Bank Customer Churn Analysis & Prediction: (EDA and Data Cleaning) " Report 1 "

- **Prepared By:** Hassan Waked –Team Leader, DEPI Graduation Project.
- **Date:** 20th Dec 2024.

1. Problem Definition:

- **The Business Goal Is To** identify customers who are likely to leave the bank (churn), so the bank can take preventive measures. We aim to build a machine learning model that can predict churn based on customer features.

---

2. Data Collection

- **The Dataset Is Based on** real bank customer information from Kaggle.
- **Original Dataset Shape:**
  - Rows: 10,002 records.
  - Columns: 14 columns.
- **The Data Includes Various** types of features: personal information, financial indicators, and behavior flags.

---

3. Exploratory Data Analysis (EDA)

General Structure:

- **Numeric Features (11):**
  - Rownumber, Customerid, Creditscore, Age, Tenure, Balance, Numofproducts, Hascrcard, Isactivemember, Estimatedsalary, Exited.
- **Categorical Features (3):**
  - Surname, Geography, Gender.

Uniqueness:

- **Surname: 2,932** unique values
- **Customerid: 10,000** unique values
- **Estimatedsalary: 9,999** unique values
- **Balance: 6,382** unique values

- **Creditscore: 460** unique values
- **Age: 73** unique values
- **Tenure: 11** unique values
- **Categorical variables like:** Geography, Gender, Hascrcard, Isactivemember, and Exited had **2–3** unique values.

Skewness (distribution shape):

- **Exited: 1.47** → skewed (imbalance in classes).
- **Age: 1.01** → right skewed.
- **Numofproducts: 0.75.**
- **Other Features**: have near-zero skewness, meaning normal distribution.

Outliers (IQR method):

- **Creditscore: 15** outliers.
- **Age: 359** outliers.
- **Numofproducts: 60** outliers.
- **Exited: 2,038** rows marked as outliers (due to class imbalance).
- **Other Columns:** had **0 outliers.**

---

4. Data Cleaning:

Actions Taken:

- **Dropped 3 Unnecessary Columns:**
  - **Customerid:** Unique identifier, not useful for modeling
  - **Rownumber:** Sequential row number
  - **Surname:** Too many unique values, low predictive power
- **Removed 2 Duplicate Rows**
- **Handled Missing Values:**
  - 4 Rows had missing values in these columns:
    - Geography, Age, Hascrcard, Isactivemember.
  - These rows were dropped to maintain data quality.
- **Changed Data Types:**
  - Hascrcard And Isactivemember Columns Converted from float type to integer type for modeling.

---

5. Data Preparation:

After cleaning:

- **Final Number Of Rows:** 9,996
- **Final Number Of Columns:** 11
- **Total Rows Removed:** 6 Rows.
    - 4 Rows due to missing values
    - 2 Duplicate Rows.
- **Total Columns Removed:** 3 Columns.

---

6. Next Steps (Features Engineering):

- **Encode Categorical Columns:** (Gender, Geography) for use in ML models.
- **Scale Numeric Features Like:** Balance, Estimatedsalary, Age, etc.
- **Handle Class Imbalance In:** The Exited Column using techniques like **SMOTE** or class .weighting.
- **Perform Features** selection and correlation analysis.
- **Train Classification Models** and evaluate them using recall, precision, and ROC-AUC.

Conclusion:

- The Dataset Has Been thoroughly cleaned and analyzed. With 9,996 reliable records and relevant features, it is now ready for training machine learning models that will help predict customer churn with better accuracy. The business can use these insights to improve retention strategies.