**ANALYSIS OF BANK CHURNERS – A DEEP DIVE UTILISING SUPERVISED LEARNING ALGORITHMS WITH THE CRISP-DM METHODOLOGY**

By: Onyiriuba Leonard & Onwumbiko Chidera

TechBros :**TEAM NAME**

Applied Techniques of Data Mining and Machine Learning :**COURSEWORK**

[Kaggle](https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) :**DATASET**

**TABLE OF CONTENT**

1 - **Business Background Understanding**

* 1. – Introduction
  2. – What is Churn?
  3. – Why Churn Analysis?
  4. – Understanding the CRISP-DM Methodology
  5. – Project Context and Objective
  6. – Literature Review: Data Mining in the Application Domain
  7. – Project Plan and Timeline

2 - **Data Understanding and Exploration**

2.1 – Data Set Overview  
2.2 – Data Characteristics and Quality Evaluation  
2.3 – Initial Insights and Patterns (Python & Pandas Profiling)  
2.4 – Data Analysis (Power BI)

3 - **Data Preparation and Pre-Processing**

3.1 – Data Cleaning and Transformation  
3.2 – Handling Missing Values and Outliers  
3.3 – Feature Engineering  
3.4 – Exploratory Data Analysis (EDA) after Pre-Processing  
3.5 – Data Scaling and Normalization

4 - **Modelling and Model Evaluation**

4.1 – Model Selection and Building  
4.2 – Training and Testing  
4.3 – Evaluation Metrics  
4.4 – Interpretation and Analysis of Results  
4.5 – Iterative Refinement

5 - **Project Evaluation and Summary**

5.1 – Summary of Findings and Insights  
5.2 – Recommendations and Future Work  
5.3 – Impact on Banks, Business, and Society  
5.4 – Lessons Learned and Reflections

6 - **References**

7 - **Appendix**

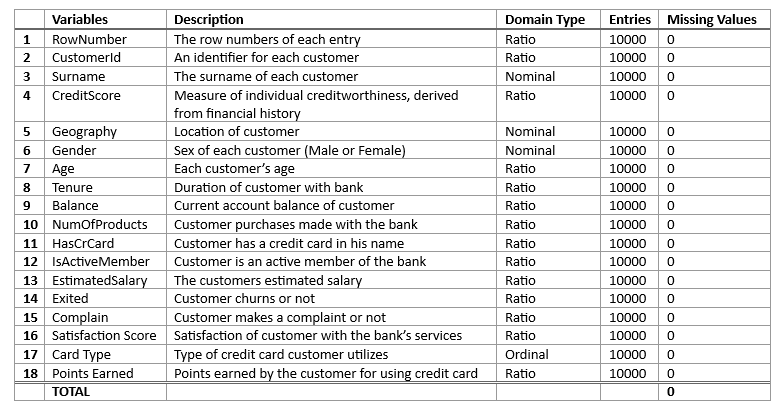
1 - **Business Background Understanding**

2 - **Data Understanding and Exploration**

2.1 – Data Set Overview

The dataset is gotten from [Kaggle](https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) which makes available, open-source datasets, and allows for collaborations among data enthusiasts. Our dataset comprises 10,000 rows and 18 columns. The domain types of each column are specified below:

**Table 1**: Variables, description, domain type, number of valid entries, and missing values in the dataset.

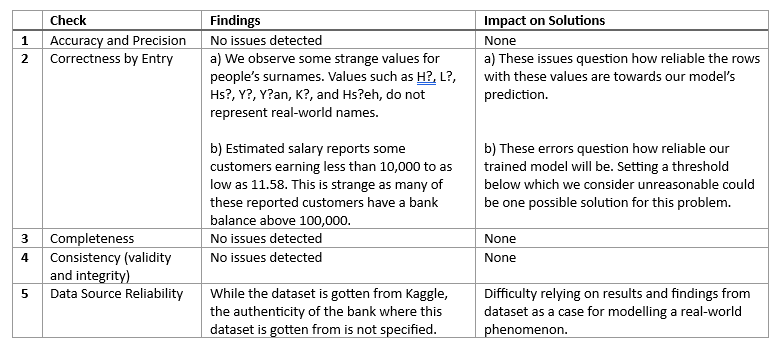


From table 1, we observe that the dataset gotten from [Kaggle](https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) without any data cleaning or pre-processing conducted has no missing values.

2.2 – Data Characteristics and Quality Evaluation

The quality of our dataset will be judged by the following - Accuracy and Precision, Correctness by Entry, Completeness, Consistency (validity and integrity), and Data Source Reliability Checks.

**Table 2**: Summary on data quality checks



Many of the issues highlighted in table 2 are further discussed in later sections during the iterative model building phase.

2.3 – Initial Insights and Patterns (Python & Pandas Profiling)

Exploratory data analysis (EDA) helps us derive insights about our dataset, understand the structure of the variables, get a glance at the level of data cleaning and preparation we need to undertake before modelling. For our analysis, we make use of Pandas Profiling – A library for EDA and some python script for our initial EDA. The EDA process was undertaking with the data quality checks as well as both help draw insights from the data.

As part of our EDA, we considered the following:

* The distribution of categories in each variable (Nominal, Ordinal, and Ratio).
* The number of duplicate entries in the rows.
* The number of missing values.
* Descriptive statistics of variables.
* Correlation among variables – Point-Biserial Correlation.

**Table 3**: Distribution of class categories among the variables – Gender, Geography, Card Type, Tenure, Satisfaction Score, NumOfProducts, IsActiveMember, HasCrCard, Exited, and Complain.

Here, we want to look at the distribution of the Gender, Geography, and Exited variables. All other variables mentioned above are provided in [Appendix A](#Appendix_A).

The colour scheme for highest and lowest values in each table are defined below:  
- Highest  
- Lowest

Figure 3.1: Geography



Figure 3.2: Gender



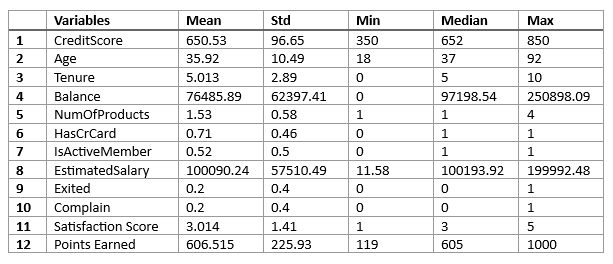
Figure 3.3: Exited



Our target variable in our dataset is Exited. Looking at the distribution of the exited category in Figure 3.3, we see 7962 instances of people who do not leave the bank and 2038 people who leave the bank. This distribution among these classes is imbalanced and create a problem for our model’s ability to capture and learn from the minority class while training. Solving this problem employs using some synthetic sampling technique to create instances for the minority class.

The difference of about a thousand customers between the gender classes is negligible as both classes are well represented in our dataset. However, the location of our customers has an uneven distribution among the classes with half of the bank customers being French. This class imbalance raises the problem of the likelihood of the model being better suited a specific demographic of the bank customers better than the rest. In our analysis of our final model, much attention needs to be drawn to this problem.

**Table 4:** Descriptive statistics of variables – CustomerId, RowNumber, Geography, Gender, Card Type, and Surname ignored as either categorical data or irrelevant for our analysis.



As seen in table 1, the dataset has no missing value. From our dataset, we also have no duplicate entries recorded in the data as each row represent a new customer record for the bank.

A comprehensive exploratory data analysis is conducted using Pandas Profiling. A link to the full report and analysis has been provided. ([Pandas Profiling Report – EDA](https://leonardleo.github.io/Analysing-Key-Performance-Indicators-Leading-to-Customer-Churn-in-Banks/pandas_profiling_report/initial_EDA.html))

2.4 – Data Analysis (Power BI)

3 - **Data Preparation and Pre-Processing**

4 - **Modelling and Model Evaluation**

5 - **Project Evaluation and Summary**

6 – **References**

Kollipara, R. (n.d.). *Bank Customer Churn*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn>

**Appendix A**

*DISTRIBUTION OF CLASS CATEGORIES AMONG VARIABLES*

Here, we want to look at the distribution among the variables – Card Type, Tenure, Satisfaction Score, NumOfProducts, IsActiveMember, HasCrCard, and Complain.

The colour scheme for highest and lowest values in each table are defined below:  
- Highest  
- Lowest

Card Type



Tenure



Satisfaction Score



NumOfProducts



IsActiveMember



HasCrCard



Complain

