**Project Title: Bank Churn Prediction**

**Dataset Link:** [**https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers**](https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers)

**Problem Statement:** Bank customer churn prediction is a critical task in the banking industry as it helps the banks to identify the customers who are likely to switch to other banks. It is important for banks to predict customer churn because losing customers can have a significant impact on their revenue and market share. The problem statement for this task is to develop a machine learning model that can accurately predict customer churn based on their historical transaction data, demographic information, and other relevant factors. The model should be able to identify the customers who are at the highest risk of churn so that the bank can take proactive measures to retain them. This involves collecting and analyzing data from various sources, selecting the relevant features, and training the model using appropriate algorithms. We will be using Machine Learning and some data analysis techniques to analyze the bank data which is open-sourced and can be found on Kaggle. We will be looking into the various factors that lead to the churn of a customer from a bank. We will be using python and Pyspark frameworks to complete this project.

**Aim:** The ultimate goal of this project is to build a reliable and accurate model that can help banks to reduce customer churn and improve customer satisfaction.

**Research Questions and Milestones:**

1. Analyze and obtain the factors which are responsible for a customer to churn
2. Is the churn of customers related to the gender of a customer?
3. Does geography or location of customer is also a factor for churn?
4. Does a person having a credit card churn?
5. Build a predictive model that tells us about the chances of a customer to churn so that some extra benefits can be provided by the bank

## **Milestone 1:** Data loading and Exploratory Data Analysis - Reading data and gaining insights from data.

## **Milestone 2:** Data Preprocessing: Removing duplicates and outliers, Null value imputation.

## **Milestone 3:** Feature Engineering - Scaling and log transformation of columns

## **Milestone 4**: Model Evaluation - Using Confusion matrix, F1-score and accuracy to evaluate the model. **Technology used:** Pyspark, Google Colab, Jupyter Notebook, Python and Machine Learning Frameworks.