5.4: Intro to Data Mining

1. To understand the data, you'll first need to assess the quality of the data, by checking for missing values, errors, and inconsistencies.
    - You'll also need to clean your data, using the techniques that you learned in previous Achievements. Fix any inconsistencies in the table and/or any errors, as far as it is possible.
    - Document your processes for assessing the data quality and cleaning the data, and note down any missing values or errors.

Answer:

Column name	Data Quality issue	Action taken	Notes
Row Number	Unnecessary column	Column dropped	
Customer_id	None	None	
Last Name	Privacy issue	Column dropped	Due to PII and not useful for Analysis
Credit score	3 Missing entries	None	
Country	Rename entries	F with France, DE with Germany, ES with Spain	Make country name consistent
Gender	Gender format is not consistent One NULL Value	Replace F with Female M with Male	Changed acronyms to spelled-out version of gender names
Age	One NULL value		
Tenure	None	None	
Balance	350 entries with $0,00	None	
NumOfProducts	None	None	
HasCrCard?	None	None	
IsActiveMember	None	None	

Column name	Data Quality issue	Action taken	Notes
Estimated Salary	One NULL value and One Blank value	None	
ExitedFromBank?	None	None	

No Duplicate records were found.

3. Now that you've cleaned the data, you're ready to calculate some basic descriptive statistics to understand the data. Remember, your goal is to identify the risk factors that have contributed to customers leaving the bank.

- Separate the clients into 2 groups: one for those who have left the bank and a second for those who have stayed (hint: "1" in the "ExitedFromBank" column represents customers who have left).

- Use pivot tables and other Excel functions to identify the top 3 to 4 factors that lead to clients leaving.

  Answer: See Excel file

  Some of the leading factors that contribute to client loss are Active Member, Age, Gender.

- Gather and analyze statistical information on both groups (e.g., find averages, means).

  See Excel Sheet.

- Determine the leading factors that contribute to client loss, based on your analysis of the data provided.

  See Excel Sheet- EDA

- Document your results and how you reached them.

  First I cleaned the data and then created multiple tables to see the descriptive statistics step by step.
  I noticed that being an active member played a big role in whether a client left or stayed with the bank, so I used that as first risk factor. Next, I explored other potential factors, and noticed that credit card ownership, gender, and age also played a huge role in whether a client left or stayed. I rearranged the risk factors around to see which had more of an impact and concluded the order to be active membership, credit card ownership, gender, and age.

4. Decision Tree