TASK 6.1 SOURCING OPEN DATA

SOURCE OF DATA

I chose a dataset from the open source Kaggle website on May 27, 2023 titled Bank Customer Churn. Here is the **DATA SET LINK**.

DATA COLLECTION

This data was collected from government sources by the author of the data set named Dhoni in collaboration with the owner Radheshyam Kollipara. There does not exist any more information about the data collection methods for this data set. The data collection occurred during the time period of March 31 2022 until April 30 2022.

DATA CONTENTS

The data set contains information from an anonymous Multinational bank with the purpose of trying to determine which customers are most likely to stay/leave the bank. The data set contains but is not limited to the following information: surname, credit score, gender, age, tenure, and tenure.

DATA PROFILE

The dataset has 10,000 columns and 17 rows. The cleaned data still has 10,000 rows but only 16 columns since the Surname column which has PII information was removed. There were no missing values, duplicates, or mixed variables found. The variable type for the columns 'HasCrCard', 'IsActiveMember', 'Exited', and 'Complain' were changed from int64 to object and the 1s were replaced with 'yes' and the 0s were replaced with 'no'. This was done to facilitate understanding and to allow for clearer visualizations. Please refer to table below for a detailed description of each column variable as well as any column name changes that were performed for consistency purposes.

Original Column Name	New Column Name	Description	Time Invariant/Variant	Data Type
CustomerId	Customer_id	Unique customer numerical identifier	Invariant	Qualitative
CreditScore	Credit_score	Customers credit score	Invariant	Quantitative
Geography	N/A	Customers location	Invariant	Qualitative
Gender	N/A	Customers gender	Invariant	Qualitative
Age	N/A	Customers Age	Invariant	Quantitative
Tenure	N/A	The number of years the customer has been a client of the the bank	Invariant	Quantitative
Balance	N/A	Account balance	Invariant	Quantitative
NumOfProducts	Number_or_products	The number of products purchased by the customer	Invariant	Quantitative

HasCrCard	Has_credit_card	Does the customer have a credit card?	Invariant	Qualitative
IsActiveMember	Is_active_member	Is the customer active?	Invariant	Qualitative
EstimatedSalary	Estimated_salary	The estimated salary of customer	Invariant	Quantitative
Exited	N/A	Has the customer left the bank?	Invariant	Qualitative
Complain	N/A	Does the customer have any complaints?	Invariant	Qualitative
Satisfaction Score	Satisfaction_score	Score represents customers satisfaction of complaint resolution	Invariant	Quantitative
Card Type	Card_type	Indicates the type of card held by the customer	Invariant	Qualitative
Point Earned	Points_earned	The amount of points the customer has earned by using his/her credit card	Invariant	Quantitative

DATA LIMITATIONS AND ETHICS

LIMITATIONS:

Since the data collection methods for this data set cannot be verified or determined, we do not know how reliable the information is. This indicates that any insights from this data can only be used for informative purposes but cannot be considered facts. In addition, there are several unrealistic values in the ‘Estimated_salary’ columns which indicate there is something wrong with the way the data was collected for this column. For example, there is no way someone’s salary is $11.58 which is the lowest value in this column. It is difficult to determine what is a realistic salary since we do not know if the customers are working part-time, full-time, or are unemployed. At this stage, I have decided to leave the ‘Estimated_salary column as is. The firs possible solutions to this problem is perhaps to find the average hourly min wage for the three countries online and assume a minimum 15 hour work week to determine a threshold for annual salary; if this is under 5%, we can remove those rows. The second solution is to create a new variable that divides Estimated_salary into Low-income, Middle-income, and High-income based on pre-set values; this way we can still derive useful insights from this information without eliminating any rows as in the first suggestion.

ETHICS:

The column ‘Surname’ which had the customers last name was removed to ensure privacy and respect the customers right to anonymity.

QUESTIONS TO CONSIDER:

Is low credit score a factor for people leaving the bank?

Does geography have anything to do with customer churn?

Does age group and/or gender have anything to do with customer churn?

Are customers with low tenure more likely to leave the bank?

Do customers with a low account balance have a greater likelihood of leaving the bank?

Are customers with credit cards more likely to leave the bank?

Are inactive members more likely to leave the bank?

Are high-income individuals more likely to stay with the bank?

Are customers who complain more likely to be unsatisfied and leave the bank?

Does card type the customer has affect customer churn?

Does the more points a customer has, affect the likelihood of them staying?