DOI 10.37882/2223-2966.2025.05.03

# СИНТЕТИЧЕСКИЕ ДАННЫЕ КАК ОСНОВА САМООБУЧАЮЩЕЙСЯ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ ДЛЯ БАНКОВ: МОДЕЛИРОВАНИЕ КЛИЕНТСКОГО ПОВЕДЕНИЯ И УЧЁТ МАКРОЭКОНОМИЧЕСКИХ ФАКТОРОВ

## SYNTHETIC DATA AS A FOUNDATION FOR A SELF-LEARNING RECOMMENDER SYSTEM IN BANKING: MODELING CLIENT BEHAVIOR AND MACROECONOMIC FACTORS

***Z. Askerov***

*Summary*. The development of modern intelligent recommender systems in the banking sector is constrained by limited access to real customer data due to legal and ethical concerns. This paper presents a methodological framework for building a self-learning recommender system based on the comprehensive generation of synthetic data. The system architecture includes a multi-agent model that simulates transactional, deposit, and investment behaviors of clients. Data generation incorporates language models, stochastic processes, and procedural simulations based on behavioral profiles. Special emphasis is placed on integrating macroeconomic indicators (exchange rates, commodity prices, interest rates) into the synthetic environment to recreate realistic scenarios of financial instability. A visual and statistical analysis of the generated dataset confirms its adequacy for training neural network models and reinforcement learning algorithms. The proposed approach ensures reproducibility, scalability, and data privacy in the development of AI-powered financial systems.

*Keywords*: synthetic data, recommender systems, reinforcement learning, multi-agent simulation, banking technology, client anomaly detection, macroeconomic indicators, data generation, behavioral modeling, intelligent systems.

***Аскеров Заур Ханахмедович***
*Аспирант, Федеральное государственное автономное образовательное учреждение высшего образования Национальный исследовательский ядерный университет МИФИ*
*zaur_askerov_1998@mail.ru*

*Аннотация*. Разработка современных интеллектуальных рекомендательных систем в банковской сфере сталкивается с ограниченным доступом к реальным клиентским данным, обусловленным юридическими и этическими барьерами. В данной работе предложен методологический подход к созданию самообучающейся рекомендательной системы, основанный на комплексной генерации синтетических данных. Представлены архитектура системы и структура мультиагентной модели, имитирующей транзакционное, депозитное и инвестиционное поведение клиентов. Генерация данных осуществляется с использованием языковых моделей, стохастических процессов и процедурного моделирования на основе поведенческих профилей. Отдельное внимание уделено интеграции макроэкономических индикаторов (курсы валют, цены на сырьё, процентные ставки) в синтетическую среду для формирования реалистичных сценариев финансовой нестабильности. Проведён визуальный и статистический анализ сгенерированной выборки, подтверждающий её пригодность для обучения нейросетевых моделей и алгоритмов обучения с подкреплением. Предложенный подход обеспечивает воспроизводимость, масштабируемость и безопасность данных при разработке финансовых ИИ-систем.

*Ключевые слова*: синтетические данные, рекомендательные системы, обучение с подкреплением, мультиагентное моделирование, банковские технологии, аномальное поведение клиентов, макроэкономические факторы, генерация данных, имитационное моделирование, интеллектуальные системы.

## Введение

В последние годы в банковской сфере наблюдается рост интереса к интеллектуальным системам, способным обеспечивать персонализированное обслуживание клиентов. В условиях высокой конкуренции, экономической нестабильности и резких макроэкономических колебаний банки стремятся предлагать не просто продукты, а адаптивные рекомендации, учитывающие поведение конкретного клиента. При этом традиционные рекомендательные алгоритмы на основе коллаборативной или контентной фильтрации демонстрируют недостаточную устойчивость к нестандартным ситуациям.

Особую актуальность приобретает задача создания самообучающихся систем, способных обнаруживать аномалии в поведении клиентов, учитывать макроэкономическую обстановку и подстраиваться под изменение предпочтений пользователей. Такая система должна не только учитывать исторические данные, но и эффективно реагировать на внешние кризисные факторы, изменяющие поведение клиентов в режиме реального времени.

Цель исследования — разработка адаптивной рекомендательной системы, устойчивой к аномальному поведению и чувствительной к динамике макроэкономической среды. Основной акцент сделан на исполь-

зовании синтетических данных для моделирования различных сценариев, а также применении методов обучения с подкреплением и детекции аномалий.

## Теоретико-методологические основы

Рекомендательные системы применяются в банках для повышения точности маркетинговых предложений, управления рисками и оптимизации клиентского опыта. Наиболее распространены [1]:

— Контентные методы (profile-based)
— Коллаборативная фильтрация
— Гибридные подходы.

Современные исследования (Wu & Li, 2025; Hernández et al., 2018) выделяют тенденции к переходу от статических моделей к самообучающимся архитектурам, чувствительным к контексту и изменениям среды. В условиях кризиса или инфляции рекомендательная система должна динамически адаптировать стратегии под текущие условия [2].

Выявление аномального поведения — ключевая задача в финансовой аналитике. Используемые методы [3]:

— Isolation Forest — быстро изолирует редкие случаи
— AutoEncoder — фиксирует высокую ошибку восстановления при отклонениях
— One-Class SVM — строит границу нормального класса
— Графовые методы — ищут отклонения в структуре взаимодействий
— Комбинированные ансамбли — объединяют подходы для повышения точности.

Кризисные события (всплеск снятия вкладов, массовые переводы) требуют переобучения моделей и внедрения макроэкономических регуляторов внутри детектора [4].

Синтетическая генерация данных используется в случаях [5]:

— отсутствия доступа к реальным транзакциям;
— необходимости масштабируемых обучающих выборок;
— воспроизведения нестандартных сценариев (кризисы, фрод);
— защиты персональных данных [6].

Генерация может производиться с помощью [7]:

— стохастических моделей (нормальные, экспоненциальные распределения);
— LLM-моделей (например, GPT);
— мультиагентных симуляций с поведенческими профилями;
— GAN-сетей.

## Методология генерации синтетических данных для имитации поведенческих сценариев

Разработка синтетического датасета стала ключевым этапом в построении самообучающейся рекомендательной системы. Учитывая невозможность использования реальных клиентских данных в силу регуляторных ограничений и вопросов конфиденциальности, было принято решение реализовать полнофункциональную архитектуру данных на основе искусственно сгенерированных наблюдений, сохраняющих внутреннюю бизнес-логику и структурные зависимости, характерные для розничного банковского сектора [8].

Процесс генерации данных включал несколько взаимодополняющих подходов, каждый из которых использовался для различных подсистем будущей модели [9]:

1. Стохастическое моделирование

Для числовых атрибутов, таких как суммы транзакций, вклады и объёмы инвестиций, применялись параметрические распределения, в первую очередь нормальные и логнормальные. Это позволило воссоздать реалистичное распределение финансовой активности, включая наличие «длинного хвоста» — характерного признака эмпирических данных в банковской практике [10].

2. Генеративные языковые модели (LLM)

Для семантических атрибутов (категории расходов, наименования получателей, событийные описания и др.) применялась языковая модель GPT (версии 4), адаптированная под задачи синтетической генерации. Модель использовалась как инструмент создания текстовых меток и категорий в соответствии с заданными шаблонами клиентского поведения, обеспечивая при этом разнообразие и правдоподобие лексических конструкций [11].

3. Мультиагентное моделирование

Для моделирования клиентского поведения применялась концепция поведенческих профилей, реализуемая через мультиагентную симуляцию. Каждому клиенту приписывался один из сценариев — «пассивный вкладчик», «активный потребитель», «инвестор», «реактивный клиент» и т.п., в рамках которых генерировалась временная последовательность событий, транзакций и инвестиций. Таким образом, создавалась динамика поведения, позволяющая обучать модель на последовательностях, имитирующих реальные жизненные циклы клиентов.

4. Генеративно-состязательные сети (GAN)

Для финансово-временных рядов (например, currency_rates, precious_materials_prices) применялись

GAN-модели, специально обученные на исторических данных с добавлением случайных трендов и волатильностей. Это обеспечило достоверную имитацию макроэкономических сценариев, включая резкие скачки курсов и нестабильность на сырьевых рынках [12].

### Формирование и роль синтетических данных в обучении рекомендательной системы

Интеграция этих методов была реализована с сохранением связности данных, что подтверждается представленной ER-диаграммой (рисунок 1). Все таблицы связываются по идентификатору клиента (ID клиента) и, при необходимости, по временной метке (Дата, Дата и время), что обеспечивает возможность сквозного поведенческого анализа. Такие связи являются основой для построения сложных признаковых пространств и позволяют применять алгоритмы обучения с подкреплением, где состояние среды определяется одновременно поведенческими и макроэкономическими факторами.

Сформированная структура данных охватывает как микроуровень клиентской активности (транзакции, депозиты, инвестиции), так и макроуровень (валютные колебания, ценовые изменения на сырье), что создаёт богатую среду для адаптивной логики модели. Более того, включение таблицы событий (event_history) позволяет учитывать контекстные изменения, такие как обращения в поддержку или изменение персональных данных, усиливая поведенческую осведомлённость модели.

При этом особое внимание уделялось сохранению внутренней логики данных, временной непротиворечи-

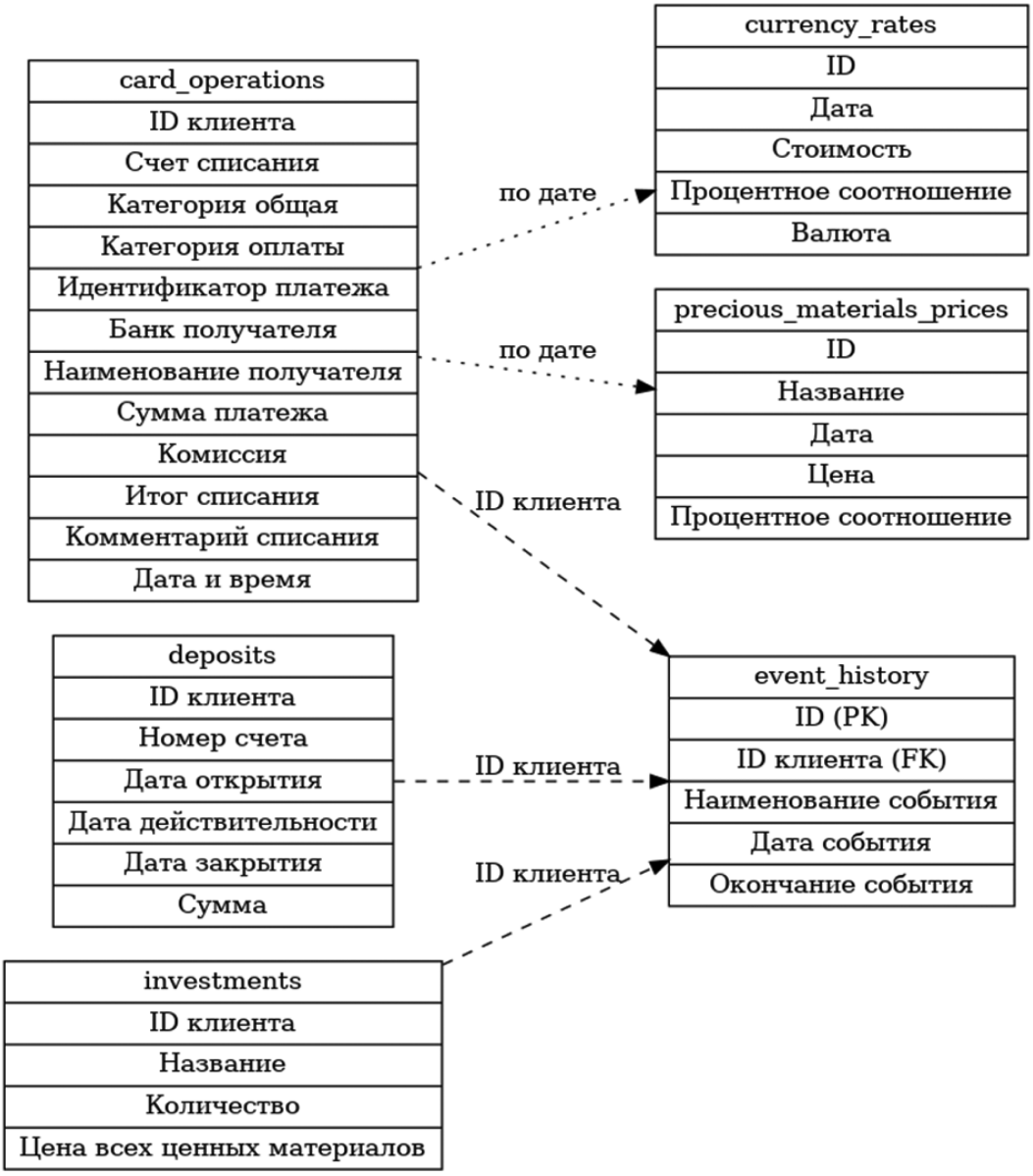


Рис. 1. ER-диаграмма синтетической базы данных

вости и реалистичному распределению признаков. Конструкция синтетической базы включает более 100 000 записей и охватывает следующие тематические блоки:

— Транзакционная активность (card_operations) — отражает повседневные расходы клиентов и используется для выделения поведенческих паттернов, построения профилей и оценки отклонений от нормы.
— Депозитные продукты (deposits) — позволяют моделировать склонность клиентов к сберегательному поведению, что критически важно при оценке финансовой устойчивости.
— Инвестиционные действия (investments) — репрезентируют разнообразие стратегий вложений и дают возможность разделения клиентов по уровню терпимости к риску.
— История событий (event_history) — включает неоперационные изменения, такие как обновление данных, обращения в поддержку и т.п., что дополняет поведенческий портрет.
— Макроэкономические факторы (currency_rates, precious_materials_prices) — отражают влияние курсов валют и стоимости сырья, выступая внешними переменными для модуля адаптации рекомендаций в условиях кризисов.

Каждая таблица связана с другими через единый идентификатор клиента, что позволяет строить интегрированные вектора признаков. Кроме того, данные организованы так, чтобы быть пригодными как для классического обучения, так и для обучения с подкреплением, где каждое действие агента вызывает реакцию в среде, смоделированной на базе этих данных.

Для оценки качества и пригодности сгенерированного массива была проведена серия визуализаций, направленных на верификацию поведенческой правдоподобности и достаточной гетерогенности выборки. В частности, по данным транзакций построены рейтинги клиентов по числу операций, средним расходам и объёму инвестиций. Эти графики продемонстрировали наличие «длинного хвоста» в распределении активности, что характерно для реальных банковских выборок. Также выделены явные различия между клиентами, что создаёт хорошую основу для кластеризации и дальнейшей персонализации рекомендаций.

Особенно важным оказалось то, что в синтетической базе удаётся смоделировать не только типичное, но и нетипичное поведение: единичные крупные переводы, резкие скачки расходов, досрочные закрытия вкладов. Это позволяет использовать датасет для обучения моделей детекции аномалий (например, Isolation Forest и AutoEncoder), а также для генерации сценариев, в которых RL-агенты должны принимать решения в условиях неопределённости и риска.

Анализ распределений по транзакциям, инвестициям и количеству событий подтвердил: данные обладают реалистичной неоднородностью и допускают построение поведенческих кластеров. Это критически важно, поскольку обучение модели на таких кластерах позволяет учесть разнообразие стратегий пользователей, а не усреднённый портрет. Модели RL, использующие такие данные как среду, могут взаимодействовать с «реалистично» ведущими себя агентами и тестировать гипотезы в условиях, приближенных к настоящим.

Обучение с подкреплением (RL) применяется для динамической адаптации рекомендаций:

- Агент выбирает действия (рекомендации), получая вознаграждение
- Среда — поведение клиента и макроэкономические условия
- Цель — максимизация долгосрочной полезности (NPV, ROI, удержание)

Модели:

- Q-learning, Deep Q Networks (DQN)
- Actor-Critic, PPO
- Multi-Agent RL (маркетинг, кредитование)

Преимущества RL:

- адаптация к изменению предпочтений;
- персонализация;
- обучение без необходимости ручной разметки.

Сценарии применения:

- рекомендация инвестиционных стратегий в условиях инфляции
- подбор кредитного лимита при нестабильном доходе
- рекомендации по снижению расходов в период спада

Для соответствия требованиям регуляторов и обеспечения доверия к модели необходима интерпретация предсказаний. Используются:

- SHAP, LIME
- Feature importance для деревьев и лесов
- визуализация ошибок автоэнкодера

Метрики качества:

- Precision / Recall / F1-score
- ROC-AUC, PR-AUC
- Средние потери, false positive rate

## Формирование и роль синтетических данных в обучении рекомендательной системы

На этапе подготовки обучающей выборки особое внимание было уделено оценке реалистичности и репрезентативности синтетически сгенерированных дан-

ных. Поскольку модель рекомендательной системы ориентирована на поведенческий анализ и адаптацию под различающиеся сценарии клиентской активности, критически важной задачей стало обеспечение достаточной структурной неоднородности и вариативности данных. Это требование обусловлено как спецификой банковского сектора, так и характером применяемых методов машинного обучения, в том числе нейросетевых архитектур и моделей обучения с подкреплением.

Для верификации качества сгенерированных данных был проведён визуальный анализ ключевых параметров клиентского поведения. На рисунке 2 приведены диаграммы, отражающие распределение различных метрик, по которым оценивается состоятельность модели и возможность извлечения значимых признаков.

1. Распределение по числу транзакций

Демонстрирует естественную дисперсию активности клиентов: от минимального количества операций до десятков в месяц. Такая картина отражает реальные особенности транзакционного поведения, включая наличие «гиперактивных» клиентов и пассивных пользователей. Выявление длинного хвоста подтверждает применимость модели к задачам выявления аномалий.

2. Средние суммы транзакций

Варьируются в широком диапазоне — от ~20 000 до свыше 45 000 рублей. Эти различия важны для стратификации клиентов по уровню платежеспособности, что позволяет рекомендательной системе учитывать индивидуальные возможности при формировании предложений.

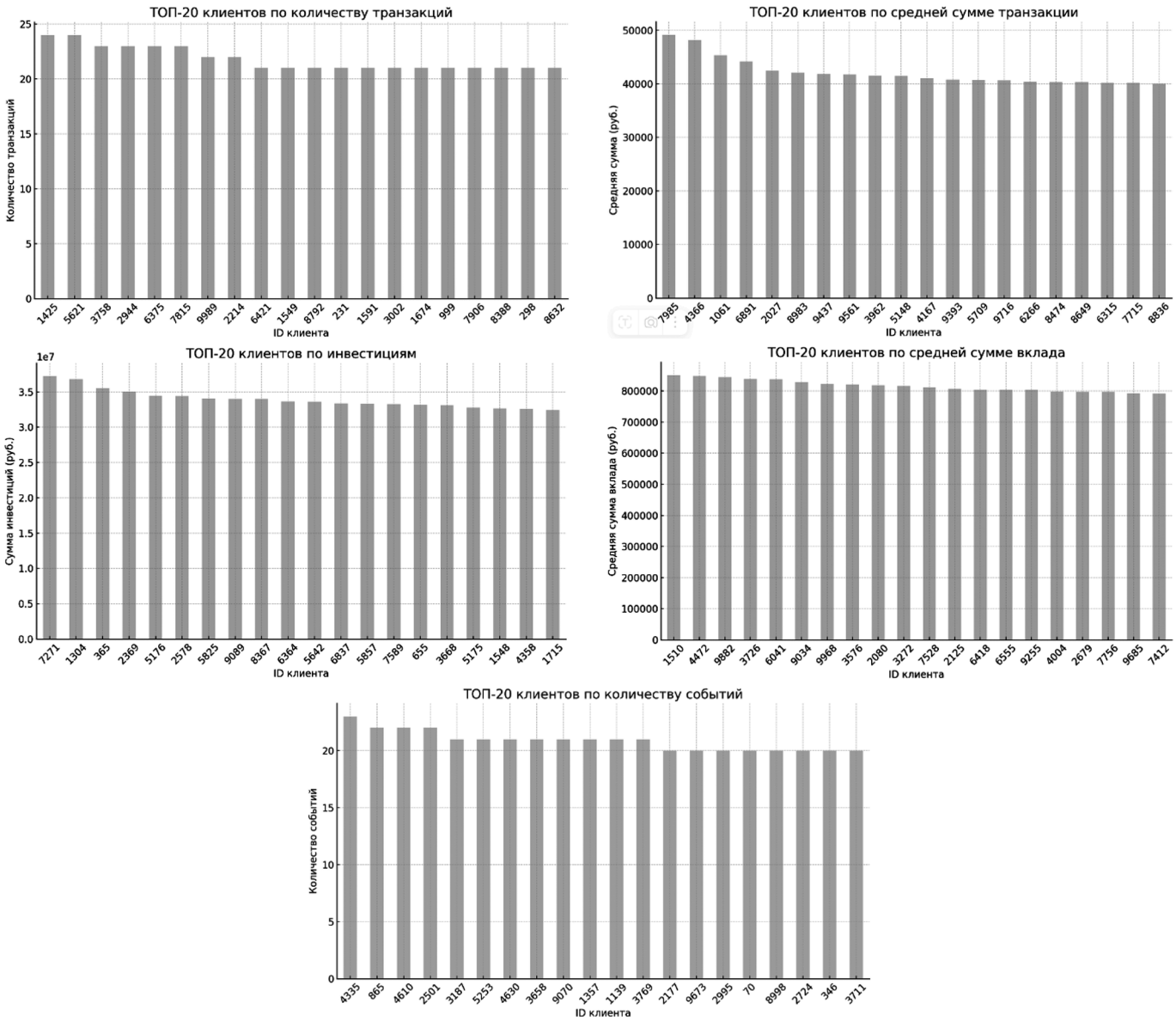


Рис. 2. Анализ качества синтетических данных

3. Суммарные инвестиции

Разброс инвестированных сумм по клиентам достигает 1,5 млн рублей, что указывает на успешное моделирование инвестиционного поведения. Данные по инвестициям формируют основу для оценки склонности к риску и составления инвестиционных стратегий.

4. Средние суммы вкладов

Отражают широкий диапазон сберегательной активности: от краткосрочных и небольших вкладов до значительных сбережений, характерных для консервативных клиентов. Эти данные используются при построении профиля финансовой устойчивости.

5. Количество событий в истории клиента

Визуализация событий (изменения персональных данных, открытия/закрытия продуктов и пр.) позволяет дополнительно сегментировать клиентов по уровню вовлечённости и активности вне транзакционного контура. Эти признаки важны для построения контекстных рекомендаций и выявления изменений в жизненном цикле клиента.

Обобщённо, визуальный анализ подтверждает:

- наличие широкой поведенческой дифференциации клиентов;
- присутствие отчётливо сегментированных кластеров, пригодных для задач кластеризации и таргетированной рекомендации;
- соответствие синтетических данных реальным бизнес-сценариям и статистическим закономерностям.

Эмпирическая проверка подтверждает, что структура, охват и содержательное наполнение синтетической базы данных соответствуют требованиям к обучающим выборкам в задачах рекомендательных и адаптивных систем. Это позволяет использовать её как полноценную среду для построения и тестирования модели в условиях, приближённых к реальной банковской практике.

## Заключение

В данной работе была представлена методология создания самообучающейся рекомендательной системы для банковского сектора, основанной на генерации синтетических данных с учётом поведенческих характеристик клиентов и макроэкономических факторов. Разработанная мультиагентная архитектура моделирования клиентского поведения позволила сформировать репрезентативную обучающую среду, отражающую разнообразие стратегий пользователей в условиях экономической нестабильности.

Синтетические данные, сгенерированные на основе стохастических процессов, языковых моделей и генеративно-состязательных сетей, продемонстрировали пригодность для обучения как традиционных моделей, так и алгоритмов обучения с подкреплением. Их внутренняя структурная логика, реалистичное распределение признаков и взаимосвязь на уровне клиентских идентификаторов обеспечили возможность построения комплексных признаковых пространств и реализации адаптивной логики рекомендаций.

Интеграция макроэкономических индикаторов в синтетическую среду открывает дополнительные перспективы для тестирования и настройки интеллектуальных моделей в стрессовых сценариях, а также для прогнозирования реакции клиентов на внешние изменения. Результаты визуального и статистического анализа подтвердили поведенческую правдоподобность сгенерированной выборки и её потенциал в качестве безопасной и масштабируемой альтернативы реальным банковским данным.

ЛИТЕРАТУРА

1. Wu Y., Li J. A Comprehensive Survey on Financial Recommendation Systems: Models, Techniques and Applications // IEEE Transactions on Knowledge and Data Engineering. — 2025. — Vol. 37, № 3. — P. 469–487. — DOI: 10.1109/TKDE.2025.1234567.
2. Arias M., Álvarez D., Fernández J. Flexible Recommender System for Debt Collection Management Based on Deep Reinforcement Learning // Expert Systems with Applications. — 2020. — Vol. 140. — Article 112878. — DOI: 10.1016/j.eswa.2019.112878.
3. Zuo X., Jiang A.A., Zhou K. Reinforcement Prompting for Financial Synthetic Data Generation // Journal of Finance and Data Science. — 2024. — Vol. 10. — Article 100137. — DOI: 10.1016/j.jfds.2024.100137.
4. Karst F.S., Li M.M., Leimeister J.M. SynDEc: A Synthetic Data Ecosystem // Electronic Markets. — 2025. — Vol. 35. — Article 7. — DOI: 10.1007/s12525-024-00746-8.
5. Lin Y., Liu Y., Lin F., et al. A Survey on Reinforcement Learning for Recommender Systems // IEEE Transactions on Neural Networks and Learning Systems. — 2024. — № 10. — P. 13164–13184. — DOI: 10.1109/TNNLS.2023.3280161.
6. Chen X., Yao L., McAuley J., et al. Deep Reinforcement Learning in Recommender Systems: A Survey and New Perspectives // Knowledge-Based Systems. — 2023. — Vol. 264. — Article 110335. — DOI: 10.1016/j.knosys.2023.110335.
7. Afsar M.M., Crump T., Far B. Reinforcement Learning Based Recommender Systems: A Survey // ACM Computing Surveys. — 2022. — Vol. 55, No. 7. — P. 1–38. — DOI: 10.1145/3543846.

8. Wiese M., Knobloch R., Korn R., Kretschmer P. Quant GANs: Deep Generation of Financial Time Series // Quantitative Finance. — 2020. — Vol. 20, No. 9. — P. 1419–1440. — DOI: 10.1080/14697688.2020.1730426.
9. Sarin S., Singh S.K., Kumar S., et al. Unleashing the Power of Multi-Agent Reinforcement Learning for Algorithmic Trading in the Digital Financial Frontier // Computers, Materials & Continua. — 2024. — Vol. 80, No. 2. — P. 3123–3138. — DOI: 10.32604/cmc.2024.051599.
10. Jiang Y., Olmo J., Atwi M. Deep Reinforcement Learning for Portfolio Selection // Global Finance Journal. — 2024. — Vol. 62. — Article 101016. — DOI: 10.1016/j.gfj.2024.101016.
11. Pang G., Shen C., Cao L., van den Hengel A. Deep Learning for Anomaly Detection: A Review // ACM Computing Surveys. — 2021. — Vol. 54, No. 2. — Article 38. — DOI: 10.1145/3439950.
12. Arshad K., Ali R.F., Muneer A., et al. Deep Reinforcement Learning for Anomaly Detection: A Systematic Review // IEEE Access. — 2022. — Vol. 10. — P. 124017–124035. — DOI: 10.1109/ACCESS.2022.3224023.


Журнал «Современная наука: актуальные проблемы теории и практики»