AI-Based Customer Behavior Prediction in Digital Banking and FinTech Platforms

Prem Pandeya, Preeti Malik
IMS University, Dehradun, U.K. India

Abstract

Digital banking and FinTech platforms produce detailed behavioral data from every login, tap, payment, transfer, and support interaction. Artificial intelligence (AI) lets institutions turn this data into predictions of what customers will do next: whether they will leave, which products they will adopt, how they will repay credit, how valuable they will become, and when they may need help. This review synthesizes research on AI-based customer behavior prediction in digital financial services, organized around three themes: (1) the predictive techniques used for churn, customer lifetime value, product propensity, credit and repayment behavior, and engagement, from ensemble learning to sequence models, recommender systems, and uplift modeling; (2) the operational and architectural foundations that turn predictions into timely action, including configurable workflows, metadata-driven platforms, zero-downtime model deployment, and reliable digital channels; and (3) the growing use of large language models (LLMs) in customer-facing banking services, together with the prompt engineering, cost management, and guardrails they require. The evidence shows that machine learning (ML) models, especially tree-based ensembles and models trained on fine-grained transaction and interaction data, predict customer behavior more accurately than traditional segmentation and regression approaches. Their business value, however, depends on acting on the right customers, adapting to changing behavior, and meeting expectations for privacy, fairness, transparency, and safety. The review proposes a layered framework for AI-driven customer behavior management and identifies directions for future research.

Keywords: customer behavior prediction, digital banking, FinTech, churn prediction, customer lifetime value, recommender systems, uplift modeling, machine learning, large language models, guardrails, conversational banking

DOI: 10.64235/ver0gd20

1. Introduction

Banking has moved from branches to screens. Customers now open accounts, apply for loans, make payments, invest, and get support through mobile apps and web platforms, often without ever speaking to a person. FinTech firms such as digital-only banks, payment providers, lending platforms, and investment apps have raised expectations further by

offering fast, personalized, and low-cost services (Gomber et al., 2018; Philippon, 2016). In this environment, understanding and anticipating customer behavior has become a core competitive capability.

Switching costs in digital finance are low. A customer can open an account with a competitor in minutes, and many hold accounts with several providers at once. Retention, engagement, and cross-selling therefore depend on recognizing customer needs and risks early. Customer relationship management (CRM) research has long shown that keeping existing customers is usually more profitable than acquiring new ones (Reichheld & Sasser, 1990), and that data mining can support customer identification, attraction, retention, and development (Ngai et al., 2009).

What has changed is the richness of the data. Digital platforms record not only what customers buy and how much they hold, but how they behave: how often they log in, which features they use, where they abandon a process, how their spending changes, and what they ask support agents. Martens et al. (2016) showed that fine-grained payment data from a bank significantly improved predictive targeting compared with traditional socio-demographic data. Berg et al. (2020) found that simple digital footprint variables predicted consumer default as well as credit bureau scores. AI and ML methods are what make such high-dimensional, sparse, and sequential data usable.

Prediction is only the first step. To create value, a prediction must reach the right system at the right time and trigger an appropriate action. It must also be refreshed as behavior changes and delivered through channels that work reliably. This makes the enterprise architecture around the model as important as the model itself, including configurable workflows (Basireddy, 2022b), metadata-driven platforms (Basireddy, 2022a; Asuri, R. (2020); Asuri, R. (2022)), safe model deployment (Sannidhanam, 2023), and reliable digital interfaces (Cherukuri, 2024). At the same time, LLMs are changing how banks interact with customers through conversational assistants and personalized communication. This raises new questions of reliability (Sannidhanam, 2024a), cost (Cherukuri, 2025), and safety (Sannidhanam, 2024b).

This review has four objectives:

1. To synthesize AI and ML techniques used to predict customer behavior in digital banking and FinTech.
2. To examine the architectural and operational foundations that turn behavioral predictions into timely, reliable action.
3. To assess the role of LLMs in customer-facing financial services and the controls they require.
4. To identify challenges and directions for future research.

2. Review Methodology

2.1 Review design

The review follows a structured, PRISMA-informed approach (Page et al., 2021) combined with thematic synthesis. This suits a literature that spans marketing science, information systems, financial technology, and ML engineering.

2.2 Research questions

- **RQ1:** Which AI and ML techniques are used to predict customer behavior in digital banking and FinTech, and for which prediction tasks?
- **RQ2:** What architectural and operational capabilities are needed to turn behavioral predictions into effective action?
- **RQ3:** How are LLMs being applied to customer interaction in financial services, and what controls do they require?
- **RQ4:** What challenges and research gaps remain?

2.3 Inclusion criteria

Studies were included if they (a) proposed or evaluated AI/ML methods for predicting customer behavior in banking, financial services, or closely comparable subscription-based services, (b) addressed architectures, workflows, deployment, or digital channel quality that enable predictive customer analytics in production, or (c) addressed the reliability, cost, or safety of LLMs in enterprise customer-facing applications. Opinion pieces without methodological content were excluded.

2.4 Data extraction and synthesis

For each study, the review recorded the prediction task, the data used, the technique, the function addressed (prediction, operationalization, or customer interaction), and the reported strengths and limitations. Findings were grouped into the themes in Sections 4 to 6.

> *Note on reproducibility:* Researchers extending this review should report database names, search strings, search dates, and record counts at each screening stage in a PRISMA 2020 flow diagram (Page et al., 2021).

3. Background: Customer Behavior in Digital Financial Services

3.1 Prediction tasks

Customer behavior prediction in digital banking covers several related tasks, summarized in Table 1.

Table 1. Main customer behavior prediction tasks in digital banking and FinTech

Prediction task	Question answered	Typical business use	Key data
Churn and attrition	Will this customer leave or become inactive?	Retention campaigns, service recovery	Balances, activity trends, complaints, app usage
Customer lifetime value	How valuable will this customer be over time?	Acquisition spend, service tiering, resource allocation	Transaction history, product holdings, tenure
Product propensity	Which product is this customer likely to adopt next?	Cross-sell, next-best-offer, personalization	Holdings, life events, spending categories
Credit and repayment behavior	Will this customer repay or fall behind?	Credit limits, early intervention, collections	Transactions, cash flow, digital footprint
Engagement and usage	How will this customer use the app or platform?	Feature design, onboarding, nudges	Clickstream, session logs, feature usage
Financial vulnerability	Is this customer heading toward financial difficulty?	Proactive support, hardship programs	Income shifts, overdraft use, missed payments

3.2 Data sources

Digital platforms combine several kinds of behavioral data. **Transactional data** records payments, transfers, deposits, and card spending, and shows income patterns, spending categories, and life events. **Interaction data** from app and web sessions records logins, navigation paths, feature use, and abandoned processes. **Service data** covers support tickets, chat logs, and complaints. **Product and account data** covers holdings, balances, limits, and tenure. **External and alternative data** may include credit bureau records and, where permitted, digital footprint signals such as device type (Berg et al., 2020). Martens et al. (2016) showed that the value of such data often comes from its fine granularity, meaning the specific merchants and counterparties a customer transacts with, rather than from aggregated summaries.

3.3 From segmentation to prediction

Traditional customer analytics in banking relied on static segmentation by demographics, wealth, or product holdings, and on regression-based scoring. These approaches are easy to explain but treat customers within a segment as alike and are updated infrequently. AI-based prediction instead estimates individual probabilities and values, updates them as new behavior arrives, and uses far more signals (Shmueli & Koppius, 2011; Asuri, R. (2020); Asuri, R. (2022)).

4. Findings: Techniques for Predicting Customer Behavior

4.1 Churn and attrition prediction

Churn prediction is the most widely studied task. In financial services, Van den Poel and Larivière (2004) used proportional hazard models on data from a European bank to identify the drivers and timing of customer attrition, showing the value of modeling when customers leave as well as whether they leave. Larivière and Van den Poel (2005) found that random forests outperformed linear and logistic regression for predicting customer retention and profitability in financial services. Xie et al. (2009) proposed improved balanced random forests for churn prediction on bank customer data, addressing the imbalance between the small number of churners and the large number of retained customers.

Neslin et al. (2006) compared churn models from many research teams and found that methodological choices such as estimation technique, variable selection, and time allocated to modeling had a significant effect on predictive accuracy and on the profitability of retention campaigns. Verbeke et al. (2012) argued that churn models should be evaluated by the profit they generate rather than by accuracy alone. They introduced a profit-based performance measure and showed that the best model by statistical measures is not always the most profitable. Class imbalance is a recurring issue and is commonly addressed through resampling methods such as SMOTE (Chawla et al., 2002) and cost-sensitive learning.

4.2 Customer lifetime value

Customer lifetime value (CLV) prediction estimates the future value of a customer relationship. Probabilistic models developed in marketing science, such as the Pareto/NBD model (Schmittlein et al., 1987; Asuri, R. (2022)) and its simpler BG/NBD variant (Fader et al., 2005), estimate whether a customer is still active and how many future transactions to expect from their past purchase frequency and recency. Gupta et al. (2006) reviewed CLV modeling approaches, including probabilistic, econometric, and ML methods. Kumar and Reinartz (2016) argue that customer value should be considered in both directions, meaning the value the firm delivers to customers and the value customers deliver to the firm. In digital banking, CLV predictions guide acquisition spending, service tiers, and where retention effort is most worthwhile (Asuri, R. (2020)).

4.3 Product propensity and recommendation

Predicting which product a customer is likely to need next supports cross-selling and personalization. Propensity models, often built as gradient-boosted classifiers (Chen & Guestrin, 2016) or random forests (Breiman, 2001), estimate the likelihood that each customer will adopt each product. Recommender system techniques extend this idea. Matrix

factorization (Koren et al., 2009) learns latent preferences from patterns of customer-product interaction, and deep learning recommenders (Covington et al., 2016) can combine many behavioral and contextual signals at scale. Martens et al. (2016) showed in a banking context that fine-grained transaction data improved targeting of product offers. Financial products differ from media or retail items in important ways, however. They are adopted infrequently, carry long-term consequences for customers, and are subject to suitability and fair-treatment rules. Recommendations therefore need constraints that ordinary recommender systems do not.

4.4 Credit and repayment behavior

Behavioral data also predicts how customers will handle credit. Khandani et al. (2010) combined transaction and credit bureau data for a large bank's customers and forecast credit card delinquencies months ahead, enabling earlier and better-targeted risk management. Berg et al. (2020) showed that digital footprints, such as the device used and the type of email address, carried predictive information about default comparable to credit bureau scores. Such models let FinTech lenders assess customers with limited credit histories. They also raise concerns about fairness and privacy (Section 7.3).

4.5 Sequence models and deep learning

Customer behavior unfolds over time, and the order of events often carries meaning. A series of declining logins, falling balances, and a large outbound transfer may signal imminent churn more clearly than any single event. Recurrent neural networks, particularly long short-term memory (LSTM) networks (Hochreiter & Schmidhuber, 1997), can model these sequences directly from transaction and clickstream data. Deep learning models are especially useful when data are high-dimensional and sequential. They are, however, more data-hungry and harder to interpret than tree-based models, which remain the dominant choice for tabular banking data.

4.6 Uplift and causal modeling

Predicting who is likely to leave is not the same as knowing whom to target. Ascarza (2018) showed through field experiments that targeting customers with the highest churn risk can be ineffective, because many high-risk customers do not respond to retention offers. Targeting customers whose behavior is most likely to change because of the intervention produced better results. Uplift modeling estimates this incremental effect by modeling the difference between outcomes with and without treatment (Gutierrez & Gérardy, 2017; Asuri, R. (2022)). For digital banks running retention campaigns, product offers, or financial wellness nudges, uplift modeling moves the question from "who will act?" to "who will act differently because of what we do?"

4.7 Text and conversational data

Complaints, chat transcripts, app reviews, and support calls hold rich signals about dissatisfaction and intent. Natural language processing can extract sentiment, topics, and intent from these sources and feed them into churn and engagement models. Finance-specific language models (Wu et al., 2023) and general LLMs make this analysis more accurate and easier to deploy, though it requires care with sensitive customer information.

4.8 Synthesis for RQ1

Table 2. AI techniques for customer behavior prediction in digital banking

Task	Dominant techniques	Strengths	Limitations	Representative studies
Churn	Random forests, gradient boosting, survival models	Accurate, handles imbalance with adjustments	Accuracy does not guarantee profitable action	Larivière & Van den Poel (2005); Verbeke et al. (2012)
Lifetime value	Pareto/NBD, BG/NBD, ML regression	Estimates long-term value from simple inputs	Assumptions may not fit banking behavior	Fader et al. (2005); Gupta et al. (2006)
Product propensity	Gradient boosting, matrix factorization, deep recommenders	Personalizes offers at scale	Suitability and regulatory constraints	Koren et al. (2009); Martens et al. (2016)
Credit behavior	Ensembles on transaction and footprint data	Early warning, credit for thin-file customers	Fairness and privacy concerns	Khandani et al. (2010); Berg et al. (2020)
Sequential behavior	LSTM and recurrent networks	Captures order and timing of events	Data-hungry and opaque	Hochreiter & Schmidhuber (1997)
Intervention targeting	Uplift and causal models	Targets customers who will respond	Needs experimental or treatment data	Ascarza (2018); Gutierrez & Gérardy (2017)

Across tasks, ML methods generally predict customer behavior better than static segmentation and linear models, especially when fine-grained behavioral data are available. The literature also stresses that prediction quality should be judged by the value of the resulting actions, not by statistical accuracy alone (Verbeke et al., 2012; Ascarza, 2018).

5. Findings: Turning Predictions into Action

5.1 Workflow integration

A churn score or product propensity has value only if it reaches the right process. Basireddy (2022b) examines **digitizing risk management through configurable enterprise workflow architecture**. In this approach, steps for identification, assessment, escalation, approval, and remediation are defined as configurable process definitions rather than hard-coded logic.

Customer behavior management needs the same capability. A high churn-risk flag may route a customer to a relationship manager or trigger a retention offer. An early sign of financial difficulty may open a proactive support case, and a propensity signal may schedule a personalized message. Behavioral predictions about credit risk and vulnerability fall squarely within risk management, so configurable workflows with audit trails ensure these interventions are consistent, traceable, and compliant. When strategies change, workflows can be reconfigured without rebuilding the underlying systems.

5.2 Metadata-driven personalization and agility

Digital banks change products, pricing, offers, and customer journeys often. If the rules connecting predictions to actions are hard-coded, each change needs development effort, and personalization falls behind the business. Basireddy (2022a) proposes a **metadata-centric platform architecture** for dynamic configuration, intelligent process automation, and operational agility. In this architecture, business rules, process behaviors, and configurations are held as metadata that can be changed without code changes. For customer behavior prediction, this allows segment definitions, eligibility rules, offer mappings, score thresholds, and journey logic to be adjusted quickly in response to model outputs, campaign results, or regulatory changes. It separates what the model predicts from how the business chooses to act on it, so each can evolve independently.

5.3 Real-time serving and zero-downtime model updates

Many customer interactions happen in real time. Personalizing an app home screen, making an in-session offer, or flagging a risky transfer requires predictions within milliseconds. Customer behavior also shifts with economic conditions, seasons, product launches, and competitor actions, so models must be retrained regularly to avoid concept drift (Gama et al., 2014). Sannidhanam (2023) addresses **zero-downtime AI model updates in real-time inference systems**, describing deployment practices that let new model versions replace old ones without interrupting service. Typical practices of this kind include shadow deployment, gradual rollout, instant switching between versions, and automatic rollback when performance degrades. For digital banks, these practices mean behavior models can be refreshed as often as needed without disrupting the customer experience. Sculley et al. (2015) warn that production ML systems accumulate hidden technical debt in their data pipelines and configuration, which makes disciplined deployment practices all the more important.

5.4 Reliable digital channels and data quality

Customer behavior prediction depends on the quality of the digital channel in two ways. First, behavioral data are only as reliable as the interfaces that generate them. Broken pages, inconsistent layouts, or rendering errors distort clickstream data and can make customers

appear disengaged when they are actually frustrated by defects. Second, the predictions themselves are delivered through these interfaces as personalized content, offers, and alerts, so errors in presentation weaken or undo their effect.

Cherukuri (2024) presents **Webchecker, a plugin based on the Epsilon Validation Language (EVL) for validating HTML pages built with Bootstrap frameworks**. Automated validation of web pages against framework rules helps keep interfaces consistent and correct across releases. This becomes more important as personalization increases the number of page variants that different customers see. For digital banking platforms, automated front-end validation supports both a dependable customer experience and cleaner behavioral data for predictive models.

5.5 Explainability and transparency

Banks must often explain decisions that affect customers, especially in credit, and they need to understand why models flag customers before acting. Post-hoc explanation methods such as SHAP (Lundberg & Lee, 2017) identify which behaviors drove a prediction, such as a drop in salary deposits or reduced app usage. This helps staff choose an appropriate response and helps compliance teams review model behavior. Rudin (2019) argues that in high-stakes settings inherently interpretable models should be preferred where they perform comparably. Customer-facing decisions such as credit limit changes are one such setting.

6. Findings: LLMs in Customer-Facing Banking

6.1 Emerging roles

LLMs are changing how banks and FinTech platforms communicate with customers. Earlier chatbots relied on scripted flows and intent classification (Adamopoulou & Moussiades, 2020; Asuri, R. (2020)). LLM-based assistants can understand open-ended questions, explain products and transactions, help customers complete tasks, and adapt their tone and content to the individual. LLMs also support customer behavior analytics by summarizing interaction histories, extracting intent and sentiment from conversations, and drafting personalized messages informed by predictive models. Retrieval-augmented generation (Lewis et al., 2020) lets assistants ground answers in the bank's own product documentation, policies, and the customer's account data.

These capabilities carry real risks. LLMs can generate plausible but incorrect statements (Ji et al., 2023). In banking, a wrong answer about fees, interest rates, eligibility, or account status can mislead customers, cause financial harm, and breach consumer protection rules.

6.2 Prompt engineering for reliability

Sannidhanam (2024a) sets out **prompt engineering patterns for reliable LLM outputs in production environments**, including clear role and task instructions, structured output formats, grounding in supplied context, explicit constraints, and validating outputs before use. Techniques such as chain-of-thought prompting can improve reasoning on complex tasks (Wei et al., 2022). In customer-facing banking, these patterns help keep responses accurate, consistent with bank policy, and tied to verified account information. They also help keep personalized messages generated from behavioral predictions appropriate and compliant.

6.3 Cost and complexity

Digital banks may handle millions of customer interactions a month, so the cost and latency of LLM calls matter. Cherukuri (2025) examines the **cost, complexity, and efficacy of prompt engineering techniques** for LLMs and shows that more elaborate prompting strategies do not always yield proportionate gains in output quality, while they can add significant cost and latency. For FinTech platforms, this suggests matching technique to task. Simple, well-structured prompts may be enough for routine queries, while more elaborate reasoning strategies are reserved for complex or high-stakes interactions where the gain in accuracy justifies the cost.

6.4 Guardrails and safety

Sannidhanam (2024b) examines **guardrails and safety mechanisms for LLM-powered enterprise applications**. These include input filtering, output validation, restrictions on topics and actions, protection of sensitive data, and escalation to human agents. In banking, guardrails can stop an assistant from giving unauthorized financial advice, disclosing another customer's data, confirming transactions without proper authentication, or responding to manipulative prompt injection. They can also make sure that customers showing signs of vulnerability or distress, which behavioral models may detect, are handed to trained staff rather than handled only by an automated system.

7. Discussion

7.1 Answering the research questions

RQ1 (techniques). Tree-based ensembles are the most consistently effective approach for churn, propensity, and credit behavior prediction on tabular banking data. Probabilistic models remain useful for lifetime value, sequence models capture the timing of behavior, and uplift models improve targeting of interventions. Fine-grained transaction and interaction data add substantial predictive value.

RQ2 (operationalization). Predictions create value when they flow into configurable workflows (Basireddy, 2022b), are acted on through metadata-driven rules that can change quickly (Basireddy, 2022a), are served and refreshed without service disruption (Sannidhanam, 2023), and are delivered through reliable, validated digital channels (Cherukuri, 2024).

RQ3 (LLMs). LLMs are becoming the main interface for customer interaction and a tool for behavioral analysis. Their safe use depends on reliable prompting (Sannidhanam, 2024a), cost-aware design (Cherukuri, 2025), and robust guardrails (Sannidhanam, 2024b).

RQ4 (challenges). These are set out in Section 7.3.

7.2 An integrated framework

The findings point to a layered framework for AI-driven customer behavior management, summarized in Table 3.

Table 3. Layered framework for AI-based customer behavior prediction and action

Layer	Function	Key enablers	Supporting studies
Data and channels	Capture accurate behavioral data through reliable interfaces	Transaction, interaction, and service data; validated web front ends	Martens et al. (2016); Cherukuri (2024)
Prediction	Estimate churn, value, propensity, and credit behavior	Ensembles, CLV models, sequence models, uplift models	Verbeke et al. (2012); Ascarza (2018); Khandani et al. (2010)
Model operations	Serve and refresh predictions safely	Real-time inference, zero-downtime updates, drift monitoring	Sannidhanam (2023); Gama et al. (2014)
Decision and action	Turn predictions into interventions	Configurable workflows, metadata-driven rules, explainability	Basireddy (2022a, 2022b); Lundberg & Lee (2017)
Customer interaction	Engage customers through personalized conversation	LLM assistants, reliable prompting, cost-aware design, guardrails	Sannidhanam (2024a, 2024b); Cherukuri (2025)

7.3 Challenges and limitations

1. **Prediction versus profitable action.** Accurate predictions do not guarantee effective interventions. Targeting high-risk customers can waste resources if they do not respond (Ascarza, 2018; Verbeke et al., 2012).
2. **Privacy and consent.** Behavioral data are intimate and revealing. Data protection laws such as the EU General Data Protection Regulation (European Parliament & Council of the European Union, 2016) limit how such data may be used and require transparency about automated decision-making.

3. **Fairness.** Models trained on historical behavior can reinforce existing inequalities, and alternative data may act as proxies for protected characteristics (Barocas & Selbst, 2016; Fuster et al., 2022).
4. **Concept drift.** Customer behavior changes with economic conditions and product changes, so models degrade unless they are monitored and retrained (Gama et al., 2014).
5. **Class imbalance.** Churn, default, and product adoption are often rare events, which makes models harder to train and evaluate (Chawla et al., 2002; Xie et al., 2009).
6. **Explainability.** Customers and regulators may require explanations for decisions affecting access to credit or services, which constrains the use of opaque models (Rudin, 2019).
7. **LLM reliability, cost, and safety.** Hallucination, prompt injection, data leakage, and high inference costs limit where LLMs can be used without human oversight (Ji et al., 2023; Cherukuri, 2025; Sannidhanam, 2024b).
8. **Manipulation risk.** Highly personalized nudges could push customers toward products that serve the institution more than the customer, raising ethical and consumer protection concerns.
9. **Operational complexity.** Maintaining data pipelines, many models, configuration, and channels creates technical debt that can undermine reliability (Sculley et al., 2015).

8. Future Research Directions

1. **Causal and uplift methods at scale.** Wider use of experimentation and causal ML to measure the incremental effect of interventions rather than raw propensity.
2. **Customer-centric outcomes.** Evaluation frameworks that measure customer financial well-being alongside institutional metrics such as retention and revenue.
3. **Privacy-preserving behavioral analytics.** Federated learning (Yang et al., 2019) and differential privacy (Dwork & Roth, 2014; Asuri, R. (2022)) to model behavior without centralizing sensitive raw data.
4. **Fairness-aware personalization.** Methods that detect and reduce disparate impact in offers, pricing, and credit decisions driven by behavioral models.
5. **Continuous learning systems.** Empirical studies of how frequent, zero-downtime retraining affects prediction quality and customer outcomes over time.

6. **Agile decisioning architectures.** Research on how metadata-driven and configurable workflow architectures shorten the time from new insight to deployed customer action.
7. **LLM evaluation in banking.** Benchmarks for accuracy, safety, cost, and customer satisfaction of LLM-based assistants, including the cost-efficacy trade-offs of prompting strategies.
8. **Channel quality and behavioral data.** Studies linking front-end quality and automated interface validation to the reliability of behavioral signals and model performance.

9. Conclusion

AI has become central to understanding and anticipating customer behavior in digital banking and FinTech. The literature shows that ML models built on fine-grained transaction and interaction data predict churn, lifetime value, product adoption, and credit behavior more accurately than traditional segmentation and regression. It also shows that the value of these predictions depends on what institutions do with them: targeting the customers whose behavior can actually be changed, routing predictions into configurable workflows, adapting rules quickly through metadata-driven platforms, refreshing models without disrupting service, and delivering actions through reliable digital channels.

LLMs are adding a conversational layer that can make banking more personal and accessible, but only if they are reliable, affordable at scale, and protected by strong guardrails. For practitioners, the main lesson is to invest in the full chain from data to action, not only in predictive accuracy. For researchers, the priorities are causal evaluation of interventions, privacy-preserving and fair personalization, continuous learning in production, and rigorous evaluation of LLMs in customer-facing financial services.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications, 2*, 100006.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research, 55*(1), 80–98.

Asuri, R. (2020). An Intelligent Enterprise Workflow Automation Framework for SAP ECC Standard Business Processes through the Integration of Robotic Process Automation and Machine Learning Techniques. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology*, *12*(02), 181-195. https://doi.org/10.18090/samriddhi.v12i02.17

Asuri, R. (2022). Leveraging SAP S/4HANA and SAP Analytics Cloud to Build an Intelligent Enterprise Framework for Real Time Decision Intelligence Digital Innovation and Sustainable Business Value Creation. International Journal of Science, Technology & Society, 6(1), 85-100. https://doi.org/10.18091/ijsts.v6i1-113

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732.

Basireddy, S. (2022a). Pioneering a future-ready metadata-centric platform architecture for dynamic configuration intelligent process automation operational agility and sustainable enterprise modernization. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 14*(03), 415–429. https://doi.org/10.18090/samriddhi.v14i03.27

Basireddy, S. R. (2022b). Digitizing risk management through configurable enterprise workflow architecture. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 14*(4), 231–239.

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. *The Review of Financial Studies, 33*(7), 2845–2897.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Cherukuri, M. (2024, November). Webchecker: A versatile EVL plugin for validating HTML pages with Bootstrap frameworks. In *International Conference on Big Data and Artificial Intelligence Applications* (pp. 285–293). Springer Nature Switzerland.

Cherukuri, M. (2025). Cost, complexity, and efficacy of prompt engineering techniques for large language models. *IJSAT – International Journal on Science and Technology, 16*(2).

Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 191–198).

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science, 9*(3–4), 211–407.

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). *Official Journal of the European Union, L 119*, 1–88.

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science, 24*(2), 275–284.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance, 77*(1), 5–47.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44.

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. *Journal of Management Information Systems, 35*(1), 220–265.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research, 9*(2), 139–155.

Gutierrez, P., & Gérardy, J.-Y. (2017). Causal inference and uplift modelling: A review of the literature. *Proceedings of Machine Learning Research, 67*, 1–13.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., … Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer, 42*(8), 30–37.

Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. *Journal of Marketing, 80*(6), 36–68.

Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications, 29*(2), 472–484.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

Martens, D., Provost, F., Clark, J., & de Fortuny, E. J. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS Quarterly, 40*(4), 869–888.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204–211.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592–2602.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71.

Philippon, T. (2016). *The FinTech opportunity* (NBER Working Paper No. 22476). National Bureau of Economic Research.

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review, 68*(5), 105–111.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Sannidhanam, A. H. (2023). Zero-downtime AI model updates in real-time inference systems. *International Journal of Engineering Science & Humanities, 13*(2), 70–78. https://www.ijesh.com/j/article/view/934

Sannidhanam, A. H. (2024a). Prompt engineering patterns for reliable LLM outputs in production environments. *Journal of Multidisciplinary Knowledge, 4*(1), 29–39. https://jmk.datatablets.com/index.php/j/article/view/168

Sannidhanam, A. H. (2024b). Guardrails and safety mechanisms for LLM-powered enterprise applications. *International Journal of Engineering Science & Humanities, 14*(3), 273–285. https://www.ijesh.com/j/article/view/933

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science, 33*(1), 1–24.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., … Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems, 28*, 2503–2511.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly, 35*(3), 553–572.

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196–217.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*(1), 211–229.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., … Mann, G. (2023). *BloombergGPT: A large language model for finance* (arXiv:2303.17564). arXiv.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 5445–5449.

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology, 10*(2), Article 12.