Project Documentation: "Preventing Bank Customer Churn"

1. Introduction

The objective of this project is to analyze and predict customer churn in a bank using exploratory data analysis (EDA) and machine learning techniques. The primary goals are:

1. Identify and visualize factors contributing to customer churn.
2. Build a prediction model capable of classifying whether a customer is likely to churn and, if possible, assign a probability to facilitate targeted preventive efforts.

Project Author: Keldine Malit

2. Data Set Review & Preparation

2.1 Required Libraries

The project utilizes the following libraries for data manipulation and visualization in Python:

```
pythonCopy code
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

2.2 Data Loading and Initial Inspection

The dataset, named 'Churn_Modelling.csv', is loaded into a Pandas DataFrame, and its structure is briefly reviewed. No missing values are observed, and unnecessary columns (RowNumber, CustomerId, Surname) are dropped.

2.3 Data Exploration

Exploratory Data Analysis (EDA) is performed to gain insights into the dataset. Categorical and continuous variables are analyzed separately.

2.3.1 Categorical Variables

- Geographic distribution: Majority of data is from France, and the churn proportion is inversely related to the customer population.
- Gender: The proportion of female customers churning is higher than that of males.
- Has Credit Card: Customers with credit cards are more likely to churn.
- Active Membership: Inactive members have a higher churn rate.

2.3.2 Continuous Variables

- Credit Score: No significant difference in credit score distribution between retained and churned customers.
- Age: Older customers are more likely to churn.
- Tenure: Extreme tenures (very short or very long) correlate with higher churn.
- Balance: Customers with significant bank balances are more likely to churn.
- NumOfProducts and EstimatedSalary show no significant impact on churn.

3. Feature Engineering

New features are introduced to improve model performance:

- **BalanceSalaryRatio:** Ratio of bank balance to estimated salary.
- **TenureByAge:** Standardized tenure over age.
- **CreditScoreGivenAge:** Credit score given age to account for credit behavior over adult life.

4. Data Preparation for Model Fitting

- Columns are arranged by data type for easier manipulation.
- Categorical variables are one-hot encoded, and continuous variables are min-max scaled.

5. Model Fitting and Selection

Several models are fit to the data, including Logistic Regression, SVM (RBF and Polynomial kernels), Random Forest, and XGBoost. Grid search is used for hyperparameter tuning.

6. Best Model Selection and Evaluation

The Random Forest model is selected as it strikes a good balance between precision and recall for predicting customer churn.

6.1 Model Evaluation Metrics

Precision on 1's (churned customers): 0.88 (88% of predicted churners actually churn). Recall on 1's: 0.53 (53% of actual churners correctly identified).

6.2 ROC Curve

The ROC curve illustrates the model's true positive rate against the false positive rate.

7. Conclusion

The Random Forest model demonstrates good performance on the training set, and its precision and recall metrics on the test set are promising. However, ongoing retraining with new data is recommended to improve accuracy over time. The model's ability to identify potential churners can help the bank focus preventive efforts on at-risk customers.