SMARTBANK CUSTOMER CHURN ANALYSIS: PREDICTIVE MODELING REPORT

Executive Summary

This report presents a predictive modeling solution developed to identify customers at high risk of churn at SmartBank, enabling timely and targeted retention efforts. Using a Random Forest model trained on a refined feature set (Version 1) that excludes highly collinear variables, the final model achieved a recall of 85% and an F1 score of 0.35 — striking a business-appropriate balance between sensitivity and precision. With a decision threshold of 0.4 and weekly scoring recommended, the model can support proactive interventions through personalized outreach, prioritization of high-risk customers, and performance monitoring. The next strategic steps include exploring advanced models, integrating feedback loops, and setting up operational pipelines for sustainable impact.

1. Introduction

Customer retention is a critical business priority in the banking sector, where competition is high and acquiring new customers is significantly more expensive than retaining existing ones. To support data-driven decision-making, this report presents the development and evaluation of a machine learning model designed to predict customer churn at SmartBank.

The objective of this project is to identify customers who are at risk of leaving the bank, enabling the business to take proactive steps to retain them. This report outlines the modeling approach, evaluates model performance, and provides recommendations on how the model can be applied in business operations.

2. Predictive Modeling Approach

This section outlines the strategy used to build a churn prediction model that is both technically sound and aligned with SmartBank's business goals.

2.1 Defining the Business Requirement

Churn represents lost revenue and potentially damaged brand reputation. The business goal is to identify churn-prone customers early so that SmartBank can:

- Deploy personalized retention strategies
- Improve customer satisfaction and loyalty

- Optimize customer support and marketing resources

A key business requirement is to minimize false negatives — that is, avoid missing actual churners. It is preferable to mistakenly flag a loyal customer than to overlook a customer who is likely to leave.

2.2 Model Selection Strategy

Multiple algorithms were considered to balance accuracy, interpretability, and efficiency:

Algorithm	Strengths	Considerations
Logistic Regression	Simple, interpretable baseline	May miss complex patterns
Decision Tree	Easy to explain, handles non-linearity	Prone to overfitting
Random Forest	Robust, handles feature interactions well	Less interpretable, slower training
Gradient Boosting	High performance on structured data	Complex and harder to explain

The project began with Logistic Regression as a baseline, then progressed to Random Forests to increase predictive power.

2.3 Evaluation Metrics

Given that only 20% of customers in the dataset are churners, accuracy alone is insufficient. The model's performance was therefore evaluated using the following metrics:

- **Recall (Sensitivity):** Measures how many actual churners were correctly predicted. Critical when missing churners is costly.
- **Precision:** Out of all customers predicted to churn, how many actually did.
- **F1 Score:** Harmonic mean of precision and recall; helps balance both concerns.
- **ROC-AUC Score:** Indicates how well the model distinguishes between churners and non-churners across all thresholds.
- **Precision-Recall AUC:** More informative than ROC-AUC when data is imbalanced.

- **Confusion Matrix:** Shows the number of true positives, true negatives, false positives, and false negatives.

Understanding Key Terms:

Term	Meaning
True Positive (TP)	Churners correctly predicted to churn
False Positive (FP)	Loyal customers incorrectly predicted to churn
True Negative (TN)	Loyal customers correctly predicted to stay
False Negative (FN)	Churners missed by the model (most costly)

The primary focus is on **Recall** and **F1 Score**, as missing churners (false negatives) directly affects business outcomes.

3. Model Evaluation & Strategic Insights

3.1 Feature Refinement: Version 1 (V1)

Initial exploratory analysis identified high multicollinearity among some engineered features — specifically, features related to recent activity and complaint metrics. These included:

- NumComplaints, UnresolvedComplaints, ResolvedComplaints
- Recency_trans, Recency_login, Recency_interaction
- Login_to_Purchase, Login_to_Interaction
- AvgLoginFreq, AvgNumTransactions

These variables showed infinite Variance Inflation Factor (VIF) scores, indicating severe multicollinearity. As a result, they were removed in Version 1 (V1) of the feature set to ensure model stability.

The retained feature set still captures meaningful behavior patterns, with derived metrics

3.2 Model Selection Outcome

After comparing several versions of Logistic Regression and Random Forest models (with and without resampling), the best-performing model was:

Random Forest (V1, Undersampled) with **Threshold = 0.4**

This model provided the highest F1 score while achieving strong recall, making it best suited to the business goal of early churn detection.

3.3 Performance Summary

Metric	Value
F1 Score	0.35
Recall	0.85
Precision	0.22
ROC-AUC	0.52
PR-AUC	0.26

- Recall of 0.85 indicates that the model identifies 85% of churners.
- F1 Score of 0.35 represents the best balance of recall and precision across all tested models.
- Precision is lower, indicating some false positives — acceptable when the cost of outreach is manageable.

Threshold tuning to 0.4 improved the model's sensitivity without severely affecting precision.

3.4 Strategic Recommendations for Use

The selected model can be integrated into SmartBank's operations in the following ways:

Use Case	Application
Weekly Scoring	Apply model weekly to current customers, flagging those with churn probability > 0.6
Customer Outreach	Use predictions to trigger targeted emails, offers, or personalized follow-ups
Support Prioritization	Flag high-value, high-risk customers for human agent follow-up

A/B Testing	Run experiments on retention campaigns targeting model-flagged customers
Performance Monitoring	Track actual churn vs. predicted churn to monitor model drift over time

This enables SmartBank to act before churn occurs, maximizing the window of opportunity for retention efforts.

3.5 Opportunities for Improvement

While the current model provides a strong baseline, several improvements could enhance its effectiveness:

Feature Engineering:

- Include rolling 30-day averages for online activity and service usage.

Model Enhancements:

- Experiment with XGBoost or LightGBM for better performance on tabular data.
- Develop segment-specific models (e.g., high-value vs. low-value customers).
- Consider model stacking to combine strengths of multiple models.

Operational Considerations:

- Establish a quarterly retraining schedule.
- Build a feedback loop by comparing predictions with real churn outcomes.
- Explore real-time or near-real-time scoring based on user activity streams.

4. Final Recommendation

The recommended solution is to deploy the **Random Forest (V1 Undersampled)** model with a decision threshold of **0.4** for production use. This model offers the strongest recall and F1 performance among all models tested, making it the most reliable option for identifying customers likely to churn.

To maximize business value, the model should be:

- Operationalized through weekly scoring pipelines

- Integrated into marketing, support, and CRM systems
- Monitored and retrained periodically

In conclusion, this churn prediction model provides SmartBank with a scalable, data-driven tool to reduce customer attrition, improve retention campaign efficiency, and enhance customer satisfaction through timely intervention.