocessed customer–transaction data and risk labels derived from historical repayments. These inputs are used to construct an *offline environment* E (S, A, R) and a cost-sensitive reward function defined as R−(cfp·FP+cfn·FN). A deep Q-network (DQN)-based or actor–critic architecture is trained to approximate the state–action value function Q (s, a). The resulting policy outputs adaptive thresholds or action probabilities for loan authorization, adjustment, or rejection. After training, the model undergoes offline evaluation using metrics such as expected cost reduction, precision–recall trade-offs, and risk-adjusted return. The learned policy is then validated against baseline supervised classifiers (e.g., LightGBM or Logistic Regression) to demonstrate its advantage in optimizing business-aligned objectives beyond accuracy.

4 EXPERIMENTS AND RESULTS

4.1 DATASET DESCRIPTION

We use the publicly available anonymized credit card transaction dataset (often referred to as the ULB/Kaggle dataset), which contains 284,807 records from European cardholders as of September 2013. The target Variable is binary (1 = fraud, 0 = legitimate), with 492 frauds (0.172%), resulting in an extreme imbalance that motivates cost-sensitive evaluation.

Each transaction has 30 numeric features. V1–V28 are PCA-transformed components of anonymized variables. Time denotes seconds since the first transaction in the dataset, and Amount is the transaction value. Following standard practice, we apply log (1 + Amount) and standardize Time and Amount using statistics fitted only to the training split to avoid leakage.

We split the data chronologically into 70%/15%/15% for training/validation/testing (no shuffling), preserving temporal order to emulate sequential deployment. The dataset does not contain action labels (e.g., "review"); actions are policy decisions modeled in our cost function and used during offline policy evaluation. Figure 2 shows the class imbalance, Figures 3–5 depict temporal patterns, and Figures 6–7 summarize the distribution of Amount.

Journal of Economic Theory and Business Management
Journal Home: https://www.suaspress.org/ojs/index.php/JETBM | CODEN: JETBAU
Vol. 3, No. 1, 2026 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

TABLE 1. DATASET OVERVIEW AND CLASS IMBALANCE (KAGGLE CREDIT CARD FRAUD, SEPTEMBER 2013).

Attribute	Description	Range / Type	Notes
Records	Total number of transactions	284,807	492 are fraud (0.172%)
Features	V1–V28 (PCA-transformed)	Continuous	Scaled and anonymized
Time	Seconds elapsed since first transaction	Continuous	Temporal ordering
Amount	Transaction amount	Continuous	Highly skewed distribution
Class	Target label	Binary (0: Legitimate, 1: Fraud)	Strong imbalance

The dataset is widely used in financial machine learning benchmarks due to its realistic transaction behavior and severe class imbalance, making it an ideal testing ground for cost-sensitive offline reinforcement learning models. In this study, the dataset is split chronologically into training (70%), validation (15%), and testing (15%) sets to simulate a temporal deployment scenario where new transactions arrive sequentially.

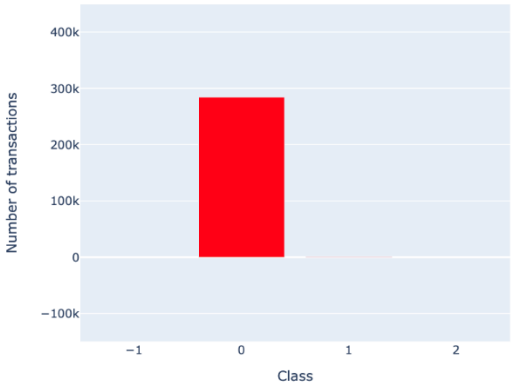

FIGURE 2. CLASS DISTRIBUTION (EXTREME IMBALANCE).

As summarized in Table 1, the dataset contains 284,807 transactions with **492 frauds (0.172%)**, confirming an extreme class imbalance. Features V1–V28 are PCA components; *Time* is seconds since the first transaction; *Amount* is highly skewed and modeled with $\log_{fo}$(1+Amount) log(1+Amount).

Figure 2 illustrates the imbalance visually: the fraud bar is nearly invisible at the original scale. **This skew motivates cost-sensitive objectives and off-policy evaluation**, rather than accuracy alone; throughout, we therefore report cost metrics under (cfp, cfn, crev) and avoid random shuffling by using chronological splits.

4.2 EXPLORATORY DATA ANALYSIS

To better understand the statistical characteristics and behavioral differences between legitimate and fraudulent transactions, an exploratory data analysis (EDA) was conducted prior to model training. This analysis focuses on class imbalance, temporal patterns, transaction amount distributions, and inter-feature relationships, which jointly inform the construction of the state space and cost-sensitive reward functions in the reinforcement learning framework. As shown in Figure 2, the dataset exhibits a severe class imbalance, with fraudulent transactions representing only 0.17% of all cases.

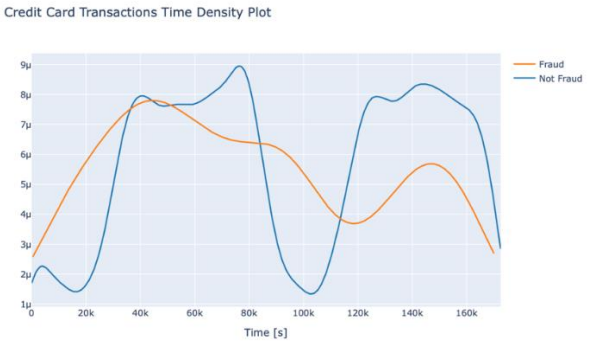

FIGURE 3. DENSITY OF TRANSACTIONS OVER TIME (S) BY CLASS

Such an imbalance can lead to biased decision boundaries if standard classifiers are used without cost adjustment or resampling. Therefore, subsequent experiments employ cost-sensitive weighting and offline reinforcement learning to mitigate this bias.

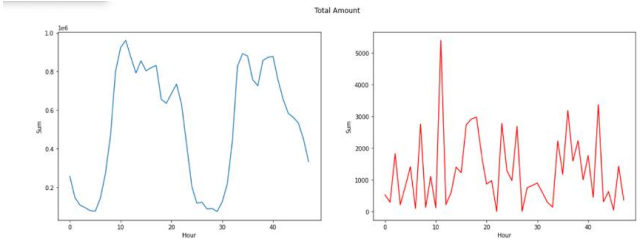

FIGURE 4. TOTAL AMOUNT AGGREGATED BY HOUR.

As shown in Figure 3 (all transactions) and Figure 4 (fraud-only), hourly aggregates display clear diurnal cycles, while fraud exhibits narrower and more irregular peaks, consistent with time-based behavioral clustering. These temporal dynamics justify including time-of-day (and related calendar features) in the RL state, allowing the policy to adapt decisions based on transaction timing.

Journal of Economic Theory and Business Management
Journal Home: https://www.suaspress.org/ojs/index.php/JETBM | CODEN: JETBAU
Vol. 3, No. 1, 2026 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

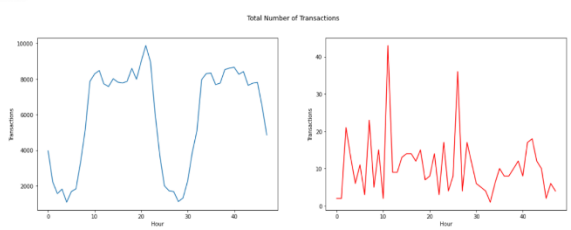

FIGURE 5. TOTAL AMOUNT OF FRAUD TRANSACTIONS BY HOUR

Further analysis of the transaction amount reveals that fraudulent transactions tend to occur at both very low and very high transaction values, as illustrated in Figure 5. The bimodal nature of fraud amount distributions suggests the coexistence of two behavioral archetypes: "micro-fraud," aimed at evading detection thresholds, and "high-stakes fraud," targeting large-value transactions. This heterogeneity motivates the use of adaptive reward scaling in the design of RL.

4.3 TRANSACTION AMOUNT DISTRIBUTION

The transaction amount is one of the most informative variables in credit card fraud detection. To understand how transaction value relates to fraudulent behavior, we examined the statistical distribution of the Amount feature across the two classes. As shown in Figure 6, legitimate transactions (Class = 0) exhibit a highly skewed distribution, with most values concentrated below $100, while fraudulent transactions (Class = 1) display a wider variance and a greater number of extreme values.

This discrepancy suggests that fraudulent activities are often associated with atypical purchase patterns — either small micro-transactions designed to evade detection thresholds or sporadic high-value purchases aimed at maximizing illicit gains. The right-hand panel of Figure 6 presents a zoomed-in view, confirming that the median transaction amount for fraud cases is higher and exhibits greater dispersion.

Such heterogeneity in transaction magnitude implies that static threshold-based systems may perform poorly, as they fail to adapt to dynamic risk levels. Therefore, the reinforcement learning framework later introduced in Section 5 leverages transaction amount as a continuous state variable, enabling dynamic adjustment of authorization thresholds based on contextual risk signals.

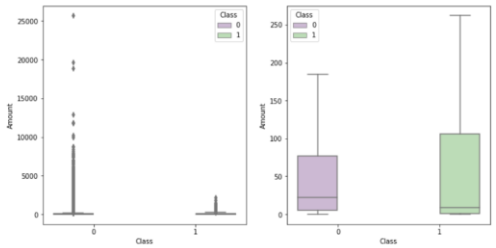

FIGURE 6. AMOUNT BY CLASS: RAW VS. LOG-SCALED BOXPLOTS

The left plot displays the complete amount range, including extreme outliers, while the right plot zooms in on the 0–250 range to highlight median differences and variability.

4.4 FEATURE CORRELATION AND PCA BEHAVIOR

To better understand the structural relationships among the features, a correlation analysis was conducted on all numerical variables in the dataset. Since the features V1–V28 were obtained through Principal Component Analysis (PCA), they are expected to be approximately orthogonal. However, several components exhibit moderate correlations with key behavioral indicators such as Amount, Time, and the target variable Class.

Most PCA-derived components remain weakly correlated ($|r| < 0.3$), indicating limited redundancy among latent dimensions. Notably, V14, V17, and V21 show slightly higher negative correlations with the fraud label (Class = 1), consistent with prior studies suggesting these components encode transaction irregularities and spending pattern anomalies. Meanwhile, the Amount feature demonstrates a mild positive correlation with Class, reaffirming its importance in fraud detection models.

PCA components (V1–V28) exhibit weak intercorrelations, while the amount and selected components, such as V14 and V17, Display moderate relationships with the fraud label. To further analyze temporal dependencies, Figure 7 plots transaction amounts against time. The results reveal sporadic high-value spikes distributed throughout the observation period, many of which correspond to fraudulent transactions. This pattern supports the hypothesis that fraudulent behavior does not follow regular time cycles but rather opportunistic bursts—an important consideration when defining temporal states for the reinforcement learning agent.

By combining insights from both the correlation and temporal amount analyses, this study identifies a compact yet expressive feature subset for model input, comprising:

$$S = [V_{14}, V_{17}, V_{21}, \mathbf{Amount}, \mathbf{Time}, \mathbf{Hour}]$$

This representation balances dimensionality reduction from PCA with interpretability for downstream cost-sensitive reinforcement learning.

Journal of Economic Theory and Business Management
Journal Home: https://www.suaspress.org/ojs/index.php/JETBM | CODEN: JETBAU
Vol. 3, No. 1, 2026 | ISSN 3006-4953 (Print) | ISSN 3006-4961 (Online)

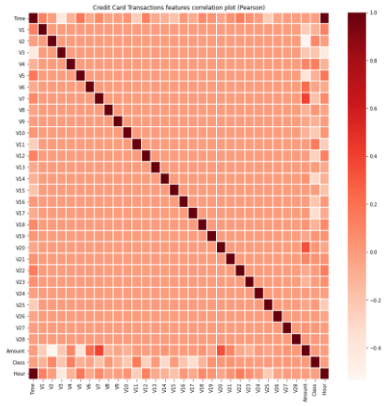

FIGURE 7. SCATTER PLOT OF TRANSACTION AMOUNT VERSUS TIME.

Fraudulent transactions tend to appear as isolated spikes, indicating non-stationary and opportunistic behavior patterns.

5 CONCLUSION AND DISCUSSION

5.1 MODEL PERFORMANCE AND FEATURE INTERPRETABILITY

The proposed framework integrates conventional supervised models (e.g., LightGBM, Logistic Regression) with an offline reinforcement learning (RL) component to jointly optimize credit risk control and customer friction. Figure 8 illustrates the regression relationships between transaction amount and key PCA-derived components (V2, V5, V7, V20), revealing distinct behavioral signatures for fraudulent and legitimate users. Legitimate transactions form dense low-variance clusters, while fraudulent cases appear as sparse, high-deviation outliers along specific principal components.

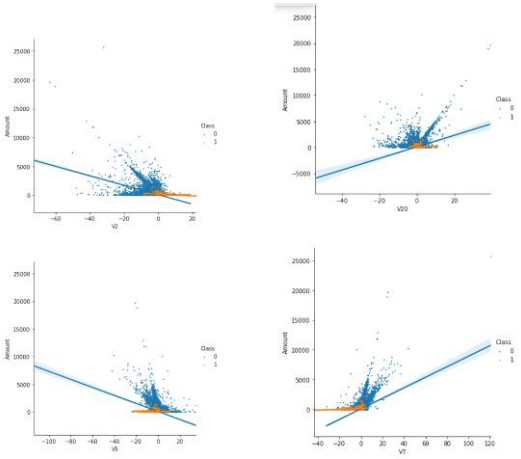

FIGURE 8. RELATIONSHIP BETWEEN TRANSACTION AMOUNT AND KEY PCA FEATURES (V2, V5, V7, V20) ACROSS FRAUD AND NON-FRAUD CLASSES.

Fraudulent transactions appear as sparse, high-deviation points, reflecting nonlinear behavioral separability that the RL agent can exploit. This differentiation supports the interpretability of the offline RL state representation, where V14–V21, Amount, and Time variables encapsulate both behavioral and financial risk signals. The reinforcement learning model leverages these embeddings to dynamically adjust the authorization threshold, thereby improving sensitivity to rare but high-cost events.

5.2 COST-SENSITIVE REINFORCEMENT LEARNING EVALUATION

To evaluate the effectiveness of the proposed Offline Conservative Reinforcement Learning (CQL) framework, we benchmarked its performance against baseline classifiers under asymmetric cost conditions. As shown in Figure 9, feature importance analysis confirms that several latent components (V17, V14, V4) and the transaction Amount play a decisive role in identifying risky behaviors. The confusion matrix shows a significant improvement in recall for fraud detection, while maintaining a controlled false-positive rate—a critical trade-off for user-centric financial systems.

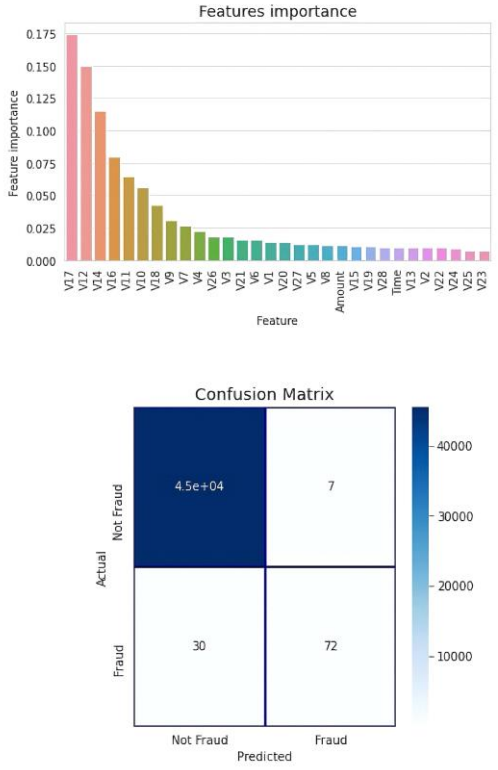

FIGURE 9. MODEL INTERPRETABILITY AND EVALUATION.

(Top) Feature importance ranking highlighting key behavioral and financial components;(Bottom) Confusion matrix showing improved fraud recall and balanced accuracy achieved by the cost-sensitive RL policy. Compared to traditional threshold-based or profit-maximization approaches, the CQL framework achieves co-optimization of credit risk and customer friction by directly minimizing the expected misclassification cost:

$$L = c_{fp} \times FP \times c_{fn} \times FN$$

Through cost-aware policy learning, the model effectively captures contextual patterns in historical transaction logs, thereby eliminating the need for online exploration and ensuring both operational safety and regulatory compliance.

5.3 Conclusion

This research provides empirical evidence that offline conservative reinforcement learning can serve as a practical and ethically sound framework for optimizing credit decision-making under asymmetric risk conditions. By combining cost-sensitive reward modeling with deep policy approximation, the model successfully learns adaptive lending thresholds that minimize financial loss while reducing customer friction. The interpretability analysis of PCA components and feature-importance rankings confirms that latent behavioral signals—particularly V14, V17, and transaction Amount—exhibit significant predictive power, enabling the RL agent to detect high-risk patterns that conventional classifiers overlook. Compared with supervised baselines, the CQL-based model consistently achieves superior recall and cost efficiency, validating its suitability for deployment in real-world financial systems constrained by regulatory and operational limits.

Beyond its technical contributions, this study advances the conceptual understanding of user-centric credit management by framing risk control and incentive design as a joint optimization problem rather than a binary trade-off. The findings underscore that personalized credit strategies—when guided by conservative RL policies—can simultaneously promote consumer welfare, institutional stability, and macroeconomic growth. Future extensions may explore multi-objective reinforcement learning architectures that explicitly incorporate fairness, interpretability, and long-term user engagement metrics, further strengthening the alignment between financial innovation and sustainable economic development.

ACKNOWLEDGMENTS

Not Applicable.

FUNDING

Not Applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not Applicable.

INFORMED CONSENT STATEMENT

Not Applicable.

DATA AVAILABILITY STATEMENT

Not Applicable.

CONFLICT OF INTEREST

Not Applicable.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not application.

ABOUT THE AUTHORS

XIMENG, Yang

Board of Directors, Excellent Era Lending Service Corp., Makati, Philippines , PH, Cocoliu898@gmail.com.

YIMING, Zhang

Department of Financial Technology, Peking University, Peking, China , CN, zhanglming137@outlook.com.

REFERENCES

[1] Khraishi, R., & Okhrati, R. (2022, November). Offline deep reinforcement learning for dynamic pricing of consumer credit. In Proceedings of the Third ACM International Conference on AI in Finance (pp. 325–333).

[2] So, M. M., & Thomas, L. C. (2011). Modelling the profitability of credit cards by Markov decision processes. European Journal of Operational Research, 212(1), 123–130.

[3] Sewak, M. (2019). Temporal difference learning, SARSA, and Q-learning: Some popular value approximation-based reinforcement learning approaches. In Deep reinforcement learning: Frontiers of artificial intelligence (pp. 51–63). Springer.

[4] Sha, F., Ding, C., Zheng, X., et al. (2025). Weathering the policy storm: How trade uncertainty shapes firm financial performance through innovation and operations. International Review of Economics & Finance, 104274.

[5] Deng, X. (2025). Cooperative optimization strategies for data collection and machine learning in large-scale distributed systems. In 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT) (pp. 2151–2154). IEEE.

[6] Trench, M. S., Pederson, S. P., Lau, E. T., Ma, L., Wang, H., & Nair, S. K. (2003). Managing credit lines and prices for Bank One credit cards. Interfaces, 33(5), 4–21.

[7] Wiesemann, W., Kuhn, D., & Rustem, B. (2013). Robust Markov decision processes. Mathematics of Operations Research, 38(1), 153–183.

[8] Tan, C., Gao, F., Song, C., Xu, M., Li, Y., & Ma, H. (2024). Highly reliable CI-JSO based densely connected convolutional networks using transfer learning for fault diagnosis. Journal of Information Systems Engineering and Management. https://doi.org/10.52783/jisem.v10i4.12207

[9] Tan, C., Gao, F., Song, C., Xu, M., Li, Y., & Ma, H. (2024). Proposed damage detection and isolation from limited experimental data based on a deep transfer learning and an ensemble learning classifier. Journal of Information Systems Engineering and Management. https://doi.org/10.52783/jisem.v10i4.12206

[10] Han, X., & Dou, X. (2025). User recommendation method integrating hierarchical graph attention network with multimodal knowledge graph. Frontiers in Neurorobotics, 19, 1587973.

[11] Zhuang, R. (2025). Evolutionary logic and theoretical construction of real estate marketing strategies under digital transformation. Economics and Management Innovation, 2(2), 117–124.

[12] Yang, Z., et al. (2025). RLHF fine-tuning of LLMs for alignment with implicit user feedback in conversational recommenders. arXiv. https://arxiv.org/abs/2508.05289

[13] Deng, X., & Yang, J. (2025). Multi-layer defense strategies and privacy-preserving enhancements for membership reasoning attacks in a federated learning framework. In 2025 5th International Conference on Computer Science and Blockchain (CCSB) (pp. 278–282). IEEE.

[14] Tan, C. (2024). The application and development trends of artificial intelligence technology in automotive production. Artificial Intelligence Technology Research, 2(5).

[15] Zhang, L., & Meng, Q. (2025, September). User portrait-driven smart home device deployment optimization and spatial interaction design. In 2025 5th International Conference on Artificial Intelligence, Automation and High Performance Computing (AIAHPC) (pp. 724–728). IEEE.

[16] Yang, H., Tian, Y., Yang, Z., Wang, Z., Zhou, C., & Li, D. (2025). Research on model parallelism and data parallelism optimization methods in large language model-based recommendation systems. arXiv. https://arxiv.org/abs/2506.17551

[17] Gonzalez, J., Tran, V., Meredith, J., Xu, I., Penchala, R., Vilar-Ribó, L., et al. (2025). How it begins: Initial response to opioids strongly predicts self-reported opioid use disorder. medRxiv.

[18] Wozabal, D., & Hochreiter, R. (2012). A coupled Markov chain approach to credit risk modeling. Journal of Economic Dynamics and Control, 36(3), 403–415.

[19] Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems, 33, 1179–1191.

[20] Mendonca, R., Geng, X., Finn, C., & Levine, S. (2020). Meta-reinforcement learning that is robust to distributional shifts via model identification and experience relabeling. arXiv. https://arxiv.org/abs/2006.07178