**Capstone Project Title: European Bank’s Customer Satisfaction Analysis & Churn Prediction**

**Name: Sathiswaran Sangaran**

## 1. Introduction & Executive Summary

**1.1 Problem Statement:**

Customer churn is a critical issue for businesses in competitive markets. Understanding and predicting churn can help companies strategize proactive measures on customer retention, improve customer satisfaction, and increase revenue while acquiring new customers. The aim is to predict customer churn (whether a customer will exit the bank or not). This is a binary classification problem where the target variable is Exited.

**1.2 Goal:**

* Analyze customer data from the bank to identify patterns and trends that can help in improving customer retention and satisfaction.
* Develop models to predict customer churn using various customer attributes.
* Provide actionable insights for bank management and marketing teams.
* Enable the bank to make informed data-driven decisions to enhance their services and reduce customer churn rates by understanding the factors that influence customer behavior.

**1.3 Audience:**

This analysis is valuable for bank management and marketing teams. Understanding churn can help in devising strategies to retain customers, thereby increasing profitability.

**1.4 Data Source:**

Data Source: <https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers>

**1.5 Success Metric:**

The primary success metric will be Accuracy. Additionally, we'll evaluate Precision, Recall, F1 score, and AUC (Area Under the ROC Curve) to get a comprehensive view of model performance.

**1.6 Data Overview:**

* Data Types: Both numerical (e.g., CreditScore, Age, Balance) and categorical (e.g., Geography, Gender).
* Distributions: We'll visualize these distributions to understand data spread and identify potential outliers.
* Missing Data: Identify and address any missing data.

**1.7 Data Details:**

10,000 rows x 14 columns

|  |  |
| --- | --- |
| **Column** | **Description** |
| RowNumber | corresponds to the record (row) number and has no effect on the output. |
| CustomerId | contains random values and has no effect on customer leaving the bank. |
| Surname | the surname of a customer has no impact on their decision to leave the bank. |
| CreditScore | can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank. |
| Geography | a customer’s location can affect their decision to leave the bank. |
| Gender | it’s interesting to explore whether gender plays a role in a customer leaving the bank. |
| Age | this is certainly relevant, since older customers are less likely to leave their bank than younger ones. |
| Tenure | refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank. |
| Balance | also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances. |
| NumOfProducts | refers to the number of products that a customer has purchased through the bank. |
| HasCrCard | denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank. |
| IsActiveMember | active customers are less likely to leave the bank. |
| EstimatedSalary | as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries. |
| Exited | whether or not the customer left the bank. |

* 1. **Metrics**

The primary metric for evaluating model performance is Accuracy, supplemented by Precision, Recall, F1 score, and ROC AUC to comprehensively assess model effectiveness in predicting churn.

**1.9 Findings**

|  |  |
| --- | --- |
| **Continuous Variables:** | **Categorical Variables:** |
| Credit Score: Slightly lower median for exited customers. Median credit score for customers who exited: 611.0 compared to 637.0 for those who did not exit.  Age: Older customers are more likely to exit. Mean age of customers who exited is 44.14 years compared to 39.75 years for those who did not exit.  Balance: Higher median balance for exited customers. Range of balance for customers who exited: 0 𝑡𝑜 250,898.09 with a median balance of $116,760.6  Estimated Salary: Similar distributions, suggesting it might not be a significant factor for churn. | Geography: Higher exits in Germany, lower in France. We can compare this percentage with the overall exit rate to understand if customers from Germany are more likely to exit.  Gender: Higher exit proportion for females.  Tenure and Number of Products: No strong trend with tenure; higher exits with only one product.  Has Credit Card: No significant difference.  Is Active Member: Lower exits for active members. |

**1.10 Risks/Limitations/Assumptions**

**Data Imbalance:** The dataset exhibits imbalance in the target variable (Exited), which was addressed using SMOTE (Synthetic Minority Over-sampling Technique).

**Model Selection:** While Random Forest and Gradient Boosting performed well, other models like Logistic Regression and K-Nearest Neighbors were also considered but showed comparatively lower performance.

**Data Quality:** Assumptions were made regarding missing data imputation and feature engineering decisions, which could impact model outcomes.

## 2. Statistical Analysis

**2.1 Implementation**

The analysis was conducted using the Python programming language, leveraging popular data science libraries including Pandas, NumPy, Matplotlib, Seaborn, and scikit-learn.

**2.2 Evaluation**

**2.2.1 Data Preprocessing:**

* **Handling Missing Values:** Missing values in categorical columns 'Geography' and 'Gender' were filled with their respective mode values.
* **Encoding Categorical Variables:** Label Encoding was applied to binary categorical variables, while One-Hot Encoding was used for multi-class categorical features.
* **Scaling Numerical Features:** StandardScaler was used to normalize numerical features to ensure they are on a similar scale, improving model performance.

**2.2.2 Exploratory Data Analysis (EDA):**

**Distributions:** Histograms and boxplots were used to visualize the distribution of continuous variables such as CreditScore, Age, Balance, and EstimatedSalary, highlighting patterns and potential outliers.

**Relationships:** Countplots and correlation heatmaps helped understand relationships between categorical variables (e.g., Geography, Gender, Tenure) and the target variable 'Exited'.

**2.2.3 Model Training:**

**Data Splitting:** The dataset was split into training and testing sets to evaluate model performance on unseen data.

**Addressing Class Imbalance:** Synthetic Minority Over-sampling Technique (SMOTE) was employed to balance the classes, preventing the model from being biased towards the majority class.

**Machine Learning Algorithms:** Various models were trained including Logistic Regression, Random Forest, Gradient Boosting, AdaBoost, and K-Nearest Neighbors (KNN).

**Cross-Validation:** Cross-validation techniques were used to ensure robust model evaluation, preventing overfitting and providing reliable performance metrics.

**2.3 Inference**

**2.3.1 Key Insights from EDA:**

**Demographic Trends:** Older customers and those with higher balances showed a higher likelihood of exiting the bank.

**Customer Behaviors:** Customers with fewer products and those who are less active members were more prone to churn.

**Geographical Influence:** Customers from Germany exhibited a higher churn rate compared to those from France and Spain.

**2.3.2 Model Evaluation:**

**Feature Selection:** Identifying and selecting significant features, such as Age, Balance, and Geography, was crucial for improving model accuracy.

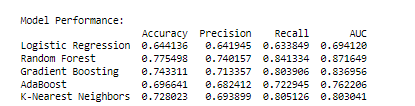
**Hyperparameter Tuning:** Fine-tuning hyperparameters for models like Random Forest and Gradient Boosting significantly enhanced predictive performance.

**Performance Metrics:** Random Forest and Gradient Boosting emerged as the top-performing models based on Accuracy, Precision, Recall, and AUC, indicating their effectiveness in predicting customer churn.

**3. Jupyter Notebook**

-Codes are and explanations are written in Jupyter Notebook.

**4. Conclusion**



From the comparison, we can observe how different models perform across various metrics.

**Accuracy:**

Accuracy gives an overall view of correct predictions. Random Forest achieves the highest Accuracy (77.5%), followed closely by Gradient Boosting (74.3%) and KNN (72.8%), indicating their ability to correctly classify instances.

**Precision:**

Precision indicates how reliable positive predictions are. Random Forest has the highest Precision (74.0%), meaning it identifies positives with the highest reliability. KNN and Gradient Boosting also perform well in Precision.

**Recall:**

Recall shows how well the model captures positive instances.Random Forest (84.1%) and KNN (80.5%) excel in Recall, showing their ability to capture true positives effectively. Gradient Boosting follows closely with 80.4%.

**AUC:**

AUC assesses the model's ability to distinguish between classes.Random Forest leads in AUC (0.872), indicating superior discrimination ability between positive and negative classes. Gradient Boosting (0.837) and KNN (0.803) also demonstrate good discrimination.

**Summary**

**Random Forest emerges as the top performer across all metrics, demonstrating high Accuracy, Precision, Recall, and AUC.** It is particularly strong in correctly classifying instances and distinguishing between classes.

Gradient Boosting and KNN also perform well, showing competitive metrics across the board, making them viable alternatives depending on specific needs such as interpretability (Gradient Boosting) or instance-based learning (KNN).

Logistic Regression and AdaBoost show moderate performance compared to the ensemble methods, with AdaBoost slightly outperforming Logistic Regression.

**Conclusion:**

The analysis highlighted the critical factors influencing customer churn and demonstrated the effectiveness of advanced machine learning techniques in predicting churn. The insights gained can aid the bank's management and marketing teams in devising targeted strategies to improve customer retention and satisfaction.