**Name : Semir Yusuf ID : ugr/32136/14**

**Biruktawit sisay ugr/34533/14**

**Bank Customer Churn Prediction: Case Study Problem Statement**

## Introduction

Customer churn is a critical issue in the banking sector, as retaining existing customers is often more cost-effective than acquiring new ones. This project aims to predict whether a customer will churn based on various features, including demographic information, account details, and transactional behavior. By identifying the key drivers of churn, banks can take proactive steps to enhance customer satisfaction and loyalty.

## Project Overview

This project uses machine learning to analyze a dataset containing customer information and churn status. The primary goal is to build a predictive model that can accurately classify customers into 'churn' or 'non-churn' categories. The project follows a structured workflow involving data exploration, preprocessing, model building, and evaluation.

## Key Steps in the Analysis

### 1. Data Loading and Exploration

The first step involves loading the dataset and exploring its structure and contents. Key tasks include identifying the types of features (e.g., categorical, numerical), checking for missing values, and generating basic statistical summaries. Exploratory Data Analysis (EDA) involves creating visualizations such as histograms, box plots, and correlation heatmaps to uncover patterns and relationships in the data.

### 2. Data Preprocessing

Data preprocessing ensures the dataset is clean and suitable for machine learning. This includes:   
**-Handling Missing Values:** Filling or removing missing entries to avoid errors during modelling   
**-Encoding Categorical Variables:** Converting non-numerical data (e.g., 'Male', 'Female') into numerical formats using techniques like one-hot encoding or label encoding.   
-Feature Scaling: Standardizing numerical features to ensure all variables are on the same scale.   
**-Data Splitting:** Dividing the dataset into training and testing sets to evaluate model performance.

### 3. Model Building

Various machine learning algorithms are applied to predict customer churn. Some common algorithms include:   
**-Logistic Regression:** A statistical model for binary classification tasks.

## About the data

The set has 10,000 rows & 18 columns. The Columns are-

1. **Row-Number** — corresponds to the record (row) number and has no effect on the output.
2. **Customer-id**— contains random values and has no effect on customer leaving the bank.

3. **Surname** —the surname of a customer has no impact on their decision to leave the bank.

4. **Credit-Score** —can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.

5. **Geography** —a customer’s location can affect their decision to leave the bank.

1. **Gender**—it’s interesting to explore whether gender plays a role in a customer leaving the bank.
2. **Age** —this is certainly relevant, since older customers are less likely to leave their bank than younger ones.
3. **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.
4. **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.
5. **Num-Of-Products**—refers to the number of products that a customer has purchased through the bank.
6. **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.
7. Is-Active-Member —active customers are less likely to leave the bank.
8. Estimated-Salary —as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
9. Exited —whether or not the customer left the bank.
10. Complain —customer has complaint or not.
11. Satisfaction Score —Score provided by the customer for their complaint resolution.
12. Card Type —type of card held by the customer.
13. Points Earned —the points earned by the customer for using credit card.

### 4. Insights and Conclusions

The analysis identifies the most influential features driving customer churn. For instance, variables such as account balance, tenure, and customer complaints may significantly impact churn likelihood. These insights help banks target at-risk customers with personalized retention strategies, such as tailored promotions, improved services, or proactive customer support.

The correlation matrix has an interesting relationship- the complaint and Exited variables are 100% positively correlated.

Churn Rate- about 20% of customers are leaving the bank, out of the 10,000.

Churn vs. Features

Gender: 25.07% of female customers vis-a-vis 16.47% male customers are exiting.

Geography: Amongst the countries, Germany has the highest churn rate of 32.44%, followed by Spain (16.67%) and France (16.17%).

Age Group: The majority of customers leaving are in the 50-60 age range, with a churn rate of 56.21%, followed by the 40-50 age group with a corresponding rate of 33.96%.

Tenure : Even loyal customers with ten years of association have a churn rate of 20%. Customers below the one year mark have a churn rate of 23%.

Active Member: Non-active members have a higher churn rate of 27% while active customers of the bank have a churn of 14%.

Number of Products: Interestingly, customers with higher number of products are more likely to churn. Customers who have bought 3 products have a churn rate of 83% while those who bought only one product have a churn rate of 28%.

## Conclusion

The Bank Customer Churn Analysis project highlights the value of data-driven decision-making in the banking industry. By leveraging machine learning, banks can identify potential churners early and take strategic actions to retain them. This not only improves customer satisfaction but also contributes to the bank's long-term profitability. Future work could involve integrating additional data sources, such as social media activity or customer feedback, to enhance the model's predictive power.

**REFERENCES**

* Customer churn dataset [kaggle]