**Customer Churn Prediction of Union Bank of India (UBI)**

*Understanding and Addressing Customer Attrition through Data Analysis*

**INFO 5307 Section 021 - Knowledge Management Tools and Technologies (Fall 2023 1)**

**Project - Part 3**

**Presented by:**  
Sai Tulasi Kolapudi

**Date:**  
12/05/2023

**Table of Contents**

1. **Introduction**
   * Brief Overview of Customer Churn
   * Objectives of the Analysis
2. **Data Overview**
   * Dataset Description
   * Preprocessing and Feature Selection
3. **Exploratory Data Analysis**
   * Key Insights and Visualizations
   * Customer Demographics and Churn Rates
4. **Feature Engineering**
   * Introduction to New Features
   * Relevance to Churn Prediction
5. **Data Preparation for Model Fitting**
   * Data Organization and Transformation
   * One-Hot Encoding and Normalization
6. **Model Development and Evaluation**
   * Overview of Predictive Models
   * Performance Metrics and Results
7. **Conclusions**
   * Summary of Findings
   * Recommendations and Strategies

**Introduction**

**Brief Overview of Customer Churn**

Customer churn, a critical challenge in the banking sector, represents the phenomenon of clients discontinuing their relationship with a financial institution. In the context of Union Bank of India (UBI), understanding churn is vital as retaining existing customers is typically more cost-effective than acquiring new ones. This analysis aims to unravel the underlying factors contributing to churn at UBI, offering insights into customer behavior and retention.

**Objectives of the Analysis**

Our study at UBI is tailored to achieve the following key objectives:

1. **Identification and Visualization of Churn Factors:**
   * We will thoroughly analyze UBI's customer dataset to pinpoint and visually present the factors that lead to customer churn. This exploratory process will illuminate various demographics, account features, and behavioral patterns that might influence a customer's decision to leave the bank.
2. **Development of a Predictive Churn Models:**
   * Our central goal is to construct an effective prediction model that serves two primary functions:
     + **Churn Classification:** Differentiating customers who are likely to churn from those who will likely continue with UBI. This classification is crucial for understanding the distinct characteristics of each group.
     + **Probability Assessment of Churn:** Depending on the model’s accuracy and reliability, we aim to implement a probability-based approach. This methodology will allow UBI’s customer service teams to strategically target customers who are on the verge of churning but can potentially be retained through focused and timely interventions.

Through this analytical venture, we intend to equip Union Bank of India with strategic, data-driven tools to enhance customer retention and mitigate churn effectively.

**Data Overview**

**Dataset Review and Preparation**

In this pivotal section, we explore the structure of our dataset to understand its input space and prepare it for the exploratory and predictive tasks outlined in the introduction.

**Dataset Characteristics:**

* **Shape and Size:** The dataset comprises 10,000 records spanning across 14 different attributes. This rich dataset provides a comprehensive view of the customer's profile within Union Bank of India.
* **Data Integrity:** A preliminary review reveals a complete dataset with no missing values, a rare but advantageous find in data analytics.
* **Unique Counts and Relevance:** Analysis of unique counts per variable helps us determine the necessity of each attribute. For instance:
  + **RowNumber**, **CustomerId**, and **Surname**: These attributes are unique to each customer and do not contribute to the predictive model. They are dropped to avoid individual customer profiling and focus on more generalizable features.

**Initial Data Manipulation:**

* **Dropping Irrelevant Columns:** Based on the above assessment, columns such as **RowNumber**, **CustomerId**, and **Surname** are excluded from the analysis.
* **Data Preview:** A glimpse into the remaining dataset reveals key attributes like **CreditScore**, **Geography**, **Gender**, **Age**, and more, each offering potential insights into the churn behavior.

**Critical Observations and Queries:**

* **Data Snapshot Nature:** It's apparent that the data represents a snapshot at a specific point in time. This raises questions about the context and relevance of the data, such as the date of the balance and its significance.
* **Customer Exit vs. Balance:** Notably, some customers who have exited still show a balance in their accounts. This intriguing observation necessitates a deeper understanding of what 'exit' means in this context and whether it pertains to specific products or the bank as a whole.
* **Active Membership Definition:** The term 'active member' in the dataset is somewhat ambiguous. Clarifying its meaning and possibly supplementing it with more granular transaction data could enhance the model's effectiveness.
* **Product Details:** A more detailed breakdown of the products used by customers, rather than just a count, could yield richer insights for the predictive model.

**Data Type Assessment:**

* The dataset is primarily composed of categorical variables, with a total of 5 continuous variables. This composition will influence our approach to data preprocessing and model selection.

**Exploratory Data Analysis**

In this analysis, we examine the relationship between various attributes and the customer's exit status from Union Bank of India. Understanding these relationships helps in identifying the traits of customers who are more likely to churn.

**Churn Proportion Analysis**

A pie chart with a number of different colored circles with Crust in the background

Description automatically generated  
  
Our initial observation reveals that approximately 20% of customers have churned. This sets a baseline for our predictive models, which aim to identify this specific subset with high precision, given its significance to the bank's customer retention strategies.

A group of blue and orange bars

Description automatically generated**Demographic Insights**

Several demographic factors stand out:

* France constitutes the majority in the dataset, yet it has a lower churn rate compared to other regions, hinting at a potential mismatch in service allocation.
* Females are more likely to churn than males, indicating possible gender-specific service gaps.
* Customers with credit cards do not exhibit a significantly different churn rate, suggesting that credit card ownership alone is not a churn predictor.
* Inactive members show a higher churn rate, which calls for initiatives to engage these customers more effectively.

**Behavioral and Financial Attributes**:

A group of graphs showing different sizes and colors

Description automatically generated

Credit score does not show a marked difference between churned and retained customers.

* Older customers are more prone to churn, suggesting a need to tailor services to this demographic.
* Customers with either very low or high tenure are more likely to churn, highlighting critical touchpoints for retention efforts.
* The loss of customers with significant bank balances is particularly concerning, as it directly impacts the bank's lending capital.

**Correlation Analysis**

A purple and white squares with black and white squares

Description automatically generated  
The heatmap analysis underscores 'Age' and 'IsActiveMember' as notable correlates of churn, with 'Balance' also playing a significant role. These findings can guide the bank in refining retention strategies, such as targeting older customers and boosting customer engagement initiatives.

A blue and orange rectangular bars with text

Description automatically generatedA graph of different colored squares

Description automatically generated**Geographical and Gender-Based Churn Rates**

Geographically, Germany shows the highest churn rate, which warrants an in-depth evaluation of the bank's market strategy in the region. Gender-wise, the higher churn rate among females could indicate the necessity for gender-specific retention approaches.

**Product Holdings and Churn**  
A graph of a distribution of products

Description automatically generated  
The analysis reveals a higher churn among customers with a single product and those with more than two products. This suggests a potential 'sweet spot' in product holdings that correlates with customer loyalty.

**Age Distribution Among Churned and Retained Customers**

A graph of a distribution of customers

Description automatically generated

The age distribution highlights a peak churn risk in the late 40s to early 50s age bracket, while younger customers are more likely to be retained. Targeting middle-aged customers with retention strategies could be beneficial in reducing overall churn rates.

**Feature Engineering**

In predictive modeling, feature engineering is a critical step that involves creating new variables from existing data to improve model performance. For Union Bank of India's churn prediction, we've introduced new features to gain a deeper understanding of the factors that may influence a customer's decision to leave the bank.

A graph of a diagram

Description automatically generated with medium confidenceA chart with blue and orange squares

Description automatically generated**Development of New Features**

1. **BalanceSalaryRatio:**  
   The ratio of a customer's bank balance to their estimated salary has emerged as a significant indicator. Customers with a higher balance relative to their salary are more likely to churn. This pattern is concerning because it could signify the loss of customers who have substantial financial assets, affecting the bank's loan capital availability. This new feature, 'BalanceSalaryRatio,' has therefore been engineered to pinpoint customers who may require targeted retention strategies.
2. **TenureByAge:**  
   Recognizing that tenure at the bank is often correlated with age, we've normalized tenure by age to create 'TenureByAge.' This feature allows us to compare tenure on a scale that accounts for the customer's lifecycle stage, potentially highlighting different patterns of churn among various age groups.
3. **CreditScoreGivenAge:**  
   We have also calculated 'CreditScoreGivenAge' to assess creditworthiness across different ages. This feature could reveal how credit behavior and customer lifecycle stage influence the likelihood of churn.

**Data Segmentation for Model Training**

The dataset has been split into training and test sets, with 8,000 records for training and 2,000 for testing. This segmentation is critical for validating the performance of our predictive models on unseen data.

**Insights from Feature Engineering**  
The preliminary analysis of these new features suggests that:

* Customers with a higher 'BalanceSalaryRatio' show an increased propensity to churn, which may require the bank to reassess their product offerings or customer engagement for high-balance customers.
* 'TenureByAge' might reveal insights into loyalty patterns across different age demographics.
* 'CreditScoreGivenAge' will help in understanding how a customer's age affects their credit score and their decision to stay with the bank.

These engineered features will be incorporated into our predictive models to enhance their accuracy and provide more nuanced insights into customer churn at Union Bank of India.

**Data Preparation for Model Fitting**

Before fitting our models, a meticulous data preparation process is carried out to ensure the data is in the optimal form for our predictive algorithms.

**Organizing Data by Type**  
For efficiency and clarity, we've organized the columns by data type. Continuous variables like 'CreditScore', 'Age', and 'Balance' are separated from categorical variables such as 'HasCrCard' and 'IsActiveMember'. This organization facilitates targeted preprocessing steps for different data types.

**Variable Transformation**  
Categorical variables are transformed to ensure they correctly represent the presence or absence of a feature:

* 'HasCrCard' and 'IsActiveMember' are recoded from 0 (no card or inactive membership) to -1 to denote the absence of these attributes explicitly.

**One-Hot Encoding Categorical Variables**  
The categorical variables 'Geography' and 'Gender' are one-hot encoded, creating binary columns for each category. This encoding transforms categorical inputs into a format that can be provided to machine learning algorithms to do a better job in prediction.

**Normalization of Continuous Variables**  
Continuous variables are normalized using Min-Max scaling, which scales the range of data to [0, 1]. This scaling is crucial for models that are sensitive to large variations in data scales.

**Data Preparation Pipeline for Test Data**  
A data preparation pipeline is established for the test set, mirroring the process performed on the training data:

* New features are added: 'BalanceSalaryRatio', 'TenureByAge', and 'CreditScoreGivenAge'.
* The test data is reordered to match the training data structure.
* Categorical variables in the test set are also transformed and one-hot encoded.
* Continuous variables are scaled using the Min-Max values derived from the training set to maintain consistency.

**Ensuring Consistency**  
It's essential to ensure that all features used in the training set are present in the test set. Therefore, any missing columns in the test set are added with a default value of -1, indicating their absence.

This rigorous data preparation ensures that our models are trained on data that is clean, well-structured, and representative of the real-world scenarios that Union Bank of India may face. With these steps completed, we can proceed with confidence to the model fitting and selection phase.

**Model Development and Evaluation**

To predict customer churn at Union Bank of India, we have developed and evaluated several predictive models. Each model's performance has been carefully assessed to ensure that the most accurate and reliable one is chosen for deployment.

**Logistic Regression**

A graph of a curve

Description automatically generated  
The logistic regression model yielded an AUC of 0.79, with a training accuracy of 82% and a test accuracy of 80%. While precision was notably higher for the retained customers, recall for churned customers was relatively low, indicating that the model is conservative in predicting churn.

**Decision Tree**  
A line graph with a blue line

Description automatically generated  
The decision tree model achieved an AUC of 0.70. It shows perfect training accuracy of 1, a common sign of overfitting, and a test accuracy of 79%. The model exhibits balanced precision and recall across classes, suggesting it could be more effective in identifying churned customers compared to logistic regression.

**Random Forest**  
A graph of a curve

Description automatically generated  
The random forest classifier outperformed the previous models with an AUC of 0.86. It also showed perfect training accuracy and a high test accuracy of 86%, indicating a strong performance. However, similar to the decision tree, it might be overfitting, as indicated by the perfect training accuracy.

**Gaussian Naive Bayes**  
A graph of a curve

Description automatically generated  
The Gaussian Naive Bayes model achieved an AUC of 0.80, with a training accuracy of 82% and test accuracy of 81%. This model provided a good balance between precision and recall, suggesting that it is capable of effectively distinguishing between churned and retained customers.

**Evaluation and Performance Metrics**  
The models were evaluated based on their accuracy, precision, recall, and AUC - metrics that together provide a comprehensive view of each model's performance. The Random Forest model, despite its potential overfitting, shows the most promise due to its high AUC and balanced accuracy. However, its complexity and the possibility of overfitting must be considered when selecting the final model.

These insights will guide the bank in selecting the most suitable model for its churn prediction tool. The chosen model will play a crucial role in identifying at-risk customers and shaping the bank's retention strategies.

**Conclusions**

**Summary of Findings**

Our comprehensive study on customer churn at Union Bank of India (UBI) has revealed several critical insights. Through meticulous exploratory data analysis, we identified key factors that contribute to customer churn. The demographic analyses suggested that geography and gender play a significant role in churn likelihood, with a notably higher churn rate observed in Germany and among female customers. The feature engineering process provided deeper insights, unveiling the BalanceSalaryRatio, TenureByAge, and CreditScoreGivenAge as influential predictors of churn.

The predictive modeling phase assessed four different algorithms: Logistic Regression, Decision Tree, Random Forest, and Gaussian Naive Bayes. Each model was evaluated based on accuracy, precision, recall, and ROC-AUC score. The Random Forest model emerged with the highest AUC score, suggesting a strong capacity to distinguish between churned and retained customers. However, the potential for overfitting observed in this model warrants a cautious approach to its deployment.

**Recommendations and Strategies**

Based on our findings, we recommend the following strategies to UBI:

1. **Tailored Customer Retention Programs:**
   * Develop targeted retention initiatives for high-risk customer segments, particularly focusing on females and customers from regions with higher churn rates like Germany.
2. **Enhanced Engagement for At-Risk Age Groups:**
   * Implement engagement strategies tailored for middle-aged customers who exhibit a higher propensity to churn. Personalized financial planning services could be beneficial.
3. **Optimization of Product Portfolios:**
   * Reassess the bank's product offerings to find the 'sweet spot' for customer retention. Our analysis suggests that customers with either too few or too many products are more likely to churn.
4. **Reactivation Strategies for Inactive Members:**
   * Design programs to convert inactive members into active customers. Our data indicates that inactivity is a strong predictor of churn.
5. **Careful Model Selection and Continuous Monitoring:**
   * Deploy the Random Forest model with regular assessments for overfitting and performance on new data. Consider ensemble approaches that might combine the strengths of multiple models to improve prediction accuracy.
6. **Further Exploration of Financial Health Indicators:**
   * Investigate financial indicators such as BalanceSalaryRatio in more depth to understand the financial health of customers and to tailor personalized banking advice and product offerings.
7. **Ongoing Data Analysis:**
   * Continue to analyze customer data regularly to identify emerging trends and to update the churn prediction models accordingly.

By implementing these recommendations, UBI can expect to strengthen customer loyalty, reduce churn rates, and maintain a competitive edge in the dynamic banking landscape.