5.4. Intro to Data Mining
By Mary Kane

ANSWERS 5.4

1. Understanding the business goal: To increase customer retention, the sales team wants to identify the leading indicators that a customer will leave the bank. You've created a table of client attributes that you believe could indicate whether customers will leave—for example, age, estimated salary, etc. You're going to use this information to identify the top risk factors that contribute to client loss and model them in a decision tree.

2. Understand the data:

a. Data Dictionary Descriptions

Row_Number
Unique sequential number used to identify and locate each customer record in the dataset.

Customer_ID
Unique numerical identifier assigned to each customer.

Last_Name
Customer's last name. This column contains **personally identifiable information (PII)**.

Credit Score
Numerical value representing the customer's creditworthiness. Higher values indicate a stronger credit history and greater ability to repay debt.

Country
The country where the customer's bank account was created.

Gender
Customer's gender identifier.

Age
Customer's age in years.

Tenure
Length of time (in years) the customer has held an account with the bank.

Balance
Current monetary balance held in the customer's bank account.

NumOfProducts
Number of bank products currently associated with the customer.

HasCrCard?
Boolean indicator showing whether the customer has an approved credit card with Pig E. Bank. (1 = Yes, 0 = No).

5.4. Intro to Data Mining
By Mary Kane

IsActiveMember

Boolean indicator showing whether the customer has been financially active during the most recent period. (1 = Active, 0 = Inactive).

Estimated Salary

The estimated annual salary of the customer, typically derived from income or deposit patterns.

ExitedFromBank?

Boolean indicator showing whether the customer has left the bank
(1 = Customer exited, 0 = Customer retained).

5.4. Intro to Data Mining
By Mary Kane

c. Data Quality Summary:

Issue ID	Issue Type	Column(s)	Rows affected	Evidence (examples)	Fix / Decision	Status
1	Inconsistent category values	Country	385	FR, ES, DE mixed with full names (France/Spain/Germany)	Standardized to full country names using mapping (FR→France, ES→Spain, DE→Germany).	Fixed
2	Inconsistent category values	Gender	68	M/F mixed with Male/Female; plus NULL	Standardized M→Male and F→Female. Null Changed to Unknown	Fixed
3	Invalid numeric values	Age	11	Row_Number=625 Age=2; Row_Number=630 Age=2; Row_Number=633 Age=2	Set invalid ages to blank Removed.	Fixed
4	Missing values	Credit Score	3	3 rows have Credit Score missing.	Imputed using the MEAN value of the column	Fixed
5	Missing values	Estimated Salary	2	2 rows have Estimated Salary missing.	Imputed using the MEAN value of the column	Fixed
6	Missing values	Last_Name	1	1 row has Last_Name missing.	Changed the blank to Unknown	Fixed
7	Missing values	Gender / Age	2	1 Gender cell and 1 Age cell were NULL.	Changed the blank to Unknown	Fixed

5.4. Intro to Data Mining
By Mary Kane

3. Identifying Leading Factors to Customer Loss:

NumOfProducts: There is a strong relationship between customers who adopted 3 or more products and those who left the Bank.

- 4 products: 100% customers left
- 3 products: 85% of customers left
- 1 product: 28% of customers left
- 2 products: 7% customers left

IsAnActiveMember?: 29% of customers who left were inactive. Only 12% were active members.

Country: Spain and Germany had higher customer loss rates than France.

- Germany: 29%
- Spain: 20%
- France: 16%

Credit Score Group: Customers with Lower scores had high rates in the Left customer group.

- Low: 24%
- Medium: 21%
- High: 18%

Age Range Group: Older and Mature customers pose a higher risk of leaving the bank than younger customers.

- Mature (40 – 60) 37%
- Older (61 – 82) 35%
- Young (18 – 39) 10%

5.4. Intro to Data Mining
By Mary Kane

4. Decision Tree: A rule-based decision tree was created to represent churn risk using the strongest drivers observed in pivot tables. The first split was NumOfProducts because churn rate increased sharply for customers with 3+ products. The second split was used in IsActiveMember because inactive customers showed substantially higher churn. Country and Age Group were used as tertiary split variables to refine risk classification.