Bank Customer Churn Analysis Report

1. Introduction

In today's competitive banking industry, retaining customers is crucial for business success. Customer churn, defined as the rate at which customers stop doing business with a company, can significantly impact a bank's revenue. Understanding why customers churn can help banks develop strategies to improve customer retention.

This project analyzes a bank's customer dataset to identify factors contributing to customer churn. The dataset is prepared for machine learning to predict which customers are at risk of leaving the bank.

---

2. Dataset Overview

The dataset contains information on customers' demographics, account details, and banking behavior. Below is a summary of the dataset:

Feature	Description
CustomerID	Unique customer identifier
Surname	Customer's last name
CreditScore	Credit score of the customer
Geography	Customer's country
Gender	Customer's gender
Age	Customer's age
Tenure	Number of years with the bank
Balance	Customer’s account balance
NumOfProducts	Number of products held by customer
HasCrCard	Whether the customer has a credit card
IsActiveMember	Whether the customer is an active member

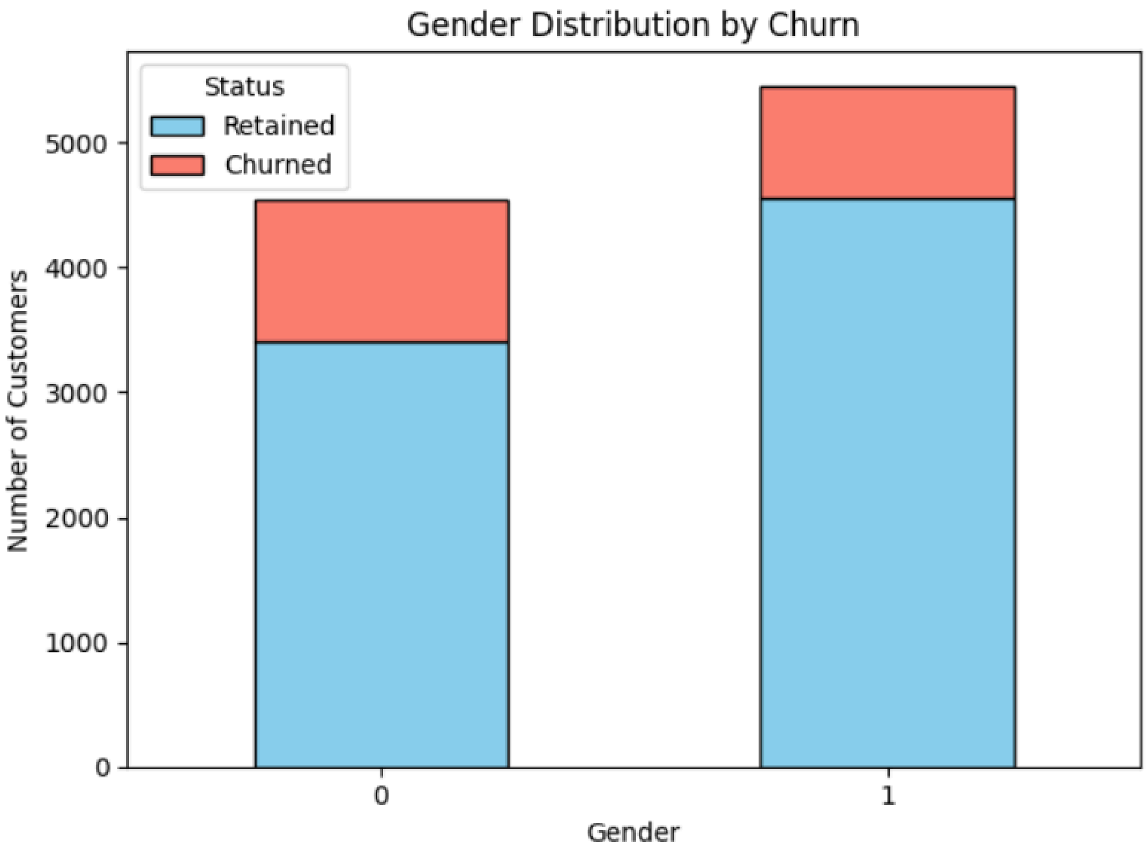
Estimated Salary	Estimated annual salary of customer
Exited	Whether the customer has churned (Target Variable)

The dataset contains 10,000 rows and 13 columns.

3. Exploratory Data Analysis (EDA)

Gender Distribution by Churn

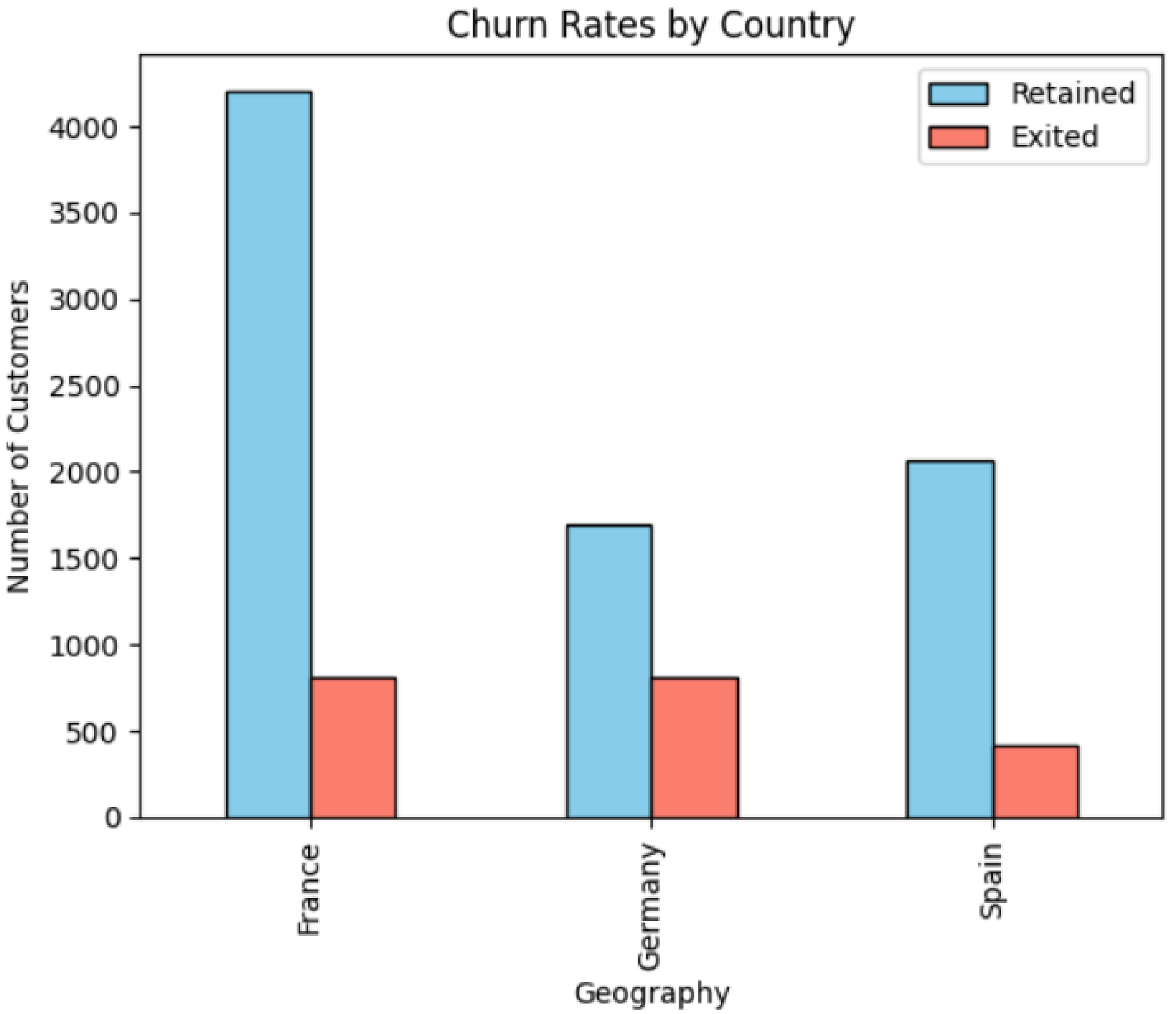
The chart below shows the distribution of churned and retained customers by gender:



- Insight: Both males and females have similar churn rates, indicating that gender is not a strong factor in predicting churn. However, visualizing this through a bar chart adds clarity and makes the insight more engaging.

Geography and Churn

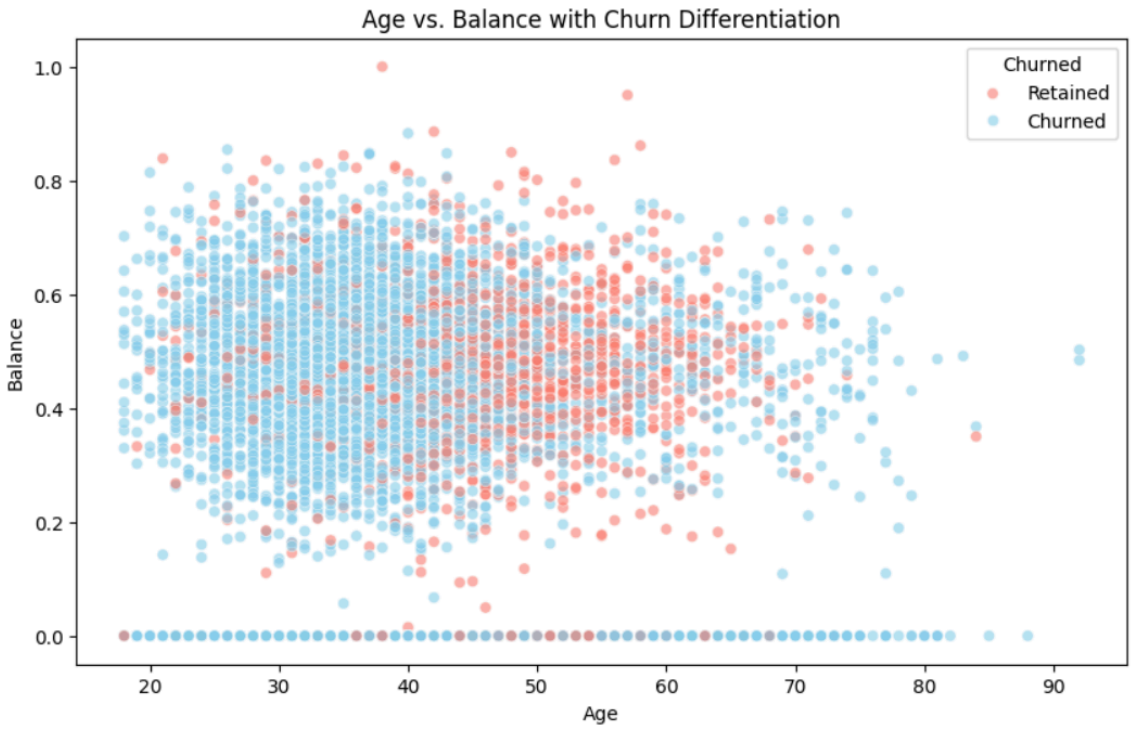
The plot below displays customer churn rates by country:



- Insight: Customers from France have a lower churn rate compared to customers from Germany and Spain.

Age and Balance Interaction

The scatterplot below shows the interaction between age and balance for churned and retained customers:



- Insight: The plot reveals that younger customers with higher balances have a higher likelihood of churning. Additionally, there is a noticeable spread in balances among older retained customers, suggesting that age and balance interact differently across customer segments.

---

4. Feature Engineering

Several new features were created to improve the dataset for machine learning models:

4.1 Balance-to-Salary Ratio

- **Formula:** balance / estimated_salary
- **Purpose:** This feature captures the relative wealth of a customer by comparing their account balance to their estimated salary. A high balance-to-salary ratio may indicate that the customer has a significant portion of their assets with the bank.
- **Insight:** Older customers with higher balances tend to be more stable, while younger customers with similar balances are more likely to churn. This feature can help banks develop targeted retention strategies for different age groups.

4.2 Age-Balance Interaction

- **Formula:** age * balance
- **Purpose:** This feature captures the interaction between a customer's age and their account balance. It reflects the financial profile of customers across different age groups.
- **Insight:** Older customers with higher balances tend to be more stable, while younger customers with similar balances are more likely to churn. This feature can help banks develop targeted retention strategies for different age groups.

4.3 IsHighBalance

- **Formula:** balance > df['balance'].median()
- **Purpose:** This binary feature identifies customers with balances above the median balance in the dataset. It helps in segmenting customers based on their financial standing.
- **Insight:** High-balance customers are often more valuable to the bank but may also have higher churn risk if they feel underserved. By identifying these customers, banks can prioritize their engagement efforts to reduce churn.

---

5. Machine Learning Preparation

The dataset is now ready for machine learning.

Steps Taken:

1. **Splitting the Data:**
   - The dataset was split into a training set (80%) and a test set (20%).
2. **Model Selection:**
   - Logistic Regression was used for initial predictions.
3. **Evaluation Metrics:**
   - Accuracy, Precision, Recall, and Confusion Matrix were used to evaluate the model.

---

6. Tools Used

To complete this project, the following tools were used:

Tool/Program	Purpose

Python	Main programming language
Jupyter Notebook	Writing and running Python code
CreditScore	Credit score of the customer
Pandas	Data manipulation and analysis
Seaborn & Matplotlib	Data visualization
Scikit-learn	Machine learning library

7. Next Steps

The next step is to improve the machine learning model by:

- Trying other models (e.g., Random Forest, Gradient Boosting)
- Performing hyperparameter tuning
- Using cross-validation to improve model performance

8. Conclusion

The Bank Customer Churn Analysis project provided valuable insights into customer behavior. By identifying key factors that contribute to churn, banks can take proactive measures to retain customers and improve overall business performance. The dataset is now prepared for machine learning, enabling the prediction of customers at risk of churning.

Appendix: Code Snippets

Below are some of the key Python code snippets used in this project:

Label Encoding:

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
df['gender'] = label_encoder.fit_transform(df['gender'])
```

- **Purpose:** Converts categorical variables (e.g., "Gender") into numerical values for machine learning models. Here, "Male" and "Female" are encoded as 0 and 1, respectively.

Creating New Features:

```
# Balance-to-Salary Ratio
df['balance_to_salary_ratio'] = df['balance'] / df['estimated_salary']

# Age-Balance Interaction
df['age_balance_interaction'] = df['age'] * df['balance']

# IsHighBalance
df['is_high_balance'] = df['balance'] > df['balance'].median()
```

- **Purpose:** The Balance-to-Salary Ratio helps identify customers whose wealth is concentrated in their bank accounts. Used to analyze churn based on financial patterns.

Bank Customer Churn Analysis Report

1. Introduction

In today's competitive banking industry, retaining customers is crucial for business success. Customer churn, defined as the rate at which customers stop doing business with a company, can significantly impact a bank's revenue. Understanding why customers churn can help banks develop strategies to improve customer retention.

This project analyzes a bank's customer dataset to identify factors contributing to customer churn. The dataset is prepared for machine learning to predict which customers are at risk of leaving the bank.

---

2. Dataset Overview

The dataset contains information on customers' demographics, account details, and banking behavior. Below is a summary of the dataset:

Feature	Description
CustomerID	Unique customer identifier
Surname	Customer's last name
CreditScore	Credit score of the customer
Geography	Customer's country
Gender	Customer's gender
Age	Customer's age
Tenure	Number of years with the bank
Balance	Customer's account balance
NumOfProducts	Number of products held by customer
HasCrCard	Whether the customer has a credit card

IsActiveMember	Whether the customer is an active member
EstimatedSalary	Estimated annual salary of customer
Exited	Whether the customer has churned (Target Variable)

The dataset contains **10,000 rows** and **13 columns**.

---

3. Exploratory Data Analysis (EDA)

Gender Distribution by Churn

The chart below shows the distribution of churned and retained customers by gender:

```
# Bar chart: Gender distribution by churn
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8, 6))
gender_churn = df.groupby(['Gender', 'Exited']).size().unstack()
gender_churn.plot(kind='bar', stacked=True, color=['skyblue', 'salmon'], edgecolor='black')
plt.title('Gender Distribution by Churn')
plt.xlabel('Gender')
plt.ylabel('Number of Customers')
plt.legend(['Retained', 'Churned'], title='Status')
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```

- **Insight:** Both males and females have similar churn rates, indicating that gender is not a strong factor in predicting churn. However, visualizing this through a bar chart adds clarity and makes the insight more engaging.

Geography and Churn

The grouped bar chart below shows customer churn rates by country:

```
# Grouped bar chart: Churn rates by country
```

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
geography_churn = df.groupby(['Geography', 'Exited']).size().unstack()
geography_churn.plot(kind='bar', stacked=False, color=['skyblue', 'salmon'],
edgecolor='black')
plt.title('Churn Rates by Country')
plt.xlabel('Country')
plt.ylabel('Number of Customers')
plt.legend(['Retained', 'Churned'], title='Status')
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```

- **Insight:** Customers from **France** have a lower churn rate compared to customers from **Germany** and **Spain**. The grouped bar chart provides a clear comparison of churn rates across different countries.

Age and Balance Interaction

The scatterplot below displays the interaction between customers' age and account balance, with color differentiation to indicate whether the customer has churned or not:

```
# Scatterplot: Age vs. Balance with churn differentiation
plt.figure(figsize=(10, 6))
sns.scatterplot(
    data=df,
    x='age',
    y='balance',
    hue='exited',
    palette={0: 'blue', 1: 'red'},
    alpha=0.6
)
plt.title('Age vs. Balance with Churn Differentiation')
plt.xlabel('Age')
plt.ylabel('Balance')
plt.legend(title='Churned', labels=['Retained', 'Churned'])
plt.show()
```

- **Insight:** The plot reveals that younger customers with higher balances have a higher likelihood of churning. Additionally, there is a noticeable spread in balances among older retained customers, suggesting that age and balance interact differently across customer segments.

---

4. Feature Engineering

To improve the dataset for machine learning models, several new features were created. These features provide deeper insights into customer behavior and enhance the predictive power of models.

4.1 Balance-to-Salary Ratio

- **Formula:** `balance / estimated_salary`
- **Purpose:** This feature captures the relative wealth of a customer by comparing their account balance to their estimated salary. A high balance-to-salary ratio may indicate that the customer has a significant portion of their assets with the bank.

```
# Boxplot: Balance-to-Salary Ratio by Churn Status
plt.figure(figsize=(10, 6))
sns.boxplot(
   data=df,
   x='Exited',
   y='balance_to_salary_ratio',
   hue='Exited',
   palette=['skyblue', 'salmon'],
   dodge=False
)
plt.title('Balance-to-Salary Ratio by Churn Status')
plt.xlabel('Churn Status (0 = Retained, 1 = Churned)')
plt.ylabel('Balance-to-Salary Ratio')
plt.tight_layout()
plt.show()
```

- **Insight:** Customers with a high balance-to-salary ratio are more likely to churn. This suggests that these customers might be managing significant portions of their wealth elsewhere or are unsatisfied with the bank's services.

4.2 Age-Balance Interaction

- **Formula:** `age * balance`
- **Purpose:** This feature captures the interaction between a customer's age and their account balance. It reflects the financial profile of customers across different age groups.

```
# Boxplot: Age-Balance Interaction by Churn Status
plt.figure(figsize=(10, 6))
sns.boxplot(
    data=df,
    x='Exited',
    y='age_balance_interaction',
    hue='Exited',
    palette=['skyblue', 'salmon'],
    dodge=False
)
plt.title(‘Age-Balance Interaction by Churn Status')
plt.xlabel(‘Churn Status (0 = Retained, 1 = Churned)')
plt.ylabel(‘Age-Balance Interaction')
plt.tight_layout()
plt.show()
```

- **Insight:** Older customers with higher balances tend to be more stable, while younger customers with similar balances are more likely to churn. This feature can help banks develop targeted retention strategies for different age groups.

4.3 IsHighBalance

- **Formula:** `balance > df['balance'].median()`
- **Purpose:** This binary feature identifies customers with balances above the median balance in the dataset. It helps in segmenting customers based on their financial standing.

```
# Creating the IsHighBalance feature
df['is_high_balance'] = df['balance'] > df['balance'].median()
```

- **Insight:** High-balance customers are often more valuable to the bank but may also have higher churn risk if they feel underserved. By identifying these customers, banks can prioritize their engagement efforts to reduce churn.

5. Machine Learning Preparation

The dataset is now ready for machine learning.

Steps Taken:

1. **Splitting the Data:**
   - The dataset was split into a training set (80%) and a test set (20%).
2. **Model Selection:**
   - Logistic Regression was used for initial predictions.
3. **Evaluation Metrics:**
   - Accuracy, Precision, Recall, and Confusion Matrix were used to evaluate the model.

---

6. Tools Used

To complete this project, the following tools were used:

Tool/Program	Purpose
Python	Main programming language
Jupyter Notebook	Writing and running Python code
Pandas	Data manipulation and analysis
Seaborn & Matplotlib	Data visualization
Scikit-learn	Machine learning library

---

7. Next Steps

The next step is to improve the machine learning model by:

- Trying other models (e.g., Random Forest, Gradient Boosting)
- Performing hyperparameter tuning
- Using cross-validation to improve model performance

---

8. Conclusion

The Bank Customer Churn Analysis project provided valuable insights into customer behavior. By identifying key factors that contribute to churn, banks can take proactive measures to retain customers and improve overall business performance. The dataset is now prepared for machine learning, enabling the prediction of customers at risk of churning.

Appendix: Code Snippets

Below are some of the key Python code snippets used in this project, along with explanations of their relevance to the dataset:

Label Encoding:

```
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
df['gender'] = label_encoder.fit_transform(df['gender'])
```

- **Purpose:** Converts categorical variables (e.g., "Gender") into numerical values for machine learning models. Here, "Male" and "Female" are encoded as 0 and 1, respectively.

Creating New Features:

Balance-to-Salary Ratio:

```
# Balance-to-Salary Ratio
df['balance_to_salary_ratio'] = df['balance'] / df['estimated_salary']
```

- **Purpose:** This feature helps identify customers whose wealth is concentrated in their bank accounts. Used to analyze churn based on financial patterns.

Age-Balance Interaction:

```
# Age-Balance Interaction
df['age_balance
```