Predicting Bank Customer Churn with Random Forest Modeling

Ashley Traore

Executive Summary and Implications

Statement of the Problem

Customer churn remains a critical challenge in the banking sector, where retaining existing clients is more cost-effective than acquiring new ones. The purpose of this project is to investigate whether a Random Forest classification model can effectively predict churn using structured demographic and account-level data, such as age, tenure, credit score, and account balance.

Hypothesis

Null Hypothesis (H_0): The Random Forest model does not achieve a predictive accuracy greater than 70% in identifying customer churn.

Alternate Hypothesis (H_1): The Random Forest model achieves a predictive accuracy greater than 70% in identifying customer churn.

This 70% threshold reflects a meaningful performance benchmark for operational deployment. Surpassing it would validate the model's utility in supporting proactive retention efforts and meeting stakeholder expectations for reliability and impact.

Data Analysis Process

The data analysis process began with the acquisition of a curated bank churn dataset from Kaggle, comprising over 10,000 customer records with structured variables such as age, tenure, credit score, account balance, and churn status. The dataset was imported into a pandas DataFrame and explored within Jupyter Notebooks to ensure flexibility and reproducibility. Initial inspection using head(), info(), and describe() provided a foundational understanding of data types, missing values, and distributional properties. The target variable (Exited) was visualized using countplot, revealing a churn rate of approximately 20%, which flagged a moderate class imbalance. Duplicate records were removed via drop_duplicates() to maintain statistical integrity, and missing values were standardized with np.nan and dropped using dropna() due to their minimal presence.

Outliers were detected using the IQR method and capped at a Z-score threshold of three standard deviations from the mean, preserving informative variance while minimizing distortion during model training. Categorical variables underwent cardinality checks using nunique() to guide the encoding strategy. Binary features were manually

mapped to 0/1, while multi-class variables like Geography were transformed using LabelEncoder, optimizing compatibility with tree-based models. Feature selection was conducted through a layered approach, beginning with correlation analysis to eliminate redundant predictors, followed by SelectKBest with f_regression to retain statistically significant features ($p < 0.05$). Multicollinearity was assessed using Variance Inflation Factor (VIF), confirming that all retained features fell below the threshold of 10, ensuring model stability and interpretability.

A stratified train-test split was applied to preserve class proportions and mitigate bias in performance metrics. The Random Forest classifier was trained with 200 estimators and a fixed random seed to ensure reproducibility. Model evaluation incorporated both standard and business-aligned metrics: accuracy, ROC-AUC, and confusion matrix diagnostics. A dual-threshold strategy was implemented, 0.5 for general churn detection and 0.75 for high-risk flagging, to balance broad awareness with precision targeting. High-risk customers were further prioritized by account balance, enabling strategic resource allocation for retention efforts. This end-to-end workflow emphasized statistical rigor, model interpretability, and stakeholder relevance, aligning technical outputs with operational decision-making.

Outline of the Findings

The Random Forest model demonstrated strong predictive capability in identifying customers at elevated risk of churn. After training on a stratified sample that preserved the original class distribution, the model achieved an accuracy of 85.2%, indicating that the majority of predictions aligned with actual outcomes in the test set. Additionally, the ROC-AUC score of 0.8511 reflected robust discriminatory power across varying probability thresholds. The confusion matrix revealed 1494 true negatives and 210 true positives, alongside 98 false positives and 198 false negatives. While the model effectively classified most customers, the presence of 198 false negatives, churned customers misclassified as retained, suggests an opportunity for further refinement, particularly in threshold tuning or feature engineering.

To support both broad detection and precision targeting, a dual-threshold classification strategy was implemented. A standard threshold of 0.5 was used for general churn detection, while a stricter threshold of 0.75 flagged high-risk customers for targeted retention efforts. This approach enabled flexible decision-making, balancing sensitivity with specificity, and aligning technical outputs with operational priorities. Customers flagged as high-risk were further prioritized by account balance, producing a ranked list of individuals whose departure would likely result in the greatest financial

impact. This risk-based targeting framework transformed raw model output into actionable intelligence, supporting ROI-driven outreach and strategic resource allocation.

The results provide compelling statistical evidence to reject the null hypothesis, which suggested that the model would not exceed 70% predictive accuracy. The alternative hypothesis, that the Random Forest model achieves greater than 70% accuracy, was supported, validating the model's utility for operational deployment. The model's performance metrics and risk-targeting capabilities offer meaningful insights for stakeholder decision-making.

Limitations of the Tools and Techniques

While the Random Forest modeling pipeline delivered strong predictive performance, several limitations were identified throughout the data preparation, modeling, and evaluation stages.

First, the use of a pre-curated dataset from Kaggle accelerated development but introduced constraints. Although the dataset was structured and labeled, it lacked behavioral or sentiment data that could enhance churn prediction. Additionally, the dataset exhibited class imbalance, with only 20% of customers labeled as churned. This imbalance can bias the model toward predicting the majority class, potentially reducing sensitivity to minority class patterns. Stratified sampling and ROC-AUC were used to mitigate this, but the imbalance remains a structural limitation.

During data cleaning, missing values were dropped due to their minimal presence. While this approach preserved data integrity, it also reduced sample size and may have discarded informative records. Similarly, outliers were capped using Z-score thresholds to limit distortion during training. Although this preserved the distributional structure, it may have obscured genuine behavioral signals, especially in features like CreditScore, Age, or NumOfProducts.

Categorical encoding introduced interpretability challenges. Label encoding was applied to multi-class variables such as Geography, which is efficient for tree-based models but does not preserve semantic meaning. This can make feature importance plots or stakeholder explanations less intuitive, as categories are represented by arbitrary integers.

Feature selection relied on statistical thresholds, correlation analysis, p-values via SelectKBest, and VIF scores to retain relevant predictors. While this layered

approach improved model stability and interpretability, it may have excluded features with contextual or business relevance that did not meet strict statistical criteria.

Random Forest itself, while robust and accurate, presents limitations in transparency. Individual decision paths within trees are difficult to interpret, which can challenge explainability in regulated environments or when communicating results to non-technical stakeholders.

Finally, the dual-threshold strategy (0.5 for general detection, 0.75 for high-risk flagging) prioritized precision over recall. While effective for targeting high-value customers, this stricter threshold may miss individuals likely to churn but fall below the cutoff, limiting the model's sensitivity.

Proposed Actions

- Targeted Retention Strategy
    - Focus outreach efforts on customers with high churn probability and substantial account value to maximize ROI. Prioritizing these individuals ensures retention resources are directed toward those with the greatest financial impact.
- Stakeholder Dashboards
    - Develop dashboards that visualize churn risk and account impact to support transparency and informed decision-making across business units.
- Segment-Specific Modeling
    - Create separate models for distinct customer segments (for example, geography, tenure, product usage) to improve targeting precision, interpretability, and fairness. This approach aligns retention strategies with segment-level business value and mitigates bias from dominant groups.
- Behavioral and Sentiment Enrichment
    - Enhance the dataset with interaction logs, support tickets, and sentiment scores to detect latent churn drivers like frustration or disengagement. These qualitative signals enable earlier, more accurate risk detection and support empathetic, loyalty-focused retention strategies.

Expected Benefits of the Study

The study demonstrates strong predictive performance, with the Random Forest classifier achieving an accuracy of 85.2% and a ROC-AUC of 0.8511, indicating reliable separation between churn and non-churn classes. The dual-threshold strategy (0.5 for general detection, 0.75 for high-risk flagging) enables tiered intervention planning. Among the 171 customers flagged as high risk, the mean account balance is $78,715.36. Retaining just 10%, approximately 17 customers, would preserve an estimated $1.34 million in account value. This quantifiable ROI directly supports the deployment of targeted retention strategies based on model outputs.

Although stakeholder-facing dashboards are not yet deployed, the model outputs are structured for seamless integration into real-time decision-support tools. Visualizing churn probability and account-level financial exposure would enhance transparency and enable business teams to act on insights with speed and precision. Future enhancements include segment-specific modeling to improve fairness and targeting accuracy, as well as the integration of behavioral and sentiment features (for example, support ticket frequency, survey-derived sentiment scores) to surface latent churn signals. These additions would strengthen early detection and support the development of context-aware retention strategies that balance financial incentives with customer experience.