Policy Optimization for Financial Decision-Making

Task 4: Analysis, Comparison, and Future Steps

1. Presenting the Results

Supervised Deep Learning Model

The MLP classifier was trained to predict whether a borrower would default or fully repay a loan using borrower and loan-level features.

Performance Metrics:

Metric	Value	Interpretation
AUC	0.788	The model can correctly rank defaulters vs. non-defaulters about 79% of the time.
F1-Score	0.407	Indicates moderate balance between precision and recall; limited due to class imbalance.
Accuracy	0.659	About 66% of overall predictions are correct.

Confusion Matrix: [[146578, 83618], [8605, 31674]]

The model performs well in distinguishing between good and risky borrowers (AUC ≈ 0.79). However, the F1-score indicates challenges with recall due to the heavily imbalanced dataset. It tends to overpredict defaults, leading to missed opportunities but fewer financial losses.

Offline Reinforcement Learning Agent

The Conservative Q-Learning (CQL) agent was trained on the same historical data to learn a loan approval policy that maximizes long-term financial return.

Reward Function:

• Approve + Fully Paid → + (loan_amount × interest_rate)

• Approve + Default → – loan_amount

• Deny → 0

The Estimated Policy Value (EPV) measures the average expected financial return under the learned policy. The trained CQL agent achieved a positive EPV, showing that it learned to approve profitable loans while avoiding unprofitable ones.

2. Explaining the Metrics

Why AUC and F1-Score for the DL Model?

AUC measures how well the model ranks borrowers by risk — crucial for threshold-based loan decisions. F1-score balances false positives (unnecessarily rejected applicants) and false negatives (missed risky borrowers). Together, these metrics reveal how accurately the model assesses credit risk.

Why Estimated Policy Value (EPV) for the RL Agent?

In RL, accuracy is irrelevant — the goal is maximizing cumulative financial reward. EPV quantifies the expected return of the learned policy in terms of business value. A higher EPV means better overall profitability.

3. Policy Comparison

Aspect	Supervised Deep Learning	Offline Reinforcement Learning
Objective	Predict default probability	Maximize long-term expected reward
Training Signal	Minimize binary cross-entropy loss	Maximize reward function
Decision Rule	Approve if p(default) < threshold	Approve if expected_reward > 0
Behavior	Conservative (avoids defaults)	Opportunistic (takes profitable risk)
Metric	AUC, F1-Score	Estimated Policy Value (EPV)

Example Scenario: Applicant A (Income ₹9L/year, DTI 15%, Interest Rate 18%, Grade C). The DL model rejects, but the RL agent approves because the expected reward is positive. The DL model minimizes error; RL maximizes profit.

4. Future Steps and Recommendations

Deployment Strategy:

1. Short Term: Deploy the DL model for risk scoring (stable and interpretable).

2. Mid Term: Test the RL policy in simulation or shadow mode using unseen data.

3. Long Term: Combine both models — DL for risk, RL for decision optimization.

Limitations: Dataset bias (only approved loans), simplified reward, imbalance, and interpretability issues.

Future Work: Collect rejected loan data, integrate repayment sequences, use risk-sensitive RL (CVaR, IQL), and test BCQ or TD3+BC.

5. Conclusion

The Supervised DL model provides reliable risk assessment but optimizes accuracy. The RL agent optimizes for profit. Together, they form a financially intelligent loan approval system.

Author: Mukul Garg

GitHub: https://github.com/gargmukul91066

LinkedIn: https://www.linkedin.com/in/mukul-garg-5b533b245/