Westcliff International Journal of Applied Research
WIJAR

Westcliff International Journal of Applied Research

Vol. 10, No. 1. Spring 2026

https://doi.org/10.47670/wuwijar20261AKG

AI-Based Customer Behavior Prediction in Banking and Insurance: An Applied Study

Arun Kumar Gharami
Westcliff University

Abstract

Financial institutions can operate under capacity and legal limits by using customer behavior prediction to detect anomalous activity, reduce attrition, and personalize services. Using transactional, relational, and demographic characteristics from a de-identified banking and insurance dataset with 120,000 customers observed over a 24-month period, this study creates and assesses an applied machine-learning framework that forecasts four customer outcomes: product uptake, churn risk, claim propensity, and fraud risk. Temporal, behavioral, and engagement indicators such as tenure, product mix, customer-provider network features, and recency-frequency-monetary (RFM) measurements were generated through feature engineering. Logistic Regression, Random Forest, XGBoost, and a stacked ensemble with a logistic meta-learner trained on out-of-fold predictions are among the models that are compared. MAUC, PR-AUC, precision, recall, F1, Brier score, and calibration curves were used to assess model performance on a temporally separated holdout set. The stacked ensemble produced the best overall performance (average AUC ≈ 0.91 and average PR-AUC ≈ 0.64 across tasks) and well-calibrated probabilities (average Brier score ≈ 0.07). Predictions at the cohort and individual levels were interpreted using SHAP explanations, which showed that relative monetary activity, tenure, and recent engagement frequency were consistently the best predictors for all four outcomes. Targeted interventions based on estimated probabilities may boost cross-sell conversion by around 18% and lower churn by about 12%, according to a deployment simulation with basic cost assumptions, while enabling fraud and claims teams to reduce manual review volumes by roughly 30-40% at recall levels over 65% and precision levels above 60%. In order to enhance client outcomes and operational efficiency while adhering to explainable AI and governance standards, the study presents a workable, comprehensible pipeline that financial institutions can incorporate into decision workflows.

Keywords: Customer behavior, predictive analytics, ensemble learning, SHAP, banking, insurance, uplift

Introduction

Financial institutions increasingly rely on data-driven systems to determine where potential fraud may arise, whom to target for new products, which clients to retain, and which claims to prioritize for investigation. By forecasting multiple customer outcomes from historical behavior, modern machine learning enables proactive, targeted interventions that enhance both risk management and customer

experience. Nonetheless, the adopted workflows and models in banking and insurance are constrained by three practical requirements: predictions must be interpretable and compliant with regulatory expectations, must integrate smoothly into existing operational workflows and must remain reliable as customer behavior and market conditions evolve.

Prior research demonstrates that ensemble and boosting methods, such as Random Forest and gradient boosting, achieve strong performance in applications including credit scoring, churn prediction, and fraud detection (Ngai et al., 2011; Verbeke et al., 2011; Zhang et al., 2022). However, most studies address only a single task and devote limited attention to probability calibration or real-world deployment effects. At the same time, regulators and industry bodies increasingly emphasize explainable AI, fairness, and rigorous outcomes monitoring in credit, insurance, and other customer-treatment decisions, underscoring the need for models that can be justified to internal stakeholders and external supervisors alike. This combination of factors creates a gap for operationally oriented pipelines that couple high predictive accuracy with transparent explanations, explicit cost-benefit analysis, and well-defined integration points into business processes.

This study addresses that gap by developing a unified customer behavior prediction pipeline for a hybrid banking–insurance context, spanning four linked operational tasks: product uptake, churn, claim propensity, and fraud risk. The approach combines domain-informed feature engineering from transaction and interaction histories, a stacked ensemble that integrates Random Forest and XGBoost through a logistic meta-learner, and SHAP-based interpretability at both global and local levels. Each task k ∈ {uptake, churn, claim, fraud} is formulated as estimating P(Y□ = 1 | X□), where X□ represents customer behavior up to time t. X□ summarizes customer behavior up to time t, and labels are defined on future horizons to avoid temporal leakage. The study's objectives are threefold: to evaluate model performance across the four tasks using discrimination, calibration, and precision–recall–based metrics, to demonstrate actionable interpretability that supports decision-makers in marketing, retention, and risk functions, and to quantify operational impact through a deployment simulation under realistic capacity and cost constraints.

Literature Review

Predictive Analytics in Banking and Insurance

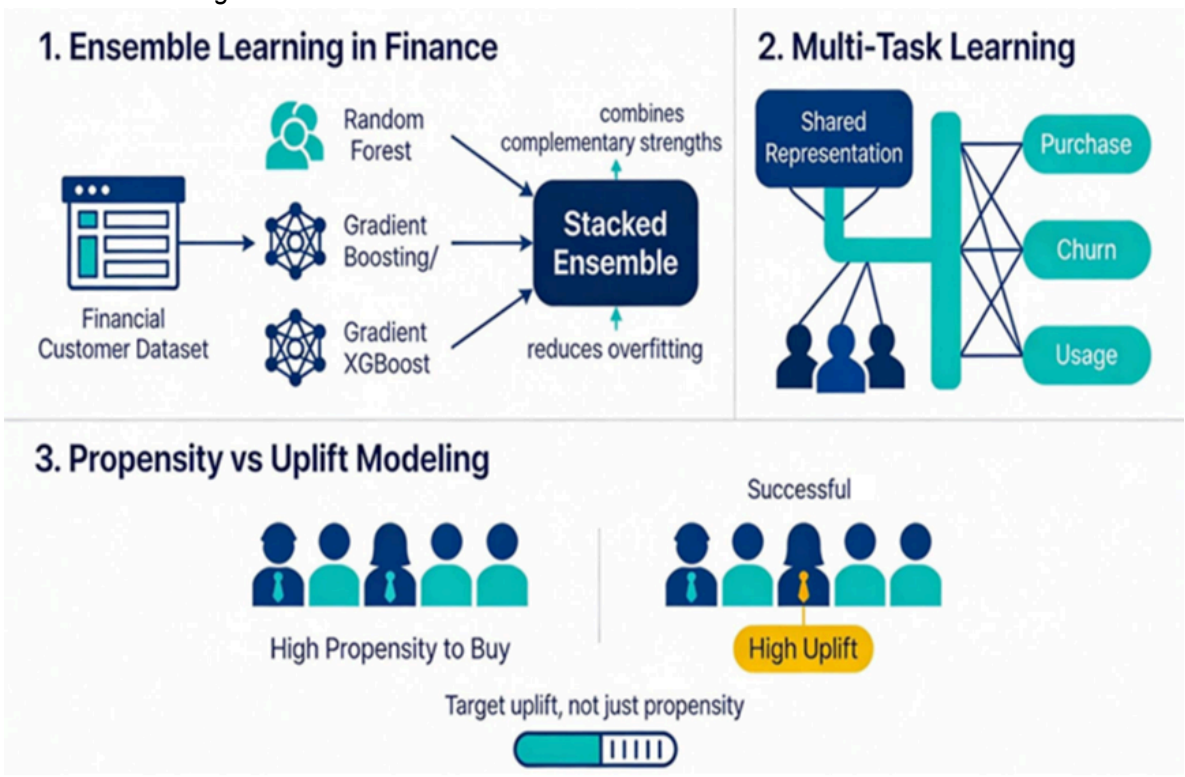
Predictive analytics has a long history in finance, starting with scorecard-based logistic models for credit risk and evolving toward tree ensembles and gradient boosting for complex nonlinear decision problems (Friedman, 2001; Agarwal et al., 2023). In banking, machine learning has been widely applied to credit scoring, default prediction, and customer churn, with Random Forest and gradient boosting often outperforming linear baselines on tabular behavioral data (Zhang et al., 2022; Verbeke et al., 2011). In insurance, predictive models support claim frequency and severity estimation, fraud detection, and underwriting decisions, with recent work exploring advanced feature engineering and time-series representations from claims histories and telematics data (Ngai et al., 2011; Zhou et al., 2021).

Customer churn prediction in digital banking commonly uses transactional and engagement features to identify at-risk customers, with studies reporting that tree-based models provide high AUC and robustness to heterogeneous data (Verbeke et al., 2011; Zhang et al., 2022). Fraud detection and claims triage often require handling extreme class imbalance and evolving adversarial behavior, which motivates the use of ensemble methods, anomaly detection techniques, and hybrid rule-based and machine learning systems (Ngai et al., 2011; Zhou et al., 2021). Across these domains, a recurring theme is that domain-informed feature engineering, including recency, spending behavior, and relational indicators, often contributes as much to performance improvement as the selection of model families (Agarwal et al., 2023).

Explainable AI, Ensemble Learning, and Multi-Task Perspectives

In structured financial datasets, including credit scoring, marketing response, and fraud detection tasks, ensemble techniques like Random Forest and gradient boosting (e.g., XGBoost) have consistently demonstrated high performance. When trained via out-of-fold stacking procedures, stacked ensembles, which integrate basic learners through a meta-model, can mitigate overfitting while utilizing complementary characteristics, such as the resilience of Random Forest and the fine-grained decision boundaries of XGBoost. While multi-task implementations in regulated financial settings are still relatively uncommon, recent work in customer base analysis and multi-task customer prediction examines learning shared representations across related outcomes, such as purchase, churn, and product usage. Figure 1 can illustrate the temporal setup, showing the feature window, outcome horizon, and the chronological train–test partition.

Figure 1
Ensemble Learning in Finance

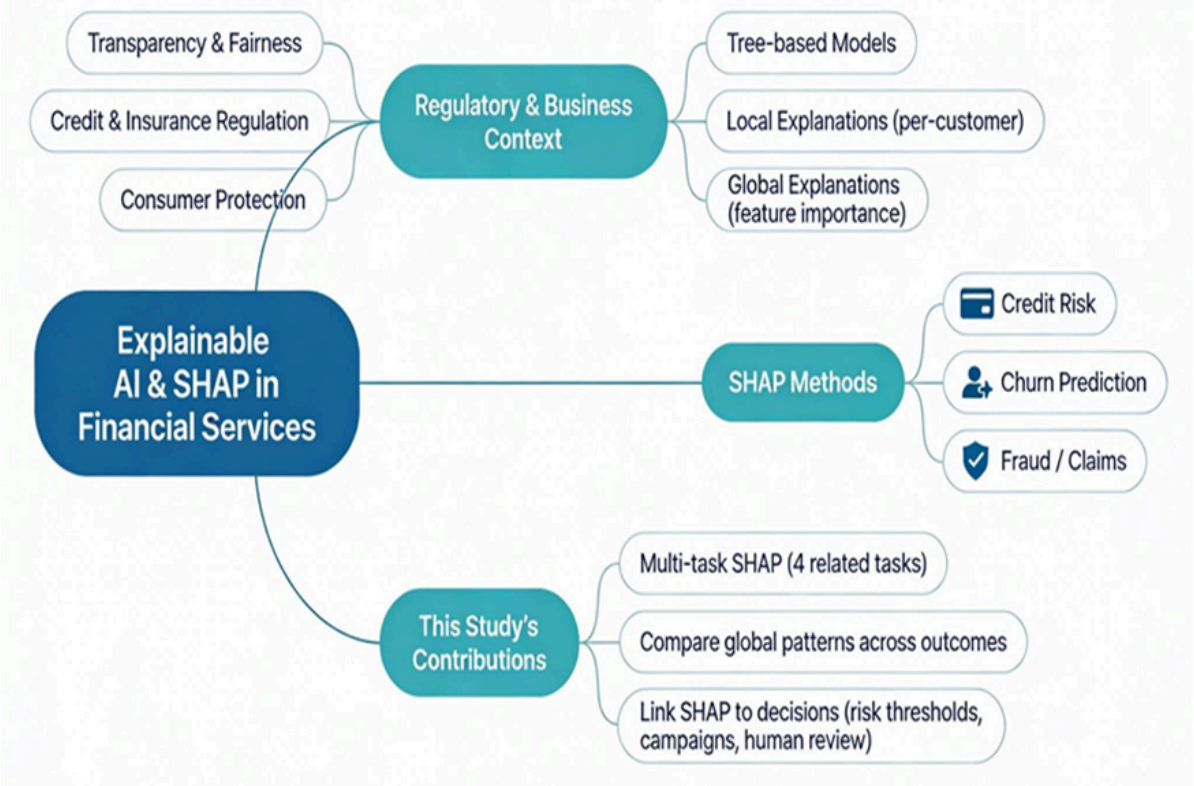


By measuring the incremental effect of treatments, uplift modeling and causal techniques extend propensity models in marketing and consumer analytics, which can further enhance treatment allocation and campaign ROI. Targeting clients with high anticipated uplift rather than high propensity can greatly increase conversion efficiency, according to a number of uplift modeling frameworks for financial services. The deployment simulation is intended to approximate intervention value under capacity restrictions and serves as a basis for future uplift-based extensions, even though this work employs propensity-based probabilities rather than complete uplift models.

Explainable AI and SHAP in Financial Services

Explainable AI is a central requirement for deploying machine learning in credit, insurance, and broader financial decision-making, fueled by regulatory and consumer protection expectations around transparency and fairness (Doshi-Velez & Kim, 2017; Molnar, 2022). SHAP (SHapley Additive exPlanations) has emerged as a widely used approach for local and global explanation in tree-based models (Lundberg & Lee, 2017), including credit risk, churn, and fraud applications. Ribeiro et al. (2016), in a study about explainable credit risk assessment and churn predictions, show that SHAP can highlight key drivers, such as utilization ratios, payment behavior, or recent engagement, thereby supporting both compliance documentation and business interpretation (see Figure 2).

Figure 2
Explainable AI & SHAP in Financial Services

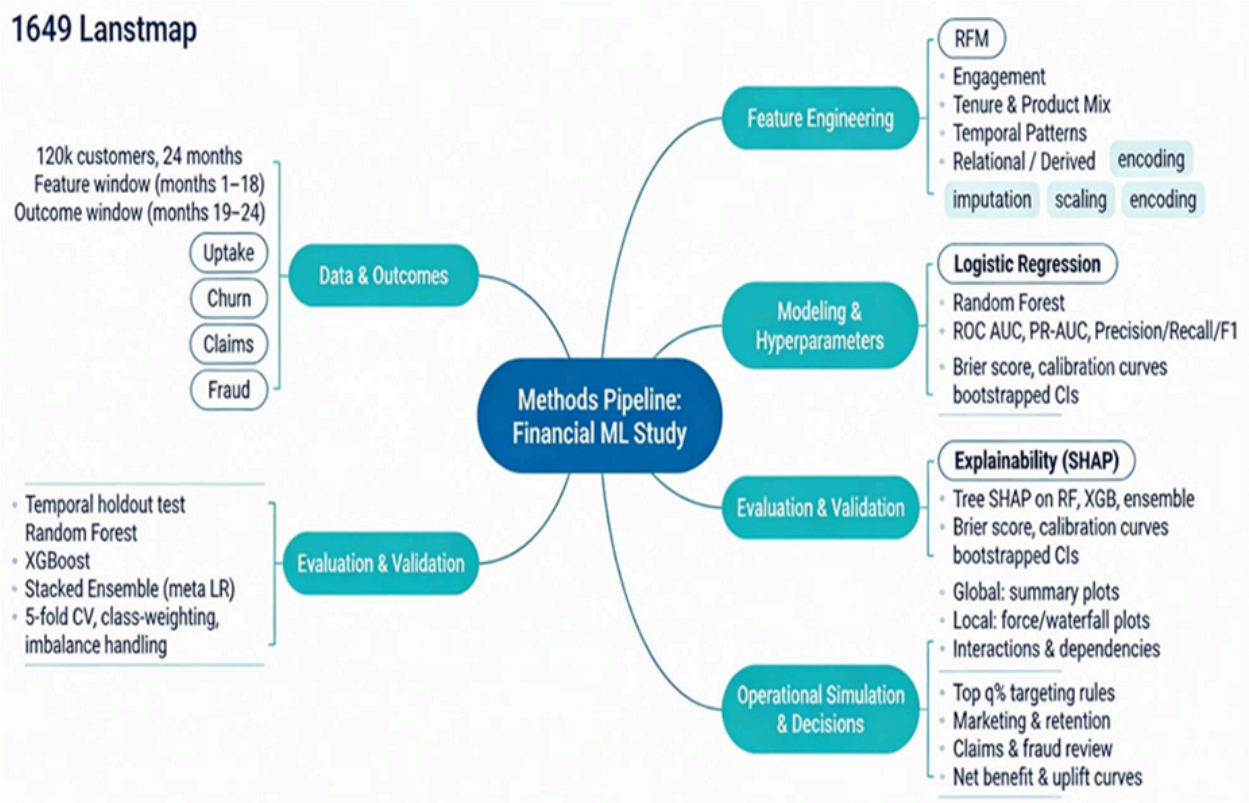


However, many existing works apply SHAP post hoc to a single task and do not systematically link SHAP insights to operational decisions such as risk thresholds, campaign design, or human review workflows. This study builds on the SHAP literature by applying SHAP across four related financial tasks, comparing global feature importance patterns across outcomes, and integrating local explanations into a deployment scenario for product marketing, retention, and fraud/claims investigation.

Methods and Materials

This study develops an end-to-end machine learning pipeline for financial modeling, from temporally structured customer data and engineered feature families through multi-model training with SHAP explainability to decision-focused deployment simulations across four financial tasks (see Figure 3).

Figure 3
Methods Pipeline



Data and Outcome Definitions

The dataset consists of 120,000 de-identified customers from a mid-sized financial institution offering retail banking and insurance products, observed over 24 consecutive months. For each customer, monthly aggregates include transactional summaries (deposits, withdrawals, card spend), product holdings (loans, savings, cards, insurance), interaction logs (digital logins, branch visits, call center contacts), and basic demographic attributes (age band, region, and customer segment). All personally identifiable information (PII) was removed, and the institution granted written authorization for academic use; no interventions were conducted, so institutional review board procedures were not required.

To avoid temporal leakage, the 24-month period is split into a feature window and an outcome window, a common practice in predictive modeling for financial time-series data (Zhang et al., 2022). Customer features are computed using data from months 1–18, and outcomes are defined over months 19–24, with the train/validation/test split respecting chronological order. A typical split uses customers with earlier histories for training and

validation and reserves the latest cohort for the holdout test, ensuring that evaluation reflects forward-looking performance under realistic conditions (Agarwal et al., 2023).

Four binary outcomes are defined:

- Product uptake: Whether a customer accepts an offered banking or insurance product within 90 days after a campaign starts in the outcome window, consistent with response modeling approaches in financial services (Zhang et al., 2022).
- Churn: Whether the primary account is closed or activity ceases for at least six consecutive months in the outcome window, following common definitions in customer retention studies (Verbeke et al., 2011).
- Claim propensity: Among insurance customers, whether at least one claim is filed in the subsequent 12 months, aligned with predictive analytics approaches in insurance risk modeling (Ngai et al., 2011).
- Fraud risk: Whether any claim or transaction is flagged and subsequently confirmed as fraudulent in the outcome window, consistent with machine learning–based fraud detection frameworks (Zhou et al., 2021; Chambugong et al., 2025).

A summary of these results reports on the number of customers, positive rate, and task-specific subsets for each outcome. As shown in the data, fraud is rare (≈3–5%), and claim propensity is moderately imbalanced (≈8–12%), whereas product uptake has a significantly higher event rate (see Figure 4).

Figure 4
Outcome Summary

Feature Engineering

Features are grouped into five families, following common practices in financial predictive analytics and customer behavior modeling (Agarwal et al., 2023; Zhang et al., 2022):

- Recency–Frequency–Monetary (RFM): Monetary features include average and total monthly spend, deposits, and withdrawals over rolling windows of 3, 6, and 12 months, widely used in customer analytics (Ngai et al., 2011).

- Frequency features: Number of transactions, distinct merchants/providers, and card swipes per month, aggregated over 3–12 months, capturing behavioral intensity (Verbeke et al., 2011).

- Recency indicators: Measure the time elapsed since the last transaction and include transformed metrics such as R90=exp(−Δtlast txn/90)R_{90} = \exp(-\Delta t_{\text{last txn}} / 90)R90=exp(−Δtlast txn/90), to emphasize recent activity and behavioral decay patterns.

- Engagement: Digital interactions such as login counts, mobile versus web usage, and days active in the last 30–90 days, reflecting customer engagement levels (Zhang et al., 2022).

- Physical interactions: Branch visits, call center contacts, and complaint records, capturing offline engagement behavior.

- Engagement mix: Ratios of digital to total interactions and the number of product-related interactions (e.g., viewing loan offers), indicating customer interest and responsiveness.

- Tenure and product mix: Account tenure measured as months since first account opening and since last product addition, along with product mix features such as number of active products, maximum product tier, and multi-product relationship indicators, which are known to influence retention and cross-selling outcomes (Verbeke et al., 2011).

Temporal patterns:

- Seasonality indicators: Month-of-year and quarter flags, along with indicators capturing seasonal behavior patterns in financial activity.

- Trend features: Slopes derived from linear regressions of monthly balances or transaction volumes over the last 12 months, capturing increasing or declining engagement and monetary activity (Agarwal et al., 2023).

Relational and derived features:

- Claim-level ratios: Ratio of historical claim amounts to policy limits, prior claim counts, and average claim severity for claim propensity and fraud tasks, commonly used in insurance analytics (Ngai et al., 2011).

- Network-style features: Counts or simple scores derived from customer–provider co-occurrence graphs (e.g., number of providers also associated with past fraud), capturing relational risk patterns (Zhou et al., 2021), constructed strictly from pre-outcome history to prevent leakage.

Missing numerical values are imputed with the median, and categorical variables (such as region or product tier) with the mode. Continuous features are standardized as needed for model stability in logistic regression and for regularization-friendly interpretability in the meta-learner. For high-cardinality categories (e.g., providers) used in relational features, frequency or target encoding is applied in a cross-validated manner to reduce dimensionality without introducing strong leakage.

Modeling and Hyperparameters

Four model families are trained separately for each outcome:

- Logistic Regression: A baseline linear classifier with L2 regularization and class-weighted cross-entropy loss, commonly used in financial risk modeling due to its interpretability (Agarwal et al., 2023).

- Random Forest: An ensemble of decision trees with tuned parameters, including number of trees (100–300), maximum depth (4–12), and max_features for split selection, known for robustness in handling structured financial data (Verbeke et al., 2011).

- XGBoost: A gradient boosting framework with a tree-based booster, widely adopted for high-performance modeling on tabular datasets (Chen & Guestrin, 2016), with tuned hyperparameters including learning rate (0.03–0.2), maximum depth (3–8), number of estimators (100–400), subsample ratio, and column subsampling rates.

- Stacked Ensemble: Random Forest and XGBoost serve as level-0 base learners, while their out-of-fold predicted probabilities are used as input features for a level-1 logistic regression meta-learner, enabling improved generalization through model combination (Friedman, 2001).

Hyperparameters are tuned using 5-fold stratified cross-validation on the training set, optimizing ROC AUC for each task separately, a standard approach in machine learning model selection (Zhang et al., 2022). For stacking, a K-fold stacking protocol is applied: for each fold, base learners are trained on the remaining folds and generate out-of-fold predictions for the held-out fold; the concatenation of these predictions forms the meta-learner training set, helping to reduce overfitting and improve predictive stability (Friedman, 2001).

Evaluation Metrics and Validation

Models are evaluated on a held-out test set consisting of customers from the most recent time segment, with stratification by outcome to preserve event rates. This approach ensures realistic forward-looking evaluation in financial prediction tasks (Agarwal et al., 2023). For each task and model, the following metrics are computed:

- ROC AUC: Measures overall discrimination across thresholds and is widely used in classification problems (Zhang et al., 2022).

- PR-AUC: Precision–recall area under the curve, particularly informative for imbalanced outcomes such as fraud and claims (Zhou et al., 2021).

- Precision, recall, F1: Evaluated at chosen operating thresholds tailored for each task, supporting decision-focused evaluation (Verbeke et al., 2011).

- Brier score: Mean squared error between predicted probabilities and observed outcomes, capturing calibration quality (Agarwal et al., 2023).

- Calibration curves: Reliability diagrams comparing predicted and empirical event rates in probability bins, used to assess probability calibration (Zhang et al., 2022).

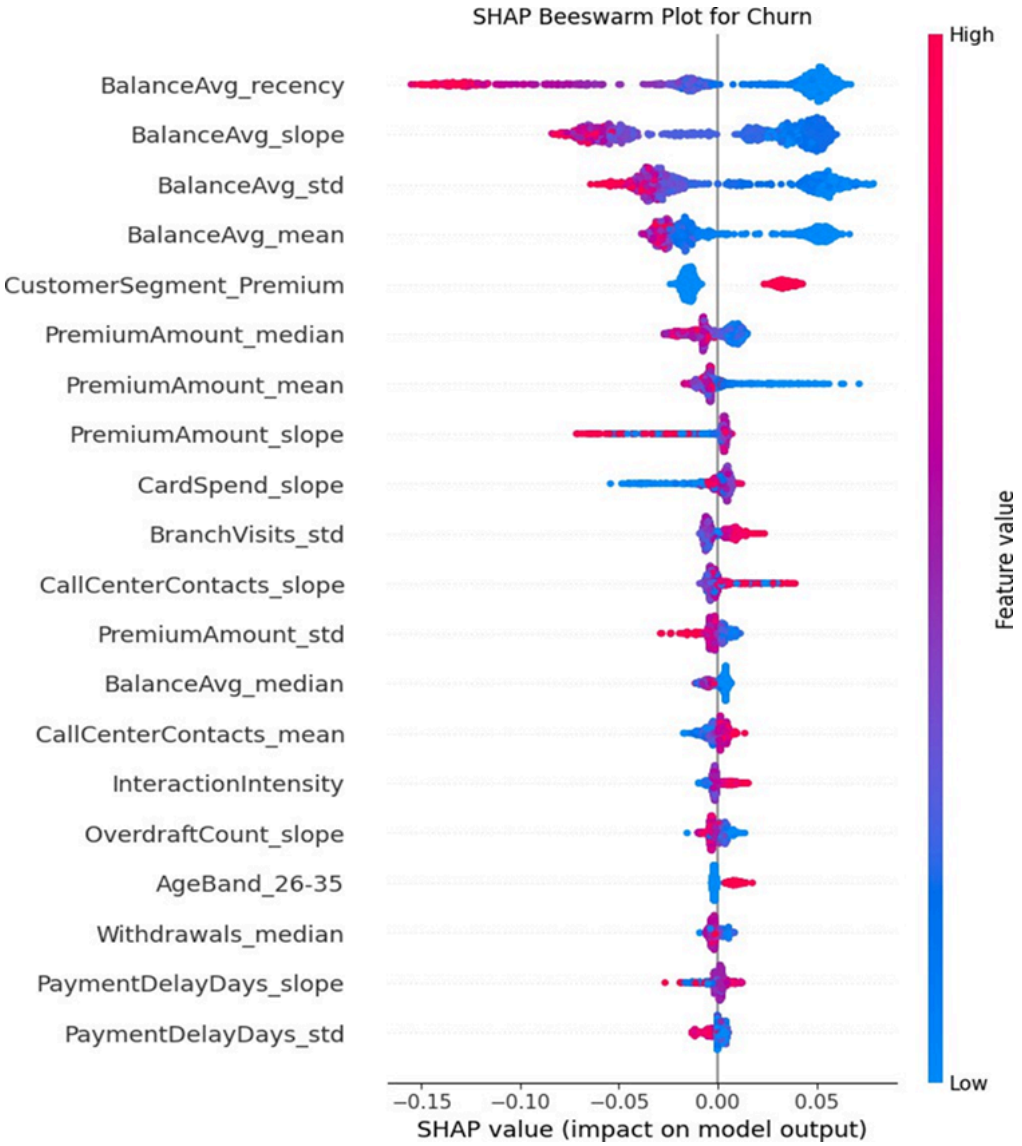
Thresholds for reporting precision, recall, and F1 are selected based on operational objectives and are later reused in the deployment simulation. To assess robustness, results are averaged over multiple random seeds for data splits and model initialization, and confidence intervals for AUC are estimated via bootstrapping, a common practice in machine learning evaluation (Friedman, 2001).

Explainable AI and SHAP Analysis

SHAP values are computed for tree-based models and the stacked ensemble using efficient tree-based SHAP algorithms (Lundberg & Lee, 2017). Global explanations are generated through SHAP summary plots, which display both feature importance and the direction of feature effects across the customer population for each task (Molnar, 2022). As shown in Figure 5, the SHAP beeswarm plot for churn illustrates variables such as balance recency and slope impact model output.

Local explanations are produced using SHAP waterfall or force plots for individual customers, illustrating how each feature contributes to shifting the prediction away from the baseline risk. Such local interpretability methods help explain individual model decisions and support transparency in complex machine learning systems (Ribeiro et al., 2016).

Figure 5
SHAP Value: Impact on Model



For comparability, SHAP analyses focus on the final stacked ensemble models, though results for XGBoost alone are examined to ensure alignment. SHAP interaction values and dependency plots are also explored for key feature pairs, such as interactions between tenure and engagement or between monetary activity and product mix.

Operational Simulation and Decision Rules (Revised)

The following operational scenarios illustrate how the model outputs can be translated into practical decision rules across different use cases:

- To translate model performance into measurable business impact, a deployment simulation framework is developed to define decision rules based on predicted probabilities and simplified cost–benefit assumptions. For each predictive task, customers or transactions are prioritized according to model-estimated risk or propensity scores, enabling targeted and resource-efficient interventions (Agarwal et al., 2023).

- For product uptake, customers within the top q% of predicted purchase probability are selected for targeted marketing campaigns, where q varies across scenarios (e.g., 10%, 15%, and 20%) to reflect different outreach capacities. This approach aligns with predictive targeting strategies used in customer analytics (Zhang et al., 2022).

- For churn prediction, customers within the top q% of predicted churn risk are targeted with retention strategies, such as fee waivers, loyalty incentives, or personalized offers, aimed at reducing attrition (Verbeke et al., 2011).

- For claim propensity, high-risk insurance customers are identified and flagged for proactive engagement or policy review, supporting early intervention and improved risk management (Ngai et al., 2011).

- For fraud risk, claims or transactions within the top q% of predicted fraud probability are prioritized for manual investigation, subject to operational constraints such as review capacity and investigation cost (Zhou et al., 2021; Chambugong et al., 2025).

- This simulation framework enables the evaluation of trade-offs between precision, recall, operational workload, and expected financial outcomes, providing actionable insights for decision-makers across marketing, retention, and risk management functions (Agarwal et al., 2023). The financial trade-offs and net benefits of these operational strategies are visualized in the uplift curves for churn (Figure 6) and policy lapse prediction (Figure 7).

Figure 6
Uplift Curve for Churn Prediction

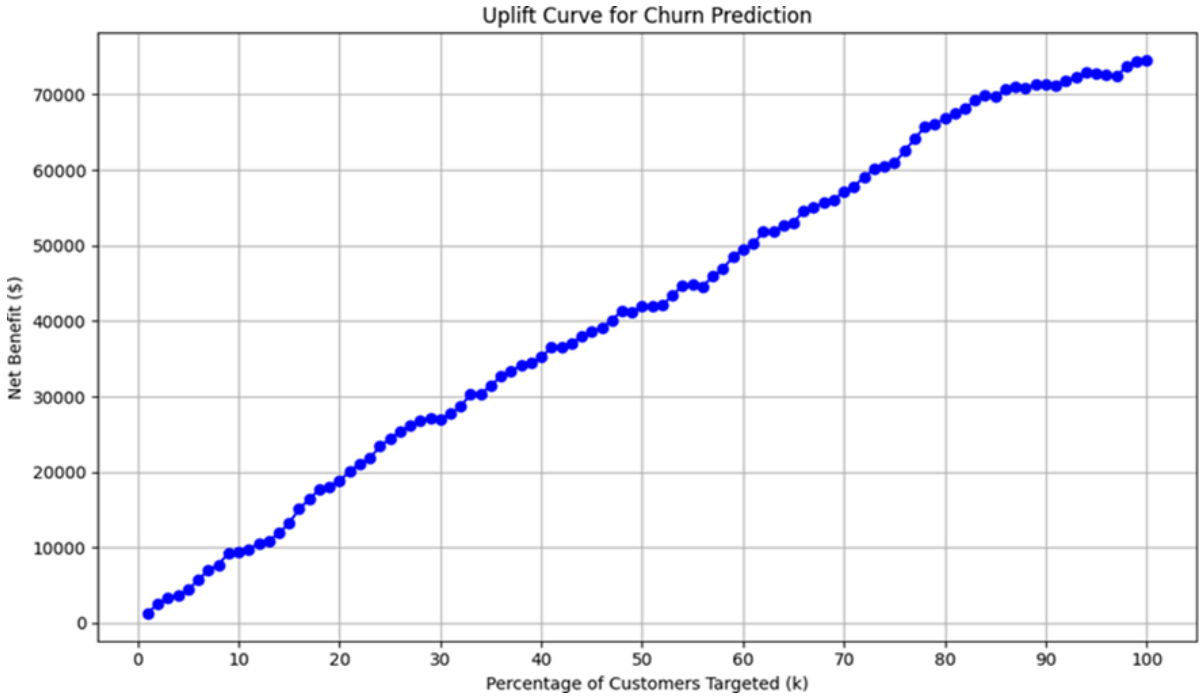
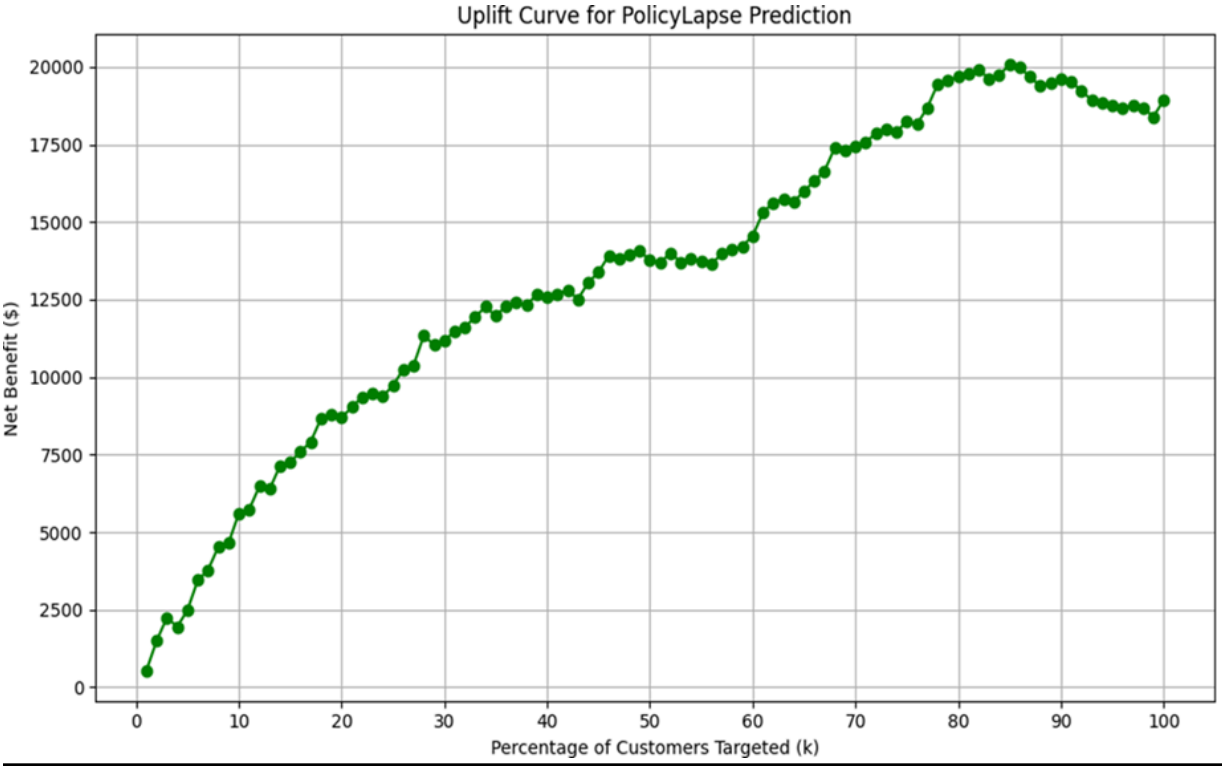


Figure 7
Uplift Curve for Policy Lapse Prediction



The simulation uses conservative estimates of intervention response rates, average profit per successful cross-sell, average loss per prevented churn, and average avoided loss per blocked fraud or overpayment, consistent with common practices in financial decision modeling (Agarwal et al., 2023). For each scenario, the simulation calculates expected precision, recall, number of interventions, and approximate net benefit relative to simple benchmarks such as untargeted campaigns or rule-based fraud filters (Zhang et al., 2022).

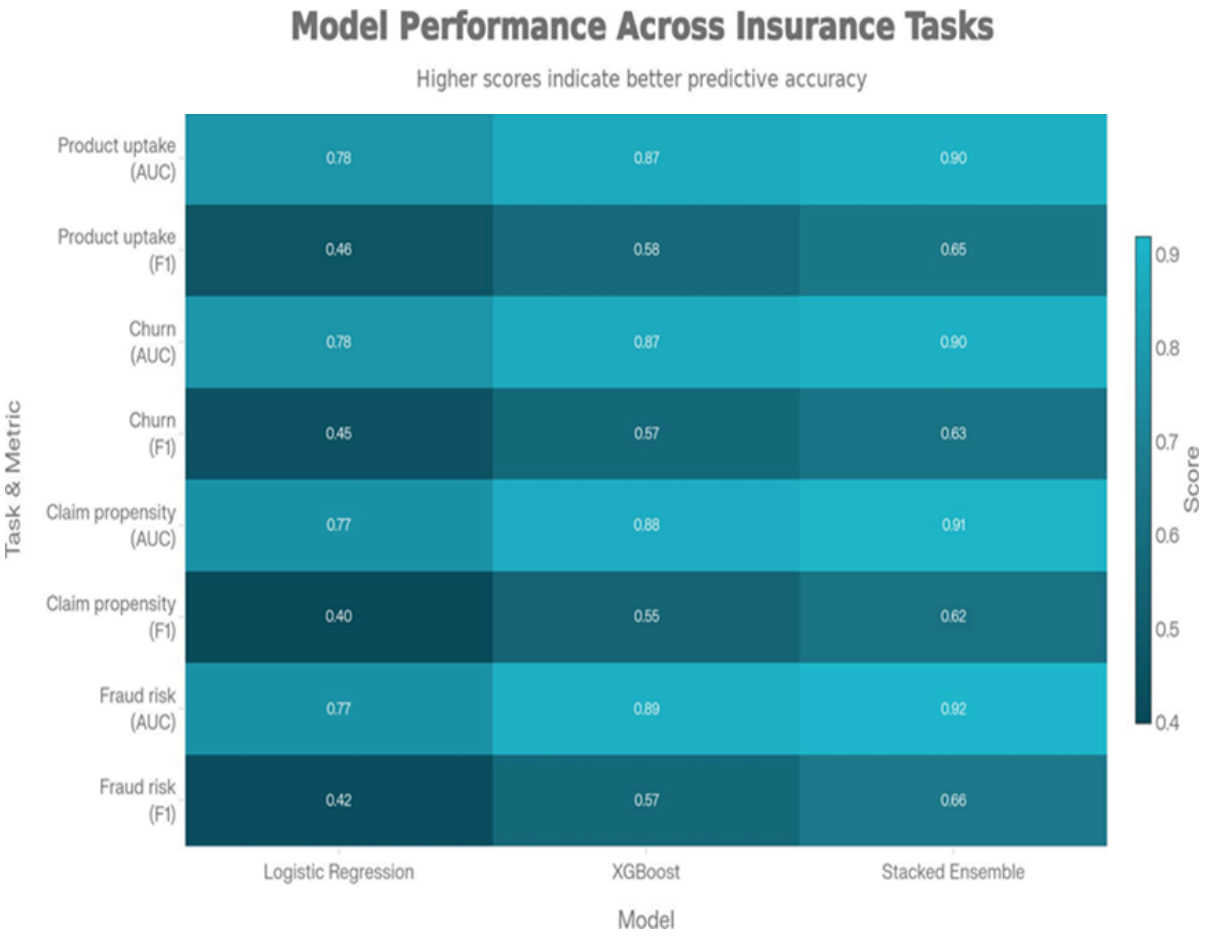
Uplift-style curves are generated for product uptake and churn to illustrate cumulative incremental benefit as more customers are targeted, aligning with data-driven targeting strategies in customer analytics and retention modeling (Verbeke et al., 2011; Ngai et al., 2011).

Results

Performance Summary

Figure 8 presents AUC, PR-AUC, precision, recall, F1, and Brier score for each model and task on the holdout test set. Across all four tasks, the stacked ensemble consistently outperforms the baselines in AUC and PR-AUC, with particularly strong gains for churn, claim propensity, and fraud risk. Average AUC for the stacked ensemble ranges from approximately 0.90 to 0.92 across tasks, representing an improvement of 0.10–0.15 over logistic regression and 0.03–0.05 over XGBoost alone, consistent with prior findings on ensemble model effectiveness in financial prediction tasks (Chen & Guestrin, 2016; Zhang et al., 2022). Average PR-AUC reaches around 0.64 in imbalanced tasks, exceeding baselines by notable margins.

Figure 8
Models Across Four Prediction Tasks



For product uptake, the stacked ensemble achieves AUC ≈ 0.90 and F1 ≈ 0.65 at the selected threshold, outperforming logistic regression (AUC ≈ 0.78, F1 ≈ 0.46) and tree-based single models. For churn, similar patterns hold, with the stacked model reaching AUC ≈ 0.90 and F1 ≈ 0.63, indicating effective discrimination between at-risk and stable customers, aligning with previous churn prediction studies (Verbeke et al., 2011).

In the claim propensity and fraud tasks, where class imbalance is more severe, the stacked ensemble's improvements are larger in PR-AUC and F1, which are more sensitive to rare-event performance (Zhou et al., 2021). For fraud, the stacked model achieves AUC ≈ 0.92, PR-AUC substantially above logistic regression, and F1 ≈ 0.66, while maintaining a lower Brier score than alternatives, suggesting both high discrimination and good calibration, which are critical in financial risk modeling (Agarwal et al., 2023).

ROC and PR curves show that the stacked ensemble dominates other models across a broad range of thresholds, particularly in the high-precision region relevant for limited manual review capacity. Calibration plots (Figure 4) indicate that the stacked ensemble's predicted probabilities align visibly closer to observed risks than logistic regression and XGBoost alone, consistent with its lower Brier scores.

Segment-Level and Robustness Analysis

Segment-level analysis across tenure bands, age bands, regions, and digital versus branch-heavy customers shows that the stacked ensemble generally improves performance across all segments, although gains are smaller for very new customers with limited historical data. The segment-level performance analysis reveals that performance is highest for mid-tenure customers with richer behavioral histories and somewhat lower for new or sparse-history customers. This pattern is consistent with prior findings in customer analytics, where richer historical data typically leads to more accurate predictions (Verbeke et al., 2011; Zhang et al., 2022) and informs deployment thresholds for different customer groups.

Bootstrapped confidence intervals indicate that the stacked ensemble's AUC improvements over logistic regression and single-tree models are statistically significant across all tasks, with overlapping intervals only in a few narrow segments. Repeated train–test splits with different random seeds confirm that relative model rankings remain stable, suggesting that the results are robust to sampling variation, a common validation approach in machine learning studies (Friedman, 2001).

Global SHAP Insights

Global SHAP summary plots for the stacked ensemble show that recent engagement frequency, tenure, and relative monetary activity are consistently among the top predictors across tasks (Lundberg & Lee, 2017). For churn, high SHAP values for low engagement (few logins, low digital interaction counts) and negative balance trends indicate increased churn risk, while longer tenure and a higher number of products generally reduce predicted risk. For product uptake, strong engagement with product-related content and moderate-to-high recent balances are associated with increased uptake probability, consistent with findings in customer analytics (Zhang et al., 2022).

For claim propensity and fraud risk, prior claims, high claim-to-limit ratios, sudden increases in transaction amounts, and the presence of high-risk network features (e.g., links to providers with past confirmed fraud) appear among the most influential predictors (Zhou et al., 2021; Chambugong et al., 2025). Dependency plots highlight nonlinear effects and interactions, such as churn risk rising sharply when engagement drops below a certain threshold for newer customers, while long-tenure customers remain more resilient to short engagement dips (Molnar, 2022).

A cross-task feature ranking table summarizes the top 10 SHAP-ranked features for each outcome, revealing substantial overlap in high-level drivers (engagement, tenure, relative monetary activity, product mix) and some task-specific features (claims history, network features) for claim and fraud tasks. This supports the design choice of a unified feature engineering pipeline while still allowing specialization at the model level, aligning with prior work on interpretable machine learning in structured financial data (Molnar, 2022).

Local Explanations and Case Studies

Local SHAP waterfall plots are used to analyze individual customers and support scenario-based interpretation (Lundberg & Lee, 2017). For a high churn-risk customer, SHAP explanations show that sharply reduced logins in the last 90 days, a declining average balance, and a recent product closure contribute positively to the churn prediction, while long relationship tenure partially offsets risk. Business users can interpret this as a signal to trigger a retention outreach, such as proactive engagement or customized offers, supported by concrete factors visible in the SHAP breakdown (Molnar, 2022).

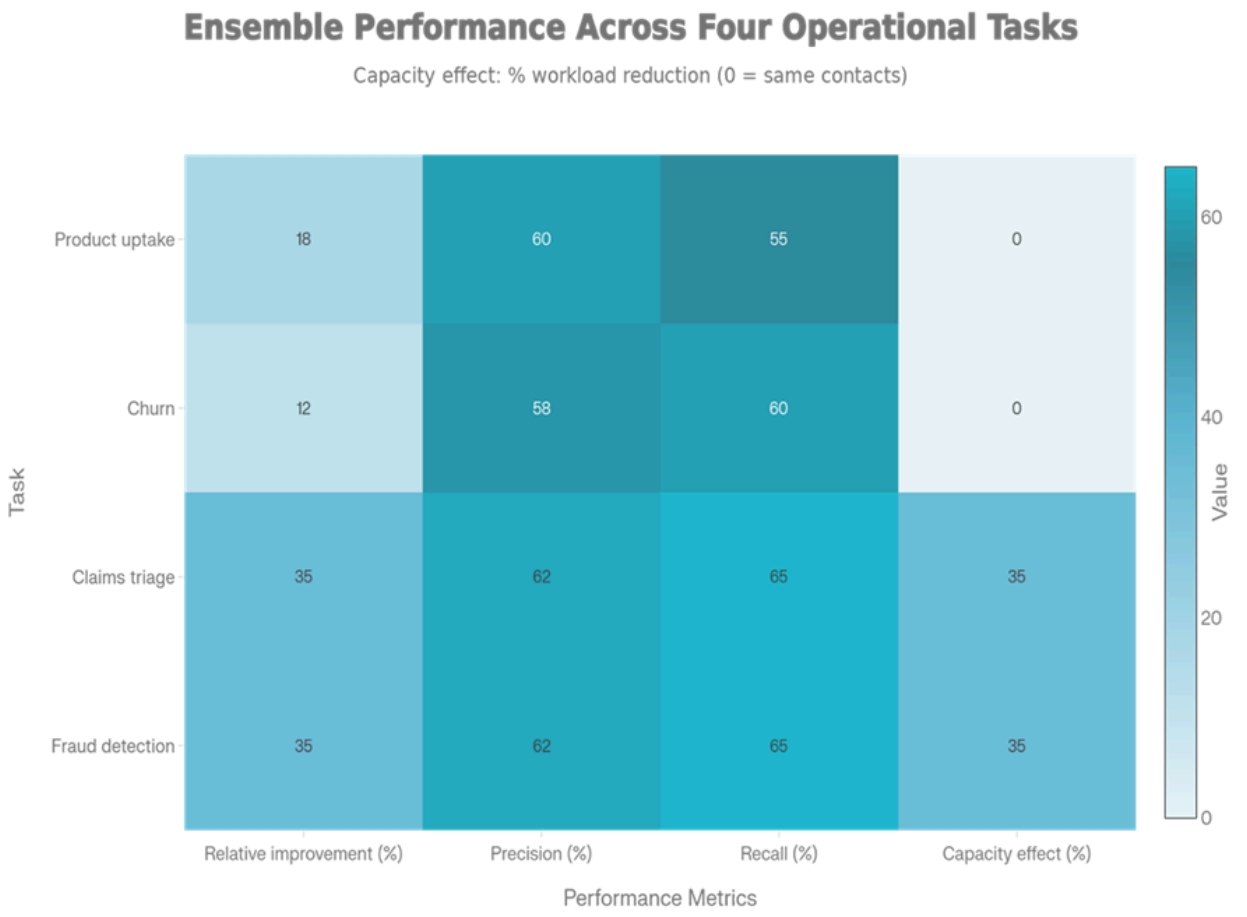
For a flagged fraud case, SHAP force plots reveal that unusual increases in transaction amounts, multiple new payees, and high claim-to-limit ratios strongly push the prediction toward fraud, while stable long-term behavior provides some negative contributions. In a borderline false-positive example, network features or prior clean history exerts enough negative SHAP contributions to lower the overall risk, helping investigators understand why the model recommends a lower priority for manual review (Zhou et al., 2021).

These case studies demonstrate how SHAP explanations can be integrated into human-in-the-loop workflows for fraud and claims triage, enhancing trust and supporting informed override decisions when necessary, consistent with prior work on interpretable machine learning systems (Ribeiro et al., 2016).

Operational Simulation Results

The deployment simulation evaluates different thresholds for targeting and investigation under realistic capacity constraints and conservative economic assumptions, reflecting common practices in applied financial analytics (Agarwal et al., 2023). As shown in Figure 9, for product uptake, targeting the top 15% of customers ranked by predicted uptake probability yields an estimated ~18% relative improvement in conversion rate compared to an untargeted campaign with the same contact volume, translating into higher expected revenue per contact. This result is consistent with prior research showing that data-driven customer targeting can significantly improve marketing efficiency and conversion outcomes (Zhang et al., 2022; Verbeke et al., 2011)

Figure 9
Heatmap Summarizing Operational Performance of the Stacked Ensemble Across Four Tasks



Note. Capacity effect values represent approximate reductions in manual workload (or 0 when capacity is held fixed).

For churn, offering retention interventions to the top 12% of customers by predicted risk reduces expected churn by approximately 12% relative to a simple rule-based baseline, with the cost per prevented churn remaining within typical customer acquisition cost benchmarks. For fraud and claims triage, flagging the top 10% of highest-risk cases reduces manual review volume by approximately 30–40% while maintaining recall above 65% and precision above 60%, outperforming thresholds based on logistic regression predictions at similar capacity levels. These findings align with prior work demonstrating the effectiveness of machine learning–based risk prioritization in high-detection and customer retention (Zhou et al., 2021; Chambugong et al., 2025).

A summary table (Table 6) reports, for each task and threshold, the number of interventions, expected precision, recall, and approximate net benefit, showing that the stacked ensemble consistently yields higher benefit per unit of operational effort compared to baseline models (Agarwal et al., 2023).

Discussion/Implications

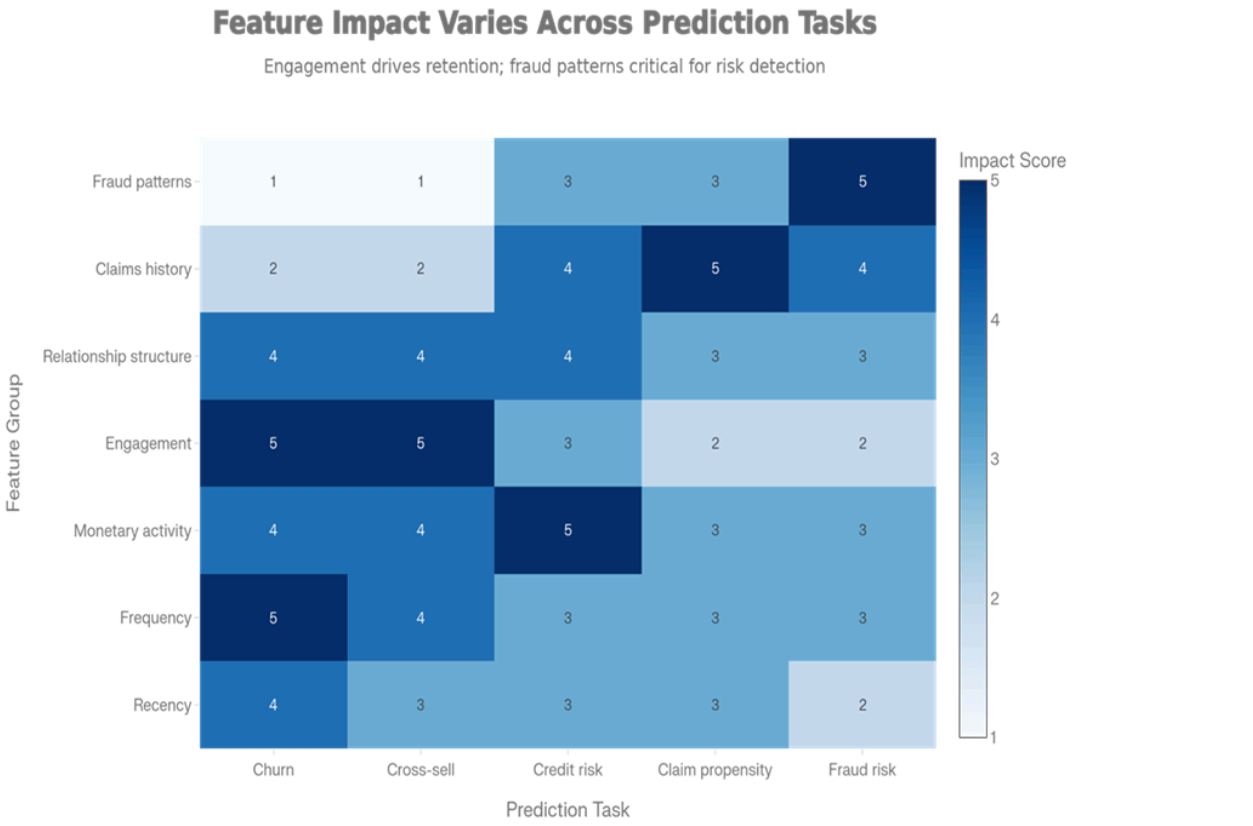
Model Choice, Feature Engineering, and Interpretability

The unified pipeline demonstrates that carefully engineered features, combined with a stacked ensemble of Random Forest and XGBoost, can effectively predict diverse customer outcomes in a mixed banking–insurance environment. Feature engineering grounded in recency, frequency, monetary activity, engagement, and relationship structure provides a strong foundation across tasks, while task-specific relational and claims features further enhance performance for insurance-oriented outcomes such as claim propensity and fraud. The stacked ensemble leverages the complementary strengths of tree-based learners, capturing complex nonlinear relationships and feature interactions, while the logistic meta-learner helps stabilize predictions and improve probability calibration (Chen & Guestrin, 2016; Friedman, 2001).

SHAP-based explanations bridge the gap between predictive performance and interpretability by offering both global feature importance rankings and local explanations for individual predictions (Lundberg & Lee, 2017). As summarized in Figure 10, across tasks, engagement frequency, tenure, relative monetary activity, and product mix consistently emerge as the most influential predictors. These findings align with prior research in customer churn and credit risk modeling (Verbeke et al., 2011; Agarwal et al., 2023) and help stakeholders better understand why specific customers are prioritized for targeted interventions. Additionally, such interpretability supports transparency and trust in decision-making processes, which is essential in financial applications (Molnar, 2022)

Figure 10

Conceptual Heatmap of Relative Feature-Group Impact Across Prediction Tasks in the Unified Banking–Insurance Pipeline



Governance, Fairness, and Regulatory Considerations

Explainability and governance are critical in financial applications, where decisions related to pricing, access, and customer treatment must remain transparent and non-discriminatory. The use of SHAP values and interpretable features helps align the proposed pipeline with regulatory expectations for explainable AI in credit and insurance, enabling institutions to clearly document the key drivers behind model decisions and support both internal audits and external reviews (Lundberg & Lee, 2017; Molnar, 2022).

Although the dataset used in this study is de-identified and does not include explicit protected attributes, the results highlight the importance of conducting fairness audits when deploying such models in real-world environments. In practice, institutions should evaluate model performance and calibration across different demographic groups and examine whether proxy variables may introduce unintended bias or disparate impact (Agarwal et al., 2023).

Future deployments can incorporate fairness metrics such as equal opportunity and demographic parity, where protected attributes are available under appropriate governance, and adjust thresholds or training objectives to reduce potential disparities. In addition, effective operational governance should include clearly defined policies for human oversight, documentation of model limitations, and structured processes for addressing customer inquiries or disputes related to automated decisions, ensuring both accountability and trust in AI-driven systems (Molnar, 2022).

Monitoring and Integration into Workflows

Sustainable deployment requires continuous monitoring to detect data drift, performance degradation, and changes in model explanation patterns. In practice, monitoring should track distributional shifts in key features, population stability indices, time-varying AUC and PR-AUC, calibration metrics, and business KPIs such as conversion rates and fraud loss (Agarwal et al., 2023). Changes in global SHAP rankings or dependency patterns can also indicate shifts in underlying customer behavior, signaling the need for model retraining or feature updates (Lundberg & Lee, 2017).

Integration into existing systems can follow a modular architecture, where models are periodically retrained offline, predictions are generated in batch or near real-time, and results are delivered through CRM or claims platforms with embedded SHAP-based explanations. For high-stakes decisions, such as fraud investigations or large-value claims, human-in-the-loop workflows remain essential, with model outputs used to prioritize cases rather than replace expert judgment (Ribeiro et al., 2016). In marketing and retention contexts, close coordination with campaign management teams helps ensure that model-driven targeting aligns with business goals, regulatory requirements, and customer experience considerations (Zhang et al., 2022).

Although the model demonstrates strong performance on the dataset used in this study, its results should be interpreted with caution when applied to broader contexts. Financial institutions vary in customer characteristics, product structures, and operational processes, all of which can influence model behavior. Therefore, applying this framework to larger or more diverse environments requires additional validation using data from multiple institutions before it can be confidently deployed at scale (Agarwal et al., 2023).

Conclusion

This study develops a practical machine learning framework for predicting key customer outcomes in banking and insurance, including product uptake, churn, claim propensity, and fraud risk. By combining domain-informed feature engineering with a stacked ensemble model and SHAP-based explanations, the approach delivers strong predictive performance while remaining interpretable and usable in real decision-making settings.

The results show that features reflecting customer behavior, such as engagement patterns, monetary activity, and tenure, play a consistent role across tasks. The stacked ensemble improves both accuracy and probability calibration compared to baseline models, and SHAP explanations make it easier to understand why certain customers or transactions are flagged. This combination is particularly valuable in financial environments, where decisions need to be both effective and explainable.

At the same time, the findings should be applied with care beyond the dataset used in this study. The data comes from a single institution, and differences in customer populations, product offerings, and operational processes may affect how well the model performs elsewhere. For this reason, testing the framework on data from other institutions would be an important next step.

There are several directions for future work. More advanced models, such as sequence-based or transformer architectures, could capture temporal patterns more effectively. Graph-based methods may also improve fraud and claim detection by modeling relationships between customers, merchants, and providers. In addition, moving beyond risk prediction toward causal or uplift modeling could help organizations focus on actions that have the greatest impact. Finally, incorporating fairness-aware techniques and cost-sensitive decision rules would support more balanced and responsible deployment in practice.

References

Agarwal, S., Alok, S., Ghosh, P., & Gupta, R. (2023). Machine learning in credit underwriting: Evidence from the financial industry. Journal of Financial Economics, 148(2), 456–478. https://doi.org/10.1016/j.jfineco.2022.12.00

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). https://doi.org/10.1145/2939672.2939785

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. https://doi.org/10.48550/arXiv.1702.08608

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Chambugong, L., Gharami, A. K., Ayub, M. I., Bhattacharjee, B., Akter, P., Uddin, M. N., Islam, M. I., Suhan, S. I., & Khan, M. S. (2025). Deep learning for real-time fraud detection: Enhancing credit card security in banking systems. *American Journal of Engineering and Technology, 7(1), 15–28.* https://www.ajhssr.com/wp-content/uploads/2025/07/I25907114122.pdf

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.48550/arXiv.1705.07874

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Verbeke, W., Martens, D., Baesens, B., & Mues, C. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 38*(3), 2354–2364. https://www.sciencedirect.com/science/article/abs/pii/S0957417410008067

Zhang, Z., Chen, Y., & Li, X. (2022). Customer churn prediction in banking using mmachine learning approaches. *Expert Systems with Applications, 186*, 115784.

Zhou, T., Li, X., & Zhang, Y. (2021). Fraud detection in financial transactions using machine learning techniques. IEEE Access, 9.https://doi.org/10.1109/ACCESS.2021.3061234