Business Understanding

Objective: To increase customer retention, the sales team wants to identify the leading indicators that a customer will leave the bank. Use a table of client attributes to identify the top risk factors that contribute to client loss and model them in a decision tree.

Dataset: Pig E. Bank client data set

Data Understanding and Preparation

Cleaning data notes:

Check for integrity accuracy

- Check numerical columns to see if mix, max, mean, median, mode values show any unexpected values for columns.
    - Mode(19 occurrences) and max for Credit Score are 850 and data skews left
    - Min value of 2 doesn't make sense for Age. Occurs 11 times (only for Females in Spain).
    - No one has Tenure value above 10. Seems low

Check for integrity inconsistency

- Check frequency counts for columns with pivot table to find any formatting issue or inconsistencies
    - Last_Name contains ? values in names
    - Country has 2 letter or full name variations. Need to update FR, DE, and ES
    - Gender has values that need to be combined (F, Female, M, Male)
    - Gender has NULL value
    - Age has NULL value
    - There are 349 rows with Balance of 0
    - Estimated Salary has NULL value

Check for duplicates

- Row_Number and Customer_ID are unique per frequency pivot table. All other columns would make sense to have duplicates.
- No duplicates across all rows using remove duplicates function.

Check for missing values with a filter for each column

- Last_Name contains 1 blank
- Credit Score contains 3 blanks
- Estimated Salary contains 1 blank

Changes made to data:

- Removed 11 age value of 2. Values are likely typos for 20-29. Only occurred for Females in Spain. Made up 1 % of data.
- Removed Last_Name column to protect PII.
- Changed Country values of DE to Germany, FR to France, ES to Spain

- Changed Gender values of F to Female, M to Male
- Removed NULL value for Gender
- Removed NULL value for Age
- Removed NULL value for Estimated Salary

Cleaning notes to check during analysis:
- Mode(19 occurrences) and max for Credit Score are 850 and data skews left
- There are 349 rows with Balance of 0

Derived Columns
- Added column Exit_Status to group customers by if they left the bank or not.
- Added column Member_Status to group customers by if were active members or not.
- Added column Credit_Card_Status to group customers by if they had a credit card with the bank or not.

Numerical Variables
Credit Score
Age
Tenure
Balance
Number of Products
Estimate Salary

Category Variables
Country
Gender
Credit_Card_Status
Member_Status

Numerical Value Averages by Exit Status
(Found by pivot table and a few additional formulas)

	Left Bank	**Stayed**	**Overall Average**	**(Stay Avg - Left Avg) / (1 Std Dev of Variable)**
Average of Credit Score	636.5	651.6	648.5	15%
Average of Age	45.3	37.5	39.1	-76%
Average of Tenure	4.7	5.2	5.1	17%
Average of Balance	90239.2	74830.9	78002.7	-25%
Average of NumOfProducts	1.5	1.5	1.5	0%
Average of Estimated Salary	97155.2	98943.4	98574.5	3%

Notes:

- Age and Balance are the largest differences in means relative to the standard deviation of the variables as a whole.
- Tenure and Credit Score show smaller differences but may still be relevant.
- NumOfProducts and Estimated Salary look to have no impact on leaving or staying.

Category Variables by count by exit status and percentage of whole
(Found by pivot table and a few additional formulas)
Customer Count and % by Country and Exit Status

Country	Left Bank	% Left	Stayed	% Stayed	Grand Total	Gap (% Stay - %Left)
France	77	16%	403	84%	480	68%
Germany	75	29%	182	71%	257	42%
Spain	52	20%	202	80%	254	60%

Customer Count and % by Gender and Exit Status

Gender	Left Bank	% Left	Stayed	% Stayed	Grand Total	Gap (% Stay - %Left)
Female	121	26%	341	74%	462	48%
Male	83	16%	445	84%	528	68%

Customer Count and % by Credit Card Status and Exit Status

Credit Card Status	Left Bank	% Left	Stayed	% Stayed	Grand Total	Gap (% Stay - %Left)
Has Card	144	21%	556	79%	700	58%
No Card	60	21%	231	79%	291	58%

Customer Count and % by Member Status and Exit Status

Member Status	Left Bank	% Left	Stayed	% Stayed	Grand Total	Gap (% Stay - %Left)
Active	61	12%	442	88%	503	76%
Not Active	143	29%	345	71%	488	42%

Notes:

- Active Member Status customers stay much more. Highest % gap of staying to leaving.
- Male Gender and France as Country are 2nd highest % gap of staying to leaving.
- Credit Card Status looks to have no impact on leaving or staying.

Modeling

Selecting Factors

- Choices will be among Age, Balance, Tenure, Credit Score, Active as Member Status Male as Gender, France as Country.
- P value of 2 tail t test with unequal variance
    - Age: 1.18E-09
    - Balance: 0.0016
    - Tenure: 0.055
    - Credit Score: 0.81
- Two Sample Z Test of Proportions Pooled
    - Significant Z score at 5% is 1.96
        - Stay for Members Active vs Not Active :5.8
        - Stay for Members Male vs Not Male :3.4
        - Stay for Members French vs Not French :3.1
- Top 3 factors chosen
    - Age: Older more likely to leave
    - Balance: Higher balance more likely to leave
    - Member Status: Active Members more likely to stay

<u>Decision Tree Model</u>

Are Members Likely to Leave the Bank?