# Predicting Customer Churn in Banking: A Binary Classification Approach

Vaishnavi Paineni

## Background/Problem Statement:

In today's highly competitive market, almost every business is losing their customers to their adversaries leading to loss of immediate revenue and increased costs associated with acquiring new customers to replace them. When it comes to the banking industry, it is no different and it is essential to maintain customer loyalty and reduce customer churn. Customer churn also affects the bank's reputation as it reflects the bank's failure to retain and satisfy its customers. This in turn erodes the customer trust in the bank and leads to a negative impact on long-term profitability. This can be prevented by predicting the churn and employing countermeasures to retain the customers.

## About the Dataset:

The dataset chosen is https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn. The dataset is about 315kB with 10,000 records. The following are the types of data :

**Categorical Data:**

- Geography: Represents the geographical location of customers.
- Gender: Signifies the gender of customers.
- Card Type: Indicates the type of card held by customers.

**Ordinal Data:**

- Satisfaction Score: Represents customer feedback on a discrete scale with a clear order but uneven intervals.

**Numeric Data:**

- CreditScore: Represents the creditworthiness of customers (continuous numeric).
- Age: Represents the age of customers (continuous numeric).
- Tenure: Indicates the number of years a customer has been with the bank (continuous numeric).
- Balance: Represents the account balance of customers (continuous numeric).
- NumOfProducts: Represents the count of products purchased by customers (discrete numeric).
- EstimatedSalary: Represents the estimated salary of customers (continuous numeric).
- Satisfaction Score (if recorded as a numeric value, it would be numeric data).

**Binary Data:**

- HasCrCard: Indicates whether a customer has a credit card.
- IsActiveMember: Represents the active membership status of customers.
- Complain: Indicates whether a customer has filed a complaint.
- Exited: The target variable, representing whether a customer left the bank.

**Text Data:** Surname: This could be considered text data.

## Methodology:

Our project follows a systematic methodology to process and analyze the dataset for customer churn prediction. The key steps in our approach are as follows:

**1. Data Preprocessing:**

- Handling Missing Values: We begin by addressing missing values, either by removing incomplete data points or filling in the gaps using appropriate techniques.

- Outlier Standardization: We identify and standardize outliers in the dataset to ensure the integrity of our analysis.

**2. Exploratory Data Analysis (EDA):**

- Data Overview: We explore the dataset using shape, dtypes, and value_counts to gain insights into its structure and distribution.

- Data Visualization: Utilizing data visualization techniques, we plot relationships between select features, such as "tenure" vs. the number of customers who churned. These visualizations help us understand the data's patterns and trends.

**3. Data Splitting:**

- Train-Test Split: We employ the train_test_split method to divide the dataset into an 80% training set and a 20% testing set. This separation ensures an unbiased assessment of model performance.

**4. Feature Engineering:**

- Feature Engineering: We engage in feature engineering to enhance the dataset's information content, potentially creating new variables or transforming existing ones for improved predictive performance.

**5. Supervised Learning Models:**

- Model Selection: We apply various supervised learning models to the preprocessed data, including: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, Artificial Neural Network (ANN)

**6. Model Evaluation:**

- Performance Metrics: We evaluate the models using key performance metrics, including accuracy, precision, recall, F1 score, and ROC-AUC curve analysis. This assessment helps determine which model best fits our dataset.

- Confusion Matrix: We visualize the model's performance using a confusion matrix, represented as a heatmap, to provide an intuitive view of true positives, true negatives, false positives, and false negatives.

**7. Result Analysis:**

- Model Comparison: We compare the actual values with the predicted values from each model to select the best-performing one based on the results.

By following this methodology, we aim to develop a comprehensive understanding of customer churn in the bank and select the most effective predictive model to help the bank improve customer retention. The results and insights derived from this process will be summarized and presented in our final project report.