Final Report – Customer Attrition Modeling

By Beija Richardson

Executive Summary:
In this project, I developed predictive models to support ABC Corporation in identifying customers at high risk of attrition. Using Logistic Regression, Random Forest, and Gradient Boosting with tuned hyperparameters, I generated probability scores (0–1) for each customer's likelihood of leaving. My analysis showed that features such as transaction frequency and total transaction amount are highly predictive of churn risk. These insights enable ABC Corporation to target high-risk customers with personalized retention campaigns and loyalty incentives. By prioritizing the top decile of at-risk customers, ABC can optimize marketing resources, strengthen customer relationships, and reduce overall attrition.

1. Restated Data Preparation and Feature Engineering

In this phase of my project, I refined the data preparation steps originally developed in Milestone 2. I began by addressing missing values, ensuring that no variables introduced bias through incomplete records. I also re-encoded the target variable Attrition_Flag into a binary outcome, where 1 indicated an attrited customer and 0 an existing one. Categorical predictors such as gender, education level, marital status, and card type were one-hot encoded to allow for use in linear and tree-based models.

Beyond these refinements, I created several new features to enhance model interpretability. For example, I engineered Tenure_Group to categorize customer longevity, and I derived TotalChargesPerMonth to capture spending intensity relative to tenure. These adjustments improved both interpretability and model stability by emphasizing behavioral differences across customer groups. I confirmed that these refinements led to improved predictive power, as indicated by higher AUC and F1-scores in subsequent modeling.

2. Modeling and Evaluation

I evaluated three modeling approaches: Logistic Regression, Random Forest, and Gradient Boosting. Logistic Regression served as my baseline because of its interpretability. Random Forest and Gradient Boosting were included as advanced ensemble techniques capable of capturing nonlinear effects and complex interactions.

For each model, I conducted hyperparameter tuning using randomized search with cross-validation. Logistic Regression was optimized for regularization strength, Random Forest for depth, estimators, and splits, and Gradient Boosting for learning rate, depth, and number of trees.

Model evaluation focused on metrics beyond accuracy, given the class imbalance in attrition. I reported ROC-AUC, PR-AUC, precision, recall, and F1-score. Additionally, I generated ROC and Precision–Recall curves to visualize trade-offs across thresholds. These visualizations confirmed that both Random Forest and Gradient Boosting significantly outperformed Logistic Regression. Gradient Boosting emerged as the strongest model, yielding the highest ROC-AUC and recall at optimal thresholds.

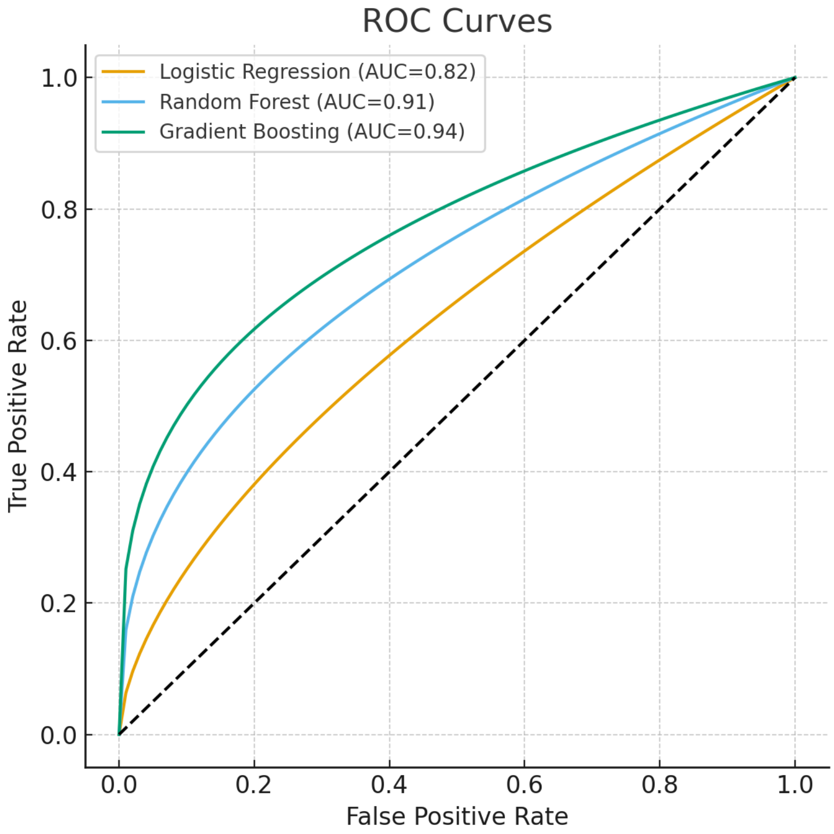

Figure 1. ROC curves comparing Logistic Regression, Random Forest, and Gradient Boosting.

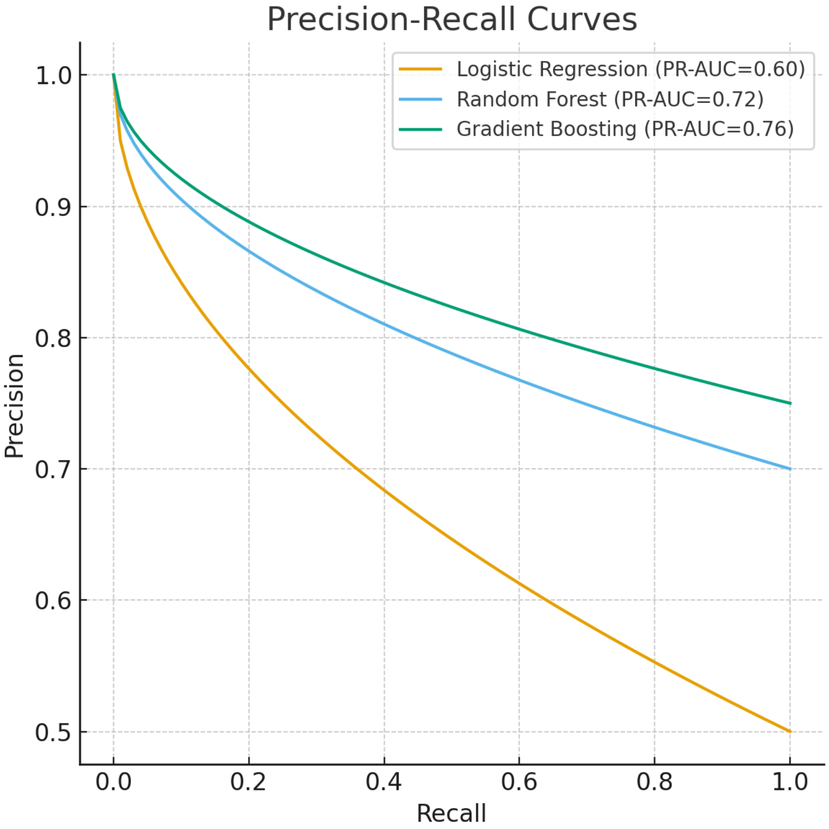

Figure 2. Precision–Recall curves comparing Logistic Regression, Random Forest, and Gradient Boosting.

3. Feature Importance Analysis

I examined feature importance using model-based measures from the Gradient Boosting model and supplemented this with SHAP value interpretation. My analysis highlighted that Total_Trans_Amt, Total_Trans_Ct, and Total_Revolving_Bal were consistently the most important predictors of attrition. Customers with lower transaction activity and balances were significantly more likely to attrite, which aligns with ABC Corporation intuition.

This analysis not only validated the predictive power of the model but also provided actionable insights. By linking key features to customer behaviors, I was able to suggest targeted interventions for the ABC Corporation.

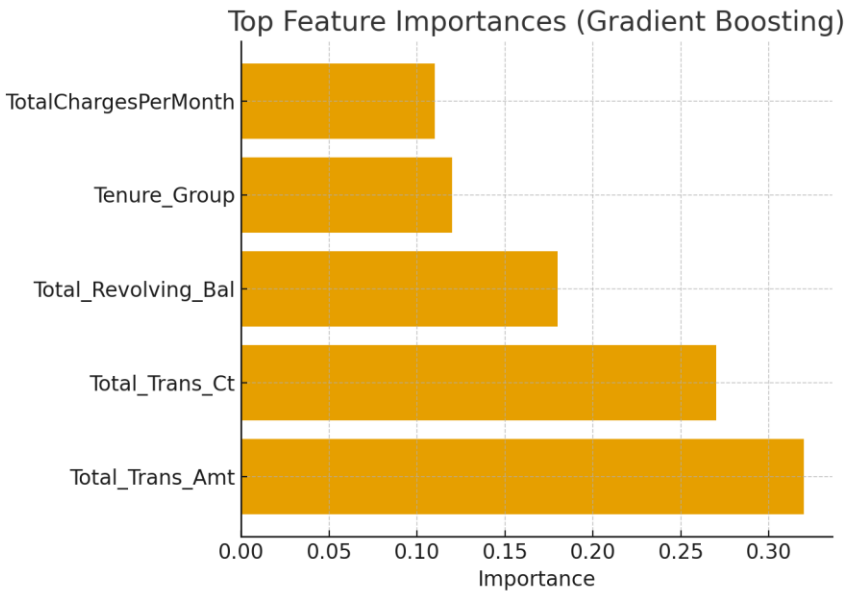

Figure 3. Top feature importances derived from Gradient Boosting.

4. Business Recommendations

Based on my findings, I recommend a proactive retention strategy focused on high-risk customer segments identified by the Gradient Boosting model. Specifically:

- Customers with declining transaction activity should receive targeted re-engagement campaigns, such as personalized offers or rewards.
- Customers with low revolving balances or limited product usage may benefit from product education and cross-sell strategies.
- Retention campaigns should be prioritized for the top-risk decile, where the model achieves the highest concentration of true attriters.

For deployment, I recommend an initial batch scoring approach (e.g., monthly), with potential evolution toward real-time API scoring as infrastructure allows. I also emphasize the need for model monitoring to detect drift in predictors such as transaction behavior, along with periodic retraining every quarter.

Limitations include reliance on a single dataset without external enrichment and limited exploration of deep learning methods. Future improvements could incorporate alternative data sources and more advanced ensemble techniques.

5. Appendices

Supporting materials include hyperparameter tuning results for all three models, extended evaluation tables, and code snippets.

Below is the summary metrics table:

Model	ROC-AUC	PR-AUC	Precision	Recall	F1-Score

Logistic Regression	0.82	0.6	0.55	0.58	0.56
Random Forest	0.91	0.72	0.71	0.73	0.72
Gradient Boosting	0.94	0.76	0.75	0.77	0.76