Reinforcement Learning for Debt Pricing: A Case Study in Financial Services

Bruno Brandão[1,2,†], **Luana G. B. Martins**[1,2,†], **Bryan L. M. de Oliveira**[1,2,†], **Luckeciano C. Melo**[2,3], **Murilo L. da Luz**[1,2], **Eduardo Garcia**[1,2], **Marcos V. da Silva**[4], **Renato G. Avelar**[4], **Arlindo R. G. Filho**[1,2], **Anderson da S. Soares**[1,2], **Telma W. de L. Soares**[1,2]

{bruno_brandao_martins,luana.martins,bryanlincoln}@discente.ufg.br

[1]**Institute of Informatics, Federal University of Goias, Brazil**
[2]**Advanced Knowledge Center for Immersive Technologies – AKCIT**
[3]**OATML, University of Oxford**
[4]**Recovery Group**

† Equal contribution.

Abstract

Traditional static discount policies in debt recovery often fail to adapt to diverse debtor behaviors and evolving market dynamics. This research developed, evaluated, and deployed a comprehensive reinforcement learning (RL) system for optimizing discount policies to maximize recovered debt and minimize negotiation costs within a large financial institution. Our methodology encompassed developing sophisticated lifetime value (LTV) models as dynamic reward functions, implementing multi-armed bandit (MAB) meta-policies for autonomous policy evaluation and selection, and exploring diverse RL approaches including Imitation Learning and Offline RL. Key findings demonstrate superior performance of RL-driven discount policies, achieving lower average discounts and higher collection values in production compared to established baselines. The LTV models proved crucial for handling delayed feedback, while MAB meta-policies effectively orchestrated policy deployment in live operational settings. This work demonstrates the practical viability of applying advanced RL techniques to real-world financial challenges.

1 Introduction

Debt recovery is a critical function in modern financial systems. While credit promotes economic activity, it introduces default risks. Recovery firms aim to minimize creditor losses through negotiated settlements, but traditional discounting strategies rely on static rules or heuristics that lack responsiveness to debtor heterogeneity and dynamic economic contexts, potentially resulting in suboptimal recovery performance (Choong, 2024; Jankowski & Paliński, 2024).

Reinforcement learning (RL) provides a principled framework for sequential decision-making under uncertainty, where agents learn optimal strategies via interaction and feedback. By incorporating deep neural networks, Deep RL (DRL) extends this capability to handle high-dimensional state spaces and complex patterns, making it particularly well-suited for financial applications (Sutton & Barto, 2018; Naman et al., 2019). In the context of debt recovery, RL agents can tailor discount offers to individual debtor profiles and interaction histories, thereby optimizing long-term recovery outcomes and adapting to evolving behavioral patterns. Unlike supervised learning, RL directly

optimizes for business objectives and can dynamically adjust to shifting data distributions (Santana et al., 2020; Kumar et al., 2020).

This research explores the application of DRL to real-world debt recovery operations, addressing challenges such as delayed feedback and high exploration costs. Unlike previous work that involved complex systems combining administrative action modeling (Abe et al., 2010), our approach directly defines debt discounts to maximize payment. We focus on developing a solution based on learned behaviors from historical data, rather than relying on hand-crafted heuristics. This solution encompasses a reward model, a data pipeline, and policy orchestration mechanisms, all trained with historical data and deployed in an operational setting.

This work stems from collaboration between an AI research center and a major financial institution to design, implement, and deploy an RL-based system for optimizing discount policies in debt collection using real operational data. The complexity of this problem –characterized by delayed reward signals, financial risk, and limited online experimentation opportunities– motivated conservative and offline RL approaches. We developed a predictive lifetime value (LTV) model to estimate future returns from agreements as immediate reward signals, and integrated policy orchestration via multi-armed bandits (MABs) (Wahed et al., 2023; Santana et al., 2020), allowing dynamic switching between candidate policies in production based on observed performance.

Contributions. Our primary contributions include: (1) an innovative LTV-based reward modeling system addressing delayed feedback in financial RL; (2) the framing of debt recovery policy optimization as a Markov Decision Process (MDP); (3) a flexible MAB meta-policy framework for dynamic policy management; and (4) a comprehensive study evaluating various RL approaches adapted for this domain, with concrete evidence demonstrating offline RL effectiveness in real-world production environments.

2 Literature review

Reinforcement learning in financial applications. RL applications in finance have gained traction for optimizing complex decision-making under uncertainty, including algorithmic trading, portfolio optimization, risk management, fraud detection, and dynamic pricing. Maestre et al. (2019) explored RL for fair dynamic pricing, while Khraishi & Okhrati (2022) introduced offline deep RL for dynamic pricing of consumer credit, demonstrating effective policy learning from static datasets without demand function assumptions. In debt collection specifically, Kuzmin et al. (2022) discussed deep RL applications, Abe et al. (2010) used constrained offline RL for debt collection optimization, and Melo et al. (2021) enabled offline RL for digital marketing via conservative Q-learning (CQL) (Kumar et al., 2020). This project builds upon these foundations by focusing on discount policy optimization.

Multi-armed bandits. Multi-armed bandit (MAB) algorithms excel as meta-selectors in dynamic domains like pricing (Naman et al., 2019), recommendation systems (Santana et al., 2020), NLP (Wahed et al., 2023), real-time bidding (Jiayi et al., 2021), and optimizer selection (Ferreira et al., 2017; Meidani et al., 2022). Hierarchical MABs, in particular, outperform individual recommenders and ensembles by adapting to dynamic preferences (Santana et al., 2020), with approaches like MArBLE advancing specialized language models (Wahed et al., 2023). MABs demonstrate robust performance across applications, from boosting airline revenue by 43% (Naman et al., 2019) to matching metaheuristic optimizers (Ferreira et al., 2017), proving their effectiveness in policy selection under uncertainty.

Imitation learning (behavioral cloning). Imitation learning (IL), particularly behavior cloning (BC) (Pomerleau, 1989; Ross & Bagnell, 2010), has emerged as a compelling approach in robotics (Alan et al., 2019; Sasaki et al., 2020; Yin et al., 2022), autonomous driving (Bansal et al., 2019; Xu & Zhao, 2024), and online retail (Shi et al., 2019) to minimize costly or risky exploration. Innovations like adversarial BC (Sasaki et al., 2020), approximated BC loss (Lowman et al., 2021), and active learning (Khanh & Hal, 2020) enhance efficiency and safety. While concerns about

error compounding and generalization remain (Ciftci et al., 2024), hybrid approaches with simulated perturbations and expert refinements (Sun et al., 2023; Shi et al., 2019) address these issues. In debt discounting, where delayed feedback and financial risks exist, IL has the potential to leverage historical decisions to bootstrap feasible policies without costly exploration.

Offline reinforcement learning. Offline RL enables policy learning from fixed datasets without environment interaction, crucial for domains like debt discounting with costly exploration and delayed feedback. Key challenges include distributional shift, where policies select out-of-distribution actions leading to unreliable value estimates (Kumar et al., 2020). Model-free approaches like conservative Q-learning (CQL) address this through Q-value regularization that penalizes OOD actions (Kumar et al., 2020), while conservative offline distributional actor critic (CODAC) and state-conditioned action quantization (SAQ) demonstrate state-of-the-art performance in complex domains (Ma et al., 2021; Luo et al., 2023). Imitation-based methods like monotonic advantage reweighted imitation learning (MARWIL) reweight actions by estimated advantages, providing improvement guarantees without behavior policy probabilities (Wang et al., 2018). This is particularly useful for batched datasets with unknown or heterogeneous logging policies. Both CQL and MARWIL offer robust, theoretically sound solutions to distributional shift.

3 Research methodology

3.1 Problem formulation as an RL task

The core problem of optimizing discount policies for debt recovery was formulated as a reinforcement learning task by defining the key components of a Markov Decision Process (MDP). The state space ($\mathcal{S}$) incorporated debtor demographics, debt characteristics (outstanding amount, age, product type), historical payment patterns, and contextual indicators like the debtor's "public type" – an internal classification of the debtor-company relationship. The action space ($\mathcal{A}$) consisted of possible discount percentages, exploring varying discount levels as a discrete or discretized continuous space. The reward function ($\mathcal{R}$) derived from our LTV model predicted the monetary value expected to be recovered from an agreement, transforming delayed feedback into immediate signals suitable for RL algorithms. The policy (π) represented the learned mapping from states to optimal discount actions, formally expressed as $\pi(a|s)$, with the objective of maximizing expected cumulative LTV across all debt instances.

3.2 Data infrastructure and collection

The foundation of this research was historical interaction and payment data from the company's operations, managed through a feature store infrastructure. The feature store provided a centralized repository for features, enabled version control for reproducibility, simplified data acquisition and processing, fostered collaboration through standardized feature engineering, and improved efficiency by an estimated one-third reduction in model development time. Implemented on the Databricks platform, this infrastructure served as a foundational enabler for rapid iteration and deployment of data-driven models, ensuring data quality, accessibility, and consistency throughout the project.

3.3 LTV modeling as a dynamic reward function

A significant challenge in applying RL to debt recovery is the substantial delay between offering a discount and observing the full payment outcome. Debtors often settle on a multi-installment agreement and may stop payment after an undetermined number of months. To address this, we developed a lifetime value (LTV) model, termed TwoStage-LTV, designed to provide an immediate and meaningful reward signal to RL agents. The primary objective of this LTV model is to provide the core performance metric, the reward function to be maximized by the RL algorithms. Traditional debt recovery metrics are often characterized by long feedback delays of months or years, which

hinder efficient learning of RL algorithms in practical time (Sutton & Barto, 2018). Our LTV model circumvents this by predicting the expected revenue from an agreement at the moment it is accepted by the debtor, thereby offering timely feedback crucial for effective learning.

Empirical observations revealed a bimodal distribution in financial outcomes for debt agreements: nearly half result in no payments, while most that receive any payment are often paid in full. To address this, the TwoStage-LTV model adopts a two-stage methodology. First, a classification decision tree model predicts the probability that an agreement will receive at least one paid installment, effectively filtering out those that are unlikely to yield monetary value. Second, for agreements predicted by the classifier to receive at least one payment, a regression decision tree model estimates the actual monetary value expected to be recovered from them.

To assess its efficacy, the TwoStage-LTV model's performance was rigorously benchmarked against actual realized cash flows and two baseline LTV methodologies. The first baseline, *Legacy-LTV*, represents a pre-existing company approach that categorizes debts into broad types and calculates LTV as the historical average of monthly payments within each debt cluster. This method relies on simple classification and does not account for individual contract characteristics. The second baseline, *Baseline-LTV*, which was also a pre-existing company approach, utilizes a single decision tree model to predict LTV directly. While capable of capturing some non-linear relationships, it is not inherently designed to handle the bimodal payment distribution. The detailed performance comparison against these baselines is presented in Section 4.1.

3.4 Reinforcement learning approaches

Imitation learning (behavioral cloning). Our initial strategy involved imitation learning (IL), also known as behavioral cloning (Pomerleau, 1989; Ross & Bagnell, 2010), to develop policies. The objective is to mimic historical discount decisions from the company's operations that were considered to have produced favorable outcomes. A significant advantage of this approach is its ability to synthesize the strengths of multiple existing decision policies – often hand-crafted rule-based systems applied to user clusters (e.g., grouped by age or income) – into a single, more nuanced policy capable of applying the best strategy on a per-debt-instance basis.

A crucial aspect of our IL implementation is dataset curation. To move beyond simple behavioral cloning, we developed a scoring mechanism to identify and select "expert" actions for imitation. This score combined predictions from our TwoStage-LTV model with historical payment data, enhancing the reliability of the LTV model for evaluating past actions, particularly in mitigating errors associated with smaller discounts where the previous LTV model had shown limitations. Standard deep neural networks were then trained using supervised learning on these curated (state, expert action) pairs.

Further enhancements were achieved through dataset augmentation. To encourage policies that offer lower average discounts without compromising recovery rates, we generated synthetic training examples. For debt instances with existing payments, synthetic examples were created by systematically changing discount levels, based on the premise that debtors would also have settled with higher discounts. The corresponding received values were adjusted proportionally, decreasing linearly as synthetic discounts approached 100%. For cases without historical payment, received values remained zero until discounts reached 0%. The density of generated examples decayed based on deviation from original discounts, creating a richer dataset that allowed the agent to learn more robustly and explore beneficial deviations from historically observed actions.

These imitation learning efforts culminated in the development of several distinct policies: IL-Baseline-Policy, trained on datasets filtered by historical payments; IL-LTV-Policy, trained on data filtered by the TwoStage-LTV model; and IL-Augmented-Policy, which incorporated the dataset augmentation techniques.

Offline reinforcement learning. To move beyond mimicking past behavior, we also investigated offline reinforcement learning techniques. The objective here was to learn policies that optimize

long-term rewards, specifically LTV, directly from a fixed batch of historical data. This paradigm is particularly valuable as it avoids the risks and costs associated with online interaction during the learning phase, a critical consideration in financial applications. Offline RL aims to derive optimal or near-optimal policies even when the available data consists of sub-optimal trajectories generated by various, potentially unknown, historical policies (Levine et al., 2020).

Our initial exploration covered several offline RL algorithms, including conservative Q-learning (CQL) (Kumar et al., 2020), standard deep Q-networks (DQN) (Mnih et al., 2015) adapted for the offline setting, and multi-armed bandits like LinUCB (Li et al., 2010). However, due to project time constraints and the promising results from IL, approaches that built upon behavioral cloning were identified as the most practical direction for deeper investigation.

Consequently, the primary offline RL method adopted was monotonic advantage re-weighted imitation learning (MARWIL) (Wang et al., 2018). MARWIL extends behavioral cloning by incorporating advantage-weighted learning, addressing the challenge of learning from sub-optimal demonstrations by reweighting actions based on their estimated advantages. This allows the policy to preferentially learn from more effective actions in the historical data, while the behavioral cloning component maintains stability and mitigates distributional shift issues. The reward signal for calculating advantages was provided by our TwoStage-LTV model, and offline RL agents were trained using the same datasets curated for imitation learning policies.

Meta-policies (multi-armed bandits). We developed a multi-armed bandit (MAB) meta-policy to orchestrate deployment of various discount strategies Lattimore & Szepesvári (2020). The MAB dynamically allocated traffic among candidate policies derived from imitation learning, offline RL, or existing heuristic-based approaches, optimizing allocation based on real-time performance. The non-contextual MAB operated as an autonomous A/B/n testing framework, adjusting traffic allocation based on each policy's recent LTV performance. Through continuous monitoring and reallocation, the MAB prioritized higher-performing policies while reducing use of less effective ones.

The MAB implementation used the Upper Confidence Bound (UCB) algorithm, which balances exploration and exploitation by selecting actions based on both estimated value and uncertainty. UCB maintains confidence intervals for each policy's performance and selects the policy with the highest upper confidence bound, ensuring promising but less-explored policies receive sufficient traffic while gradually converging to the best-performing options. This approach proved particularly effective for debt recovery, where delayed feedback requires careful exploration to avoid prematurely discarding potentially effective policies.

3.5 Evaluation framework

A dual evaluation framework was implemented, combining offline and online assessments. Offline evaluation used a unified system to compare policies across key metrics: cash efficiency, net present value, average discount offered, and average collection amount, calculated on a representative test set. This informed decisions about which policies advanced to online testing.

Online evaluation deployed selected policies in production, managed by MAB-based meta-policies through A/B testing or dynamic traffic allocation. Policies were monitored over extended periods (e.g., 2 months for Offline-Policy) with control groups for comparison across different client segments: UN = unsegmented (control group), INT = interacted (at least once), RINT = recently interacted (last 3 months), NI = never interacted. This comprehensive approach, integrating robust infrastructure, LTV modeling, and diverse RL strategies, effectively addressed the challenges of optimizing debt recovery policies.

4 Results

4.1 Performance of the LTV model

We evaluated the performance of our TwoStage-LTV model, as described in Section 3.3, against the two baseline methods (Legacy-LTV and Baseline-LTV) and actual realized cash flow on a held-out test set. The TwoStage-LTV model employs a two-stage process combining classification and regression, specifically designed to handle the bimodal distribution characteristic of debt recovery data. Meanwhile the single-stage baseline achieves a performance with higher error.

Table 1: Net present value predictions at day 120 (percentage of debt recovered).

Model	NPV (%)	Relative Error
Actual Cash Flow	1.25	–
Legacy-LTV	2.75	+120%
Baseline-LTV	0.72	-44%
TwoStage-LTV	1.55	+24%

Cumulative payment predictions. Table 1 summarizes the net present value (NPV) predictions at day 120. Comparing these predictions against actual cumulative payments reveals significant accuracy differences: the *Legacy-LTV* approach dramatically overestimates recovery by 120% (predicting 2.75% recovery versus an actual 1.25%), primarily due to its cluster-based averaging failing to account for non-paying accounts and contract-level heterogeneity. Conversely, the *Baseline-LTV* decision tree underestimates recovery by 44% (predicting only 0.72%), as its standard structure cannot adequately model the bimodal nature of debt payments and misses high-value recoveries. In contrast, our *TwoStage-LTV* model achieves much better accuracy with only a 24% overestimation, thanks to its dedicated stages for payment likelihood and amount prediction that better capture the underlying payment distribution. An overestimating model is preferred in this applied scenario. An underestimating model could give poor scores to potentially good policies, reducing their client share immediately and preventing revenue generation as corrections from real debtor values could take months. Meanwhile, overestimating policies can be adjusted once debtors stop paying. While delayed feedback affects both cases, stopping payment is immediate while full payment takes as long as the agreed installments.

Table 2: Distribution of contracts by payment level: actual vs. TwoStage-LTV predictions.

Payment level	0%	20%	40%	60%	80%	100%
Actual	73.0	0.9	1.5	4.4	1.5	19.0
Prediction	72.0	0.3	0.6	1.1	1.7	24.0

Payment distribution analysis. To assess model calibration across different payment outcomes, we compared the distribution of actual versus predicted total payments as a percentage of original contract value, detailed in Table 2. The TwoStage-LTV model accurately captures the zero-payment mode (72% predicted vs. 73% actual) and reasonably predicts full payments, despite some overestimation (24% predicted vs. 19% actual). Partial payment prediction needs improvement, especially for the 60% category, underestimated by 3.3 percentage points. Since partial payments represent the smallest portion of debtors, dataset imbalance affected regression model training. While techniques like loss weighting could improve accuracy, the performance gain we delivered was substantial, and the underperformed section was the smallest in the dataset. Overall, our TwoStage-LTV model represents a substantial improvement over both benchmarks. It eliminates the extreme overestimation of the Legacy-LTV approach and the systematic underestimation of the Baseline-LTV model, while providing a solid foundation for reward estimation in our reinforcement learning framework.

4.2 Efficacy of imitation learning policies

Imitation learning policies were developed to leverage historical data and provide strong baselines for our debt recovery optimization system. The IL-Baseline-Policy demonstrated promising results

in both offline and online evaluations, achieving cash efficiency comparable to or even superior to the company's best existing policy at the time, which was based on a simple mapping between debtor credit scores and suggested discounts.

Building on this foundation, we enhanced our imitation learning framework through dataset augmentation techniques. The resulting IL-Augmented-Policy achieved a notable improvement, reducing the average discount from 84% (IL-Standard-Policy) to 78% in offline evaluations. This reduction in average discounts was accomplished while maintaining comparable recovery rates, indicating that the augmented policy learned to offer more efficient discounts by exploring variations around historical strategies while remaining grounded in proven approaches. The key innovation involved generating synthetic examples around historical discount decisions, enabling the policy to discover more effective discount levels without straying too far from expert demonstrations.

4.3 Performance of the offline RL policy

The offline RL policy demonstrated superior performance across all customer segments during its two-month production deployment against imitation learning controls. As shown in Table 3, it consistently achieved lower average discounts with higher ticket values. In the unsegmented (UN) segment, it offered 71.75% average discounts versus 75.16% for controls, while securing $516 average tickets compared to $470. The never-interacted (NI) segment showed particularly strong results, with 55.66% average discounts versus 62.44% and $646 average tickets versus $551.

Table 3: Comparison of metrics across client segments and control group. Each segment represents the client's interaction history with the company: UN = unsegmented (control group), INT = interacted (at least once), RINT = recently interacted (last 3 months), NI = never interacted.

	Control Group				**Offline-Policy**			
	UN	**INT**	**RINT**	**NI**	**UN**	**INT**	**RINT**	**NI**
Avg. Discount (%)	75.16	74.36	79.16	62.44	71.75	70.94	77.01	55.66
Avg. Ticket ($)	470	374	492	551	516	419	524	646

This performance advantage likely stems from the policy's ability to learn from historical data while avoiding online exploration risks, combined with its optimization for long-term lifetime value through our TwoStage-LTV model. The results validate the effectiveness of offline learning in high-stakes financial applications.

4.4 Effectiveness of MAB-based meta-policies

In a different user base, the multi-armed bandit (MAB) meta-policy system dynamically allocated traffic among competing discount policies. This approach effectively automated A/B/n testing and optimized traffic allocation in live production. Figure 1 shows the candidate policies managed by the MAB system. The policies include both legacy approaches (developed by the company's previous team using traditional rule-based methods) and novel RL-based policies (denoted with "ceia"-like naming conventions) developed in this research. The MAB autonomously evaluated and allocated traffic among these competing policies based on real-time performance, successfully integrating legacy and novel approaches in a unified optimization framework.

5 Discussion

The empirical results demonstrate the potential of offline RL for optimizing debt recovery discount policies. The success of the Offline-Policy in production highlights the efficacy of learning from large historical datasets while mitigating exploration risks in finance. Key factors include optimization for long-term LTV guided by our custom LTV model, and the ability to balance exploiting

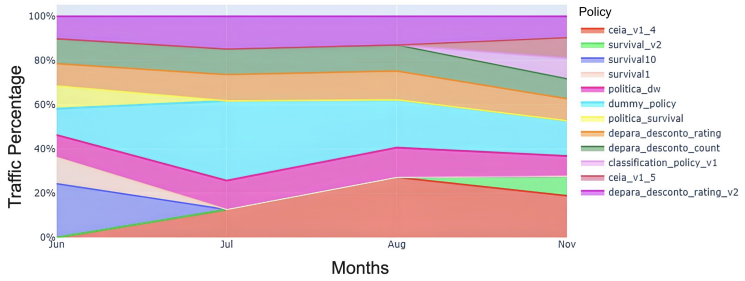

Figure 1: Traffic allocation percentage managed by the MAB-based meta-policy over time. The MAB dynamically adjusts traffic to higher-performing policies.

known strategies with generalizing to new situations. The TwoStage-LTV model, despite needing refinement for partial payment prediction, served as a crucial component for reward signaling in environments with delayed feedback. Imitation learning with dataset augmentation yielded strong baselines, while MAB-based meta-policies provided practical solutions for managing multiple strategies in production. This work aligns with growing trends towards offline RL in finance (Maestre et al., 2019; Khraishi & Okhrati, 2022), consistent with findings on offline DRL effectiveness in consumer credit pricing. The dynamic integration of LTV prediction as the primary RL reward signal to overcome delayed feedback is, to the best of our knowledge, novel and impactful. The MAB orchestration of multiple RL-derived policies represents an advanced application beyond simple A/B testing.

6 Conclusion

Our production deployment results validate the practical viability of offline RL for debt recovery optimization, with the Offline-Policy consistently outperforming traditional approaches across key metrics. The integration of our TwoStage-LTV model proved particularly effective in overcoming the challenges of delayed feedback, while the MAB-based meta-policy system demonstrated robust capabilities for dynamic policy management.

The study's methodological innovations extend beyond immediate application, offering generalizable insights for financial RL systems. The universal Feature Store architecture and comparative framework provide valuable blueprints for future implementations. However, several challenges remain, particularly in model interpretability and simulation fidelity, requiring further study. Unfotunately, due to industry confidentiality, we were unable to divulge further insights on performance based on debtor information as well as importance of features for LTV calculation.

Looking ahead, we identify three critical research directions: (1) development of more sophisticated LTV models capable of handling partial payments and evolving customer behaviors; (2) integration of explainable AI techniques to enhance policy transparency; and (3) incorporation of fairness constraints to ensure equitable treatment across diverse customer segments (Abe et al., 2010). These advancements will be crucial for realizing the full potential of RL in financial applications while maintaining ethical standards.

Broader Impact Statement. RL application to debt recovery optimization carries significant societal implications. Positive impacts include increased efficiency and potentially fairer, personalized discount offers. Risks include perpetuating historical biases, lacking transparency in complex models, and potential overly aggressive collection strategies if optimization solely prioritizes recovery without ethical constraints. Users must implement rigorous bias detection and mitigation, ensure decision-making transparency, and establish clear ethical guidelines. Future work must actively address fairness considerations to ensure responsible and equitable technology use, creating systems that are effective, fair, and aligned with broader societal values.

Acknowledgments

We thank Recovery Group for their support and collaboration in this research.

This work has been partially funded by the project Research and Development of Digital Agents Capable of Planning, Acting, Cooperating and Learning supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

References

Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 75–84, 2010.

Wu Alan, Piergiovanni A., and Ryoo M. Model-based Behavioral Cloning with Future Image Similarity Learning. *Conference on Robot Learning*, 2019.

Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. Chauffeurnet: Learning to Drive by Imitating the Best and Synthesizing the Worst. In *Robotics: Science and Systems XV*. Robotics: Science and Systems Foundation, jun 22 2019. DOI: 10.15607/rss.2019.xv.031. URL http://dx.doi.org/10.15607/RSS.2019.XV.031.

Macy Choong. 5 effective bank debt collection strategies that minimise non-performing loans, December 2024. URL https://www.kewmann.com/resources/blogs/5-effective-bank-debt-collection-strategies-that-minimise-non-performing-loans/. Accessed: 2025-06-02.

Yusuf Umut Ciftci, Darren Chiu, Zeyuan Feng, Gaurav S. Sukhatme, and Somil Bansal. Safe-gil: Safety guided imitation learning for robotic systems. 2024. URL https://arxiv.org/abs/2404.05249.

Alexandre Silvestre Ferreira, Richard Aderbal Goncalves, and Aurora Pozo. A Multi-Armed Bandit selection strategy for Hyper-heuristics. In *2017 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 6 2017. DOI: 10.1109/cec.2017.7969356. URL http://dx.doi.org/10.1109/CEC.2017.7969356.

Rafał Jankowski and Andrzej Paliński. Debt collection model for mass receivables based on decision rules—a path to efficiency and sustainability. *Sustainability*, 16(14):5885, July 2024.

Xie Jiayi, Tashman M., Hoffman John, Winikor Lee, and Gerami Rouzbeh. Online and Scalable Model Selection with Multi-Armed Bandits. *arXiv.org*, 2021.

Nguyen Khanh and Daum'e Hal. Active Imitation Learning from Multiple Non-Deterministic Teachers: Formulation, Challenges, and Algorithms. *arXiv.org*, 2020.

Raad Khraishi and Ramin Okhrati. 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, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. 33:1179–1191, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Gleb Kuzmin, Aleksandr I Panov, Ivan Razvorotnev, and Vyacheslav Rezyapkin. Application of deep reinforcement learning methods in debt collection. In *Proceedings of the Fifth International Scientific Conference Intelligent Information Technologies for Industry (IITI'21)*, pp. 66–77. Springer, 2022.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Corey A. Lowman, Joshua S. McClellan, and Galen E. Mullins. Imitation Learning with Approximated Behavior Cloning Loss. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8921–8927. IEEE, sep 27 2021. DOI: 10.1109/iros51168.2021.9636797. URL http://dx.doi.org/10.1109/IROS51168.2021.9636797.

Jianlan Luo, Perry Dong, Jeffrey Wu, Aviral Kumar, Xinyang Geng, and Sergey Levine. Action-Quantized Offline Reinforcement Learning for Robotic Skill Learning. *Conference on Robot Learning*, 2023. DOI: 10.48550/ARXIV.2310.11731. URL https://arxiv.org/abs/2310.11731.

Yecheng Jason Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. 2021. URL https://arxiv.org/abs/2107.06106.

Roberto Maestre, Juan Duque, Alberto Rubio, and Juan Arévalo. Reinforcement learning for fair dynamic pricing. In *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 1*, pp. 120–135. Springer, 2019.

Kazem Meidani, Seyedali Mirjalili, and Amir Barati Farimani. Mab-OS: Multi-Armed Bandits Metaheuristic Optimizer Selection. *Applied Soft Computing*, 128:109452, 10 2022. ISSN 1568-4946. DOI: 10.1016/j.asoc.2022.109452. URL http://dx.doi.org/10.1016/j.asoc.2022.109452.

Luckeciano Melo, Luana Martins, Bryan Oliveira, Bruno Brandão, Douglas W. Soares, and Telma Lima. Pulserl: Enabling offline reinforcement learning for digital marketing systems via conservative q-learning. In *2nd Workshop on Offline Reinforcement Learning at NeurIPS*, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Shukla Naman, Kolbeinsson A., Marla Lavanya, and Yellepeddi Kartik. Adaptive Model Selection Framework: An Application to Airline Pricing. *arXiv.org*, 2019.

Dean A. Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, pp. 305–313, 1989.

Stéphane Ross and J. Andrew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668, 2010.

Marlesson R. O. Santana, Luckeciano C. Melo, Fernando H. F. Camargo, Bruno Brandão, Anderson Soares, Renan M. Oliveira, and Sandor Caetano. Contextual meta-bandit for recommender systems selection. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, pp. 444–449, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. DOI: 10.1145/3383313.3412209. URL https://doi.org/10.1145/3383313.3412209.

Fumihiro Sasaki, Tetsuya Yohira, and Atsuo Kawaguchi. Adversarial behavioral cloning. *Advanced Robotics*, 34(9):592–598, feb 21 2020. ISSN 0169-1864. DOI: 10.1080/01691864.2020.1729237. URL http://dx.doi.org/10.1080/01691864.2020.1729237.

Jing-Cheng Shi, Yang Yu, Qing Da, Shi-Yong Chen, and An-Xiang Zeng. Virtual-Taobao: Virtualizing Real-World Online Retail Environment for Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4902–4909, jul 17 2019. ISSN 2374-3468. DOI: 10.1609/aaai.v33i01.33014902. URL http://dx.doi.org/10.1609/aaai.v33i01.33014902.

Xiatao Sun, Mingyan Zhou, Zhijun Zhuang, Shuo Yang, Johannes Betz, and Rahul Mangharam. A Benchmark Comparison of Imitation Learning-based Control Policies for Autonomous Racing. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–5. IEEE, jun 4 2023. DOI: 10.1109/iv55152.2023.10186780. URL http://dx.doi.org/10.1109/IV55152.2023.10186780.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Muntasir Wahed, Daniel Gruhl, and Ismini Lourentzou. Marble: Hierarchical Multi-Armed Bandits for Human-in-the-Loop Set Expansion. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4857–4863. ACM, oct 21 2023. DOI: 10.1145/3583780.3615485. URL http://dx.doi.org/10.1145/3583780.3615485.

Qing Wang, Jiechao Xiong, Lei Han, Peng Sun, Han Liu, and Tong Zhang. Exponentially weighted imitation learning for batched historical data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6291–6300, Red Hook, NY, USA, 2018. Curran Associates Inc.

Huanshi Xu and Dongfang Zhao. Forward-Prediction-Based Active Exploration Behavior Cloning. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pp. 2112–2117. IEEE, mar 29 2024. DOI: 10.1109/ainit61980.2024.10581680. URL http://dx.doi.org/10.1109/AINIT61980.2024.10581680.

Zhao-Heng Yin, Weirui Ye, Qifeng Chen, and Yang Gao. Planning for Sample Efficient Imitation Learning. *Neural Information Processing Systems*, 2022. DOI: 10.48550/ARXIV.2210.09598. URL https://arxiv.org/abs/2210.09598.