Predicting Customer Attrition: Data-Driven Insights for Retention

**Introduction**

Predicting customer churn enables organizations to take proactive steps in retaining valuable clients. This report focuses on analyzing banking customer data to build a model capable of identifying those likely to leave. The process involves data preparation, exploratory analysis, handling class imbalance, and training predictive models. Emphasis is placed on both accuracy and interpretability to support actionable business decisions.

**1. Data Understanding & Cleaning**

The dataset contains 10,127 customer records, each representing a unique individual with demographic, account, and transaction-related attributes. No missing values were present, ensuring complete data for analysis.

The column CLIENTNUM, a unique identifier for each customer, was dropped to avoid data leakage, as it carries no predictive value. The target variable Attrition\_Flag was mapped to numerical format (0 = Existing Customer, 1 = Churned Customer) to enable compatibility with machine learning models that require numeric inputs.

These initial preprocessing steps ensured the data was clean, structured, and ready for exploratory analysis and model development.

We’re predicting whether a customer **churned (left the bank)** — a key business KPI. This transformation prepares the data to train a churn prediction model.

**2. Data Exploration**

Visualized the categorical features wrt attrition. This step visualizes the **mean churn rate** (Attrition\_Flag = 1) for each category within every categorical feature (like Gender, Marital Status, etc.).

**Plots for all the categorical feature:**

A chart of different shades of pink

AI-generated content may be incorrect.A chart of different shades of pink

AI-generated content may be incorrect.

A chart of different shades of pink

AI-generated content may be incorrect.

|  |  |
| --- | --- |
| Feature | Notable Insight |
| Gender | Female customers show **higher churn rates** than male. Targeted retention for female segment could be considered. |
| Education\_Level | Churn increases with education — **Doctorate holders churn the most**. Perhaps they expect better services? |
| Marital\_Status | Singles and unknown marital statuses show higher churn — maybe they’re less financially stable or more transient. |
| Income\_Category | Low-income groups (<$40K) and very high-income groups ($120K+) have **above-average churn**. Indicates polar ends may be dissatisfied for different reasons. |
| Card\_Category | **Platinum cardholders have the highest churn (25%)**, even though they're premium customers — could indicate poor high-tier service quality. |

**Plots for some of the numerical features:**

In the plots, if orange and blue peaks are far apart, that feature helps separate churners from stayers.

A screenshot of a graph

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

|  |  |
| --- | --- |
| Feature | Notable Insight |
| Total\_Trans\_Amt | Customers with lower total transaction amounts are more likely to churn. Indicates inactive or disengaged users. Retention strategies can focus on increasing engagement. |
| Total\_Trans\_Ct | Fewer transactions are associated with higher attrition. This strongly signals that lower activity correlates with churn. Could suggest loss of interest or better alternatives elsewhere. |
| Total\_Revolving\_Bal | Attrited customers tend to have slightly lower revolving balances, implying they might not be leveraging credit as much — possibly due to dissatisfaction or reduced usage. |
| Total\_Ct\_Chng\_Q4\_Q1 | Lower changes in transaction count between quarters are linked to churn. Lack of growing engagement or declining activity can be a red flag. |
| Avg\_Utilization\_Ratio | Customers with very low utilization ratios are churning more. Possibly indicates they have better offers elsewhere or aren’t incentivized to use the credit facility. |
| Total\_Amt\_Chng\_Q4\_Q1 | A significant drop in the total transaction amount from Q4 to Q1 may be a strong early signal of churn. Highlights the importance of monitoring spend patterns over time. |

Customers who are less financially active (fewer or smaller transactions, declining activity) are more prone to churn. This suggests that **customer engagement** and **credit usage patterns** can be powerful indicators of dissatisfaction or disengagement. Retention efforts should focus on reviving low-usage users through personalized offers, better rewards, or service outreach.

A screenshot of a computer screen

AI-generated content may be incorrect.

**A pie chart with a number of different colored circles

AI-generated content may be incorrect.Checking for Class Imbalance**

To understand the distribution of the target variable (Attrition\_Flag), a pie chart was plotted to visually assess class balance. This step is crucial because many machine learning models assume roughly equal class representation. From the visualization, it was observed that approximately **86% of the records represent non-churned (existing) customers**, while only **14% represent churned (attrited) customers**. This indicates a significant **class imbalance** in the dataset.

Such an imbalance can be problematic during model training, as standard algorithms tend to favor the majority class. As a result, the model may achieve high overall accuracy while performing poorly on the minority class—in this case, failing to correctly predict churn. To address this, class-balancing techniques like **SMOTE (Synthetic Minority Over-sampling Technique)** was used. These will ensure that the model is exposed to enough examples of the minority class to learn meaningful patterns, ultimately leading to **better generalization and fairness** in predictions.

**Step: Encoding, Splitting, Balancing, and Modeling with Random Forest**

**1. One-Hot Encoding (OHE)**

The pandas function pd.get\_dummies() was applied to convert categorical features into numerical binary flags using **one-hot encoding**, with drop\_first=True to avoid multicollinearity. This step was essential as tree-based models like Random Forest still require numerical input and cannot natively handle strings or object data types.

**2. Train-Test Split**

To validate model generalizability, we split the dataset into 80% training and 20% testing data using train\_test\_split, with **stratification** on the target variable to maintain the original class ratio (i.e., preserve churn distribution across both sets).

**3. Handling Class Imbalance with SMOTE**

Given our earlier observation of a **significant class imbalance** (86% existing vs. 14% churned customers), we used **SMOTE (Synthetic Minority Over-sampling Technique)** on the training data. SMOTE generates synthetic examples of the minority class to ensure that the model is not biased toward the majority class during learning. This helped the model better understand churn-related patterns.

**4. Model Training and Evaluation**

After data preprocessing and exploratory analysis, two machine learning models—**Random Forest** and **XGBoost**—were trained and evaluated to predict customer churn. The evaluation focused on key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to ensure both overall and class-specific predictive performance. Additionally, interpretability methods were employed to understand which features contributed most to the model’s decisions.

**4.1 Random Forest Classifier**

The Random Forest algorithm was selected as the initial model due to its ability to handle non-linear relationships, robustness to outliers, and effectiveness with imbalanced datasets. The classifier was configured with 300 trees (n\_estimators=300) and no depth limitation (max\_depth=None) to allow trees to grow fully and capture complex feature interactions.

**Purpose of Use**:  
Random Forest serves as a reliable baseline due to its ensemble structure that reduces overfitting and handles high-dimensional data without extensive preprocessing.

**Results on Test Set:**

|  |  |
| --- | --- |
| Metric | Score |
| Accuracy | 95.9% |
| Precision (Churned) | 85.4% |
| Recall (Churned) | 85.4% |
| F1-Score (Churned) | 85.4% |
| ROC-AUC | 0.986 |

Confusion Matrix:

|  |  |
| --- | --- |
| 1472 | 36 |
| 36 | 210 |

**Interpretation**:  
The model achieved high accuracy and balanced precision-recall for the minority churn class. However, with 36 false negatives (churned customers misclassified as retained), it leaves room for improvement, especially in sensitive customer segments.

**4.2 XGBoost Classifier – Base Model**

To improve recall and overall model robustness, the **XGBoost algorithm** was introduced. XGBoost offers regularization to combat overfitting and gradient boosting capabilities, making it well-suited for structured churn datasets.

The base XGBoost model was configured with:

* n\_estimators=100
* max\_depth=3
* learning\_rate=0.1
* subsample=0.8
* colsample\_bytree=0.8

**Cross-Validation Strategy**:  
5-fold **Stratified K-Fold Cross-Validation** was applied to ensure balanced class distribution in each split, making the evaluation statistically reliable.

**Cross-validated ROC-AUC scores:**

Mean ROC-AUC: 0.9992

**Evaluation on Test Set:**

|  |  |
| --- | --- |
| Metric | Score |
| Accuracy | 95.7% |
| Precision (Churned) | 81.3% |
| Recall (Churned) | 90.2% |
| F1-Score (Churned) | 85.5% |
| ROC-AUC | 0.987 |

**Confusion Matrix:**

|  |  |
| --- | --- |
| 1457 | 51 |
| 24 | 222 |

**Observations**:  
The base XGBoost model significantly improved **recall**, correctly identifying 222 out of 246 churned customers, which is critical in churn prediction tasks. Although precision dropped slightly, this is acceptable when the business goal is to reduce customer loss risk.

**4.3 XGBoost – Tuned Model (RandomizedSearchCV)**

To further improve model performance, **hyperparameter tuning** was performed using RandomizedSearchCV with 30 parameter combinations and 5-fold cross-validation. The tuning included parameters such as:

* Number of trees (n\_estimators)
* Tree depth (max\_depth)
* Learning rate
* Subsampling ratios
* Regularization terms (reg\_alpha, reg\_lambda)
* Column sampling (colsample\_bytree)
* Minimum loss reduction (gamma)

**Best Parameters Identified:**

{

'subsample': 1.0,

'reg\_lambda': 1.5,

'reg\_alpha': 1,

'n\_estimators': 300,

'max\_depth': 7,

'learning\_rate': 0.1,

'gamma': 0,

'colsample\_bytree': 0.6

}

**Final Model Evaluation:**

|  |  |
| --- | --- |
| Metric | Score |
| Accuracy | 97.2% |
| Precision (Churned) | 89.6% |
| Recall (Churned) | 90.7% |
| F1-Score (Churned) | 90.1% |
| ROC-AUC | 0.993 |

**Confusion Matrix:**

|  |  |
| --- | --- |
| 1482 | 26 |
| 23 | 223 |

**Insights**:  
The tuned model showed the best performance overall. The **recall improved slightly**, and **precision increased substantially**, reducing false positives. The model was able to confidently and accurately detect churned customers, making it ideal for production use.

**4.4 Feature Importance & Explainability**

To ensure the model’s decisions were transparent, several interpretability techniques were applied:

**a. SHAP Summary Plot**

A SHAP (SHapley Additive exPlanations) plot visualized the impact of each feature on model predictions. Key observations:

* **Total\_Trans\_Ct** and **Total\_Trans\_Amt** had the strongest negative SHAP values — higher transaction counts and amounts significantly reduced churn probability.
* **Total\_Revolving\_Bal**, **Total\_Ct\_Chng\_Q4\_Q1**, and **Avg\_Utilization\_Ratio** were also influential indicators.
* Categorical features such as **Marital\_Status\_Single** and **Education\_Level** contributed moderately, typically increasing churn risk.

**b. Gain-Based Feature Importance (XGBoost Native)**

This bar chart showed how much each feature contributed to model performance (based on average gain). Top features aligned with SHAP results, confirming their influence on splits.

**c. Model-Based Feature Scores (Random Forest)**

The Random Forest’s feature\_importances\_ attribute ranked features by their contribution to reducing Gini impurity. Again, transaction behavior and revolving balance dominated.

**Common Top Predictors Across All Models:**

* Total\_Trans\_Ct
* Total\_Trans\_Amt
* Total\_Revolving\_Bal
* Total\_Ct\_Chng\_Q4\_Q1
* Avg\_Utilization\_Ratio
* Total\_Relationship\_Count

These consistent findings strengthen the confidence in transactional behavior as a core driver of churn.

**4.5 ROC Curve Analysis**

The ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) metric was chosen to evaluate model performance due to its ability to measure a classifier’s ability to distinguish between classes across all threshold levels. In the context of churn prediction, where correctly identifying the minority class (churned customers) is crucial, ROC-AUC provides a threshold-independent assessment of how well the model can differentiate between churned and non-churned customers. A higher AUC value indicates stronger separability, meaning the model is more effective at ranking churned customers above retained ones. This is especially valuable when false negatives (missed churn cases) carry significant business cost, and the decision threshold may be adjusted later based on risk tolerance.

A graph of a curve

AI-generated content may be incorrect.

The curve showed a strong upward arc with an AUC of 0.993, indicating excellent classification capability. This suggests that the tuned XGBoost model consistently ranks churned customers higher in risk than non-churned ones across various thresholds. The near-perfect separation reflects the model’s robustness in capturing complex patterns in transactional behavior and customer attributes. Such a high AUC score implies minimal trade-off between true positive rate and false positive rate, affirming the model’s reliability for deployment in churn risk scoring systems.

**Conclusion**

The customer churn analysis provided a comprehensive understanding of behavioral patterns linked to attrition, and successfully identified customers most at risk of leaving. Through data exploration and modeling, several key insights emerged. Customers with lower transaction activity, declining engagement, and minimal credit utilization were consistently flagged as high churn risks. These behavioral indicators, captured across various features such as Total\_Trans\_Ct, Total\_Trans\_Amt, and Total\_Revolving\_Bal, were found to be strong predictors across all models.

Three machine learning models were evaluated—Random Forest, base XGBoost, and a tuned XGBoost. While the Random Forest model provided solid baseline performance with a ROC-AUC of 0.986 and good balance between precision and recall, the XGBoost algorithm offered significant improvements. The tuned XGBoost model, optimized via RandomizedSearchCV, delivered the best results with a **ROC-AUC of 0.993**, **recall of 90.7%**, and **precision of 89.6%** on the churn class. This means it not only accurately detects most churned customers but does so with fewer false alarms. In practical terms, it allows retention teams to act confidently on the predictions.

Feature importance analysis using SHAP and gain-based methods confirmed that transactional metrics—particularly frequency and volume of transactions—are the most influential churn drivers. This reinforces the business case for focusing on **engagement-based retention strategies**, such as personalized offers or activity-triggered outreach.

From a business perspective, the tuned XGBoost model stands out as the most suitable candidate for deployment. It offers high interpretability, strong generalization, and a robust ability to identify customers at churn risk before it's too late. Importantly, the use of techniques like SMOTE helped address class imbalance, ensuring the model doesn’t overlook minority churn cases.

**Key Recommendations**:

* Prioritize at-risk customers who show declining transaction volume or frequency.
* Use churn scores from the model to trigger proactive interventions.
* Continuously monitor model performance and retrain with fresh data to adapt to customer behavior changes.

Overall, this analysis provides a reliable, explainable, and effective churn prediction framework that can support strategic decision-making and customer retention efforts.