Exercise 5.4
Data Analytics Immersion
Andrea Watson

1. Downloaded
2. Data cleaning procedures:
    a. Removed unnecessary data – see table
    b. Checked for fully duplicated rows – none found
    c. Checked data types – no conversions necessary
    d. Checked for inconsistencies – see table
    e. Checked for outliers – see table

Column Name	Data Cleaning Performed	Reasons & Notes
Row_Number	Removed column	Not needed
Last_Name	Removed column	Personal Identifying Information, not needed and potentially biasing
Credit Score	Replaced 3 blank values with 'N/A'	
Country	Replaced 244 values of 'FR' with 'France' Replaced 23 values of 'DE' with 'Germany' Replaced 136 values of 'ES' with 'Spain'	Updated for consistency
Gender	Replaced 19 values of 'F' with Female Replaced 49 values of 'M' with Male Replaced 1 value of 'NULL' with 'N/A'	Updated for consistency
Age	Replaced 11 values of '2' with 'N/A' Replaced 1 value of 'NULL' with 'N/A'	Updated to remove outliers – 2 as an age is likely an outlier, as it would be unlikely that two year olds have their own bank account and are earning a salaty
Estimated Salary	Replaced 1 blank value with 'N/A' Replaced 1 value of 'NULL' with 'N/A'	

3.

Summary Statistics & Additional Analysis

	Credit Score	Age	Tenure	Balance	NumOfProducts	Estimated Salary
Exited Customers						
Min	375	22	0	$0	1	$417.41
Max	850	69	10	$213,146.20	4	$199,725.39
Mean	637	45	5	$90,239.22	1	$97,155.20
Current Customers						
Min	411	18	0	$0	1	$371.05
Max	850	82	10	$197,041.80	3	$199,661.50
Mean	652	37	5	$74,830.87	2	$98,943.39

Country	Current	% Current	Exited	% Exited	Total
France	403	84%	77	16%	480
Germany	182	71%	75	29%	257
Spain	202	80%	52	20%	254

Gender	Current	% Current	Exited	% Exited	Total
Female	341	74%	121	26%	462
Male	445	84%	83	16%	528

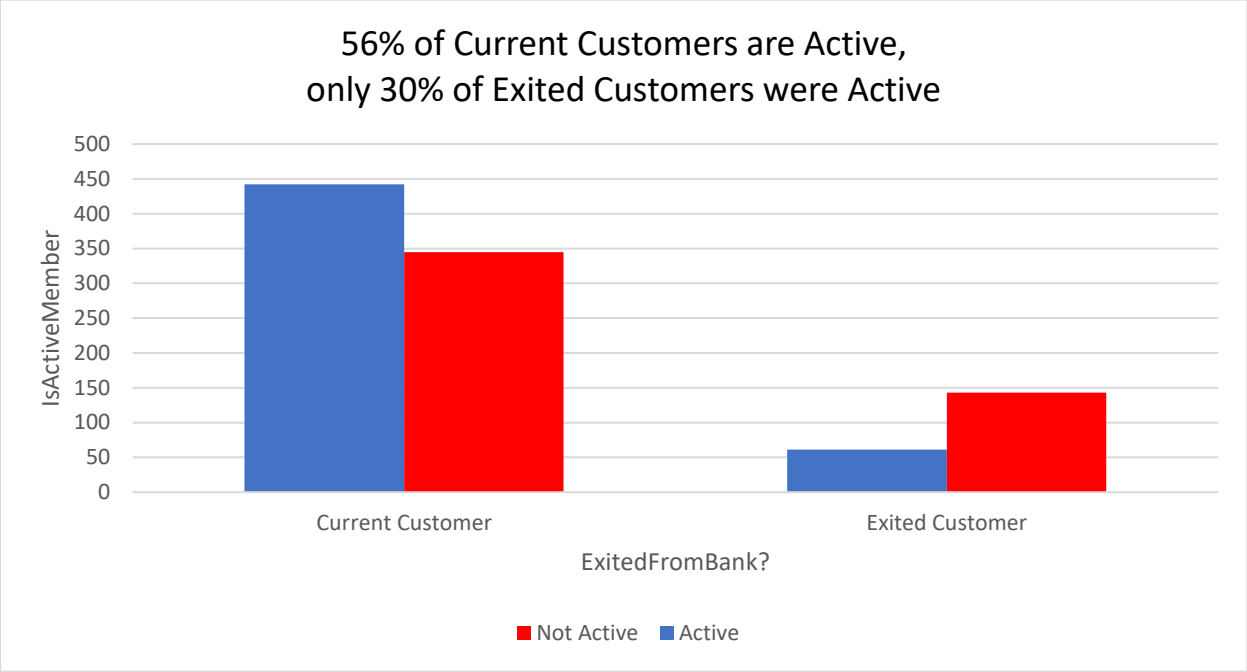
IsActiveMember	Current	% Current	Exited	% Exited	Total
No	345	71%	143	29%	488
Yes	442	88%	61	12%	503

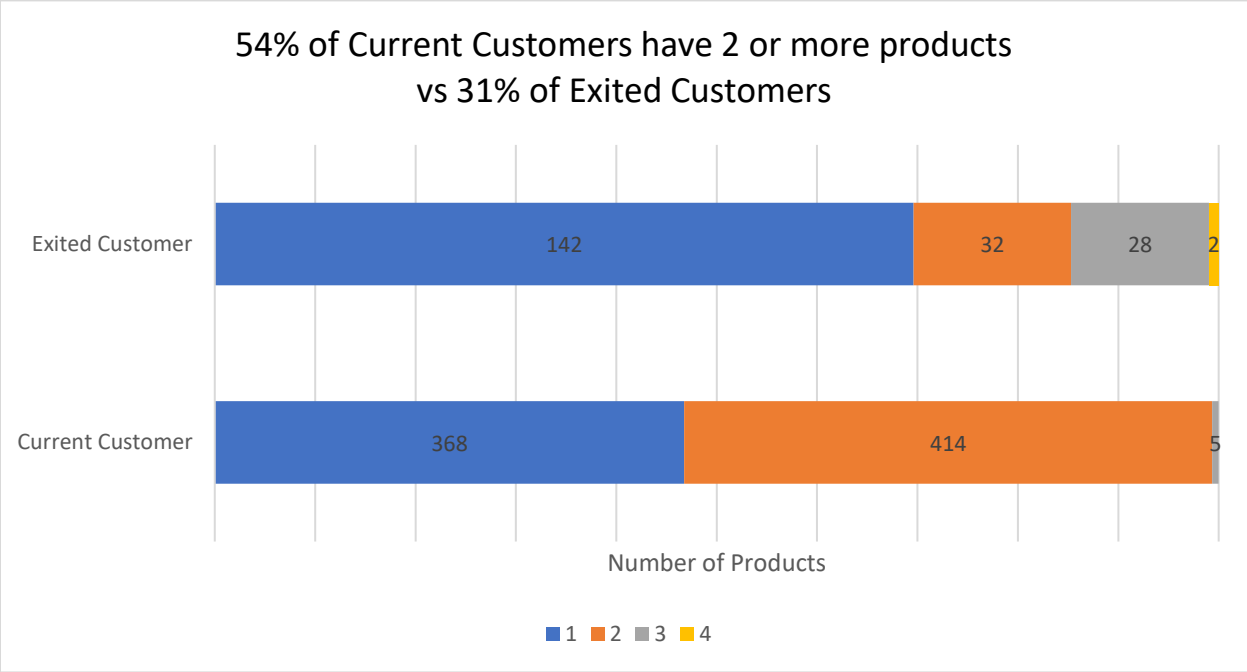
HasCrCard?	Current	% Current	Exited	% Exited	Total
No	231	79%	60	21%	291
Yes	556	79%	144	21%	700

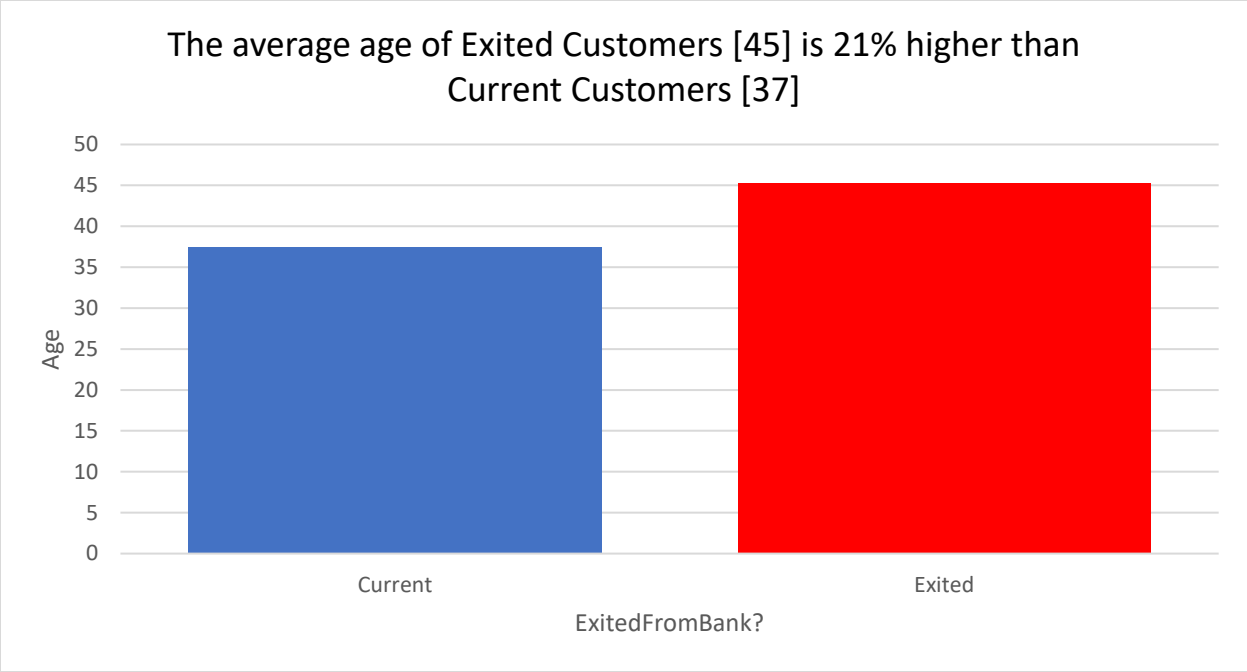
NumOfProducts	1	% 1	2	% 2	3	% 3	4	% 4	Total
Current Customer	368	47%	414	53%	5	1%	0	0%	787
Exited Customer	142	70%	32	16%	28	14%	2	1%	204

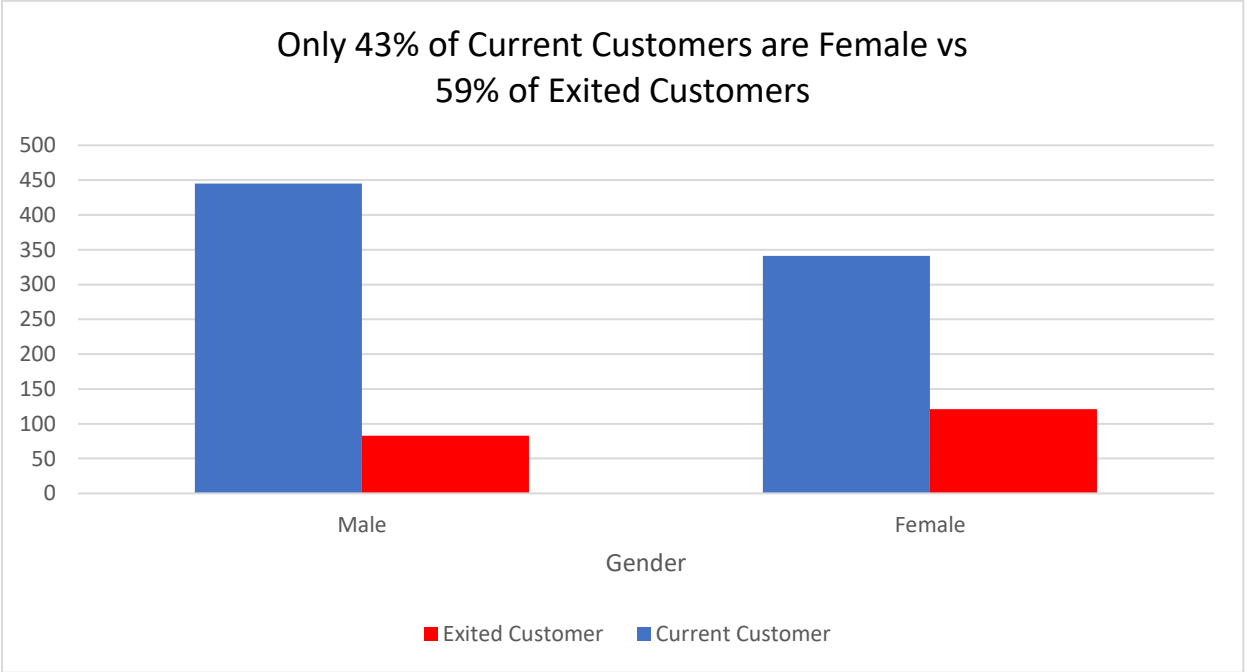
	Exited Customers		Current Customers	
Total Count	204		787	
	Count	% of Exited	Count	% of Current
HasCrCard?	144	71%	556	71%
France	77	38%	403	51%
Germany	75	37%	183	23%
Spain	52	25%	202	26%
IsActiveMember	61	30%	442	56%
Female	121	59%	341	43%
Male	83	41%	445	57%

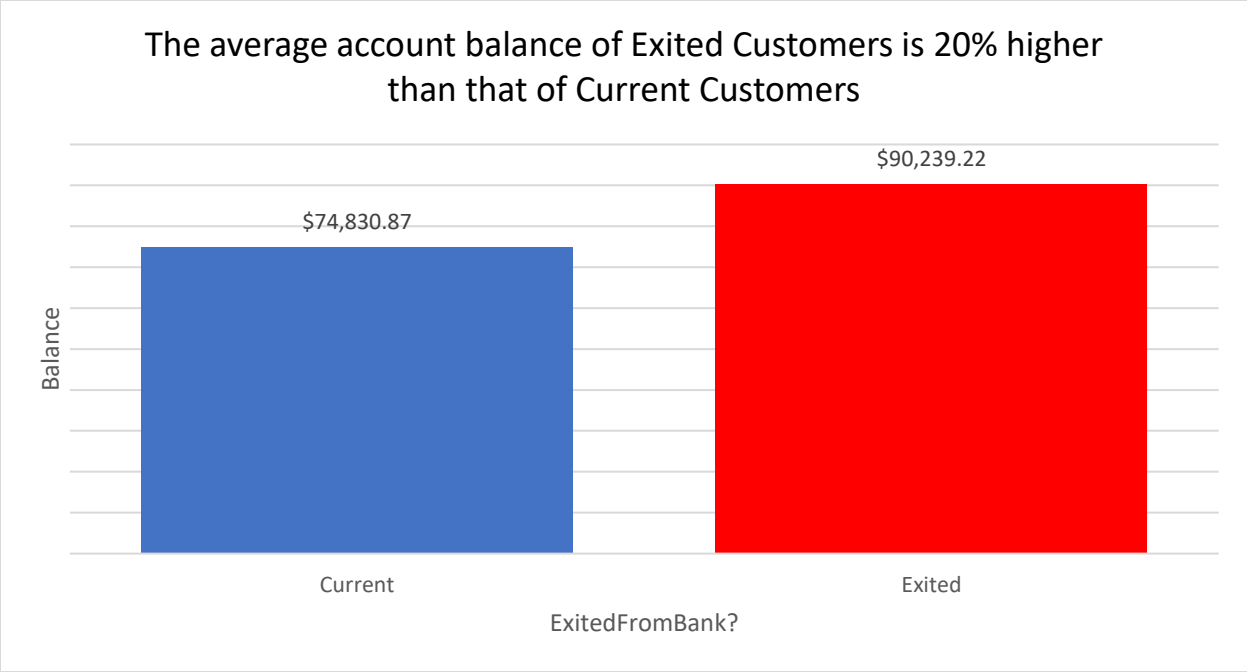
Insights and Visualizations

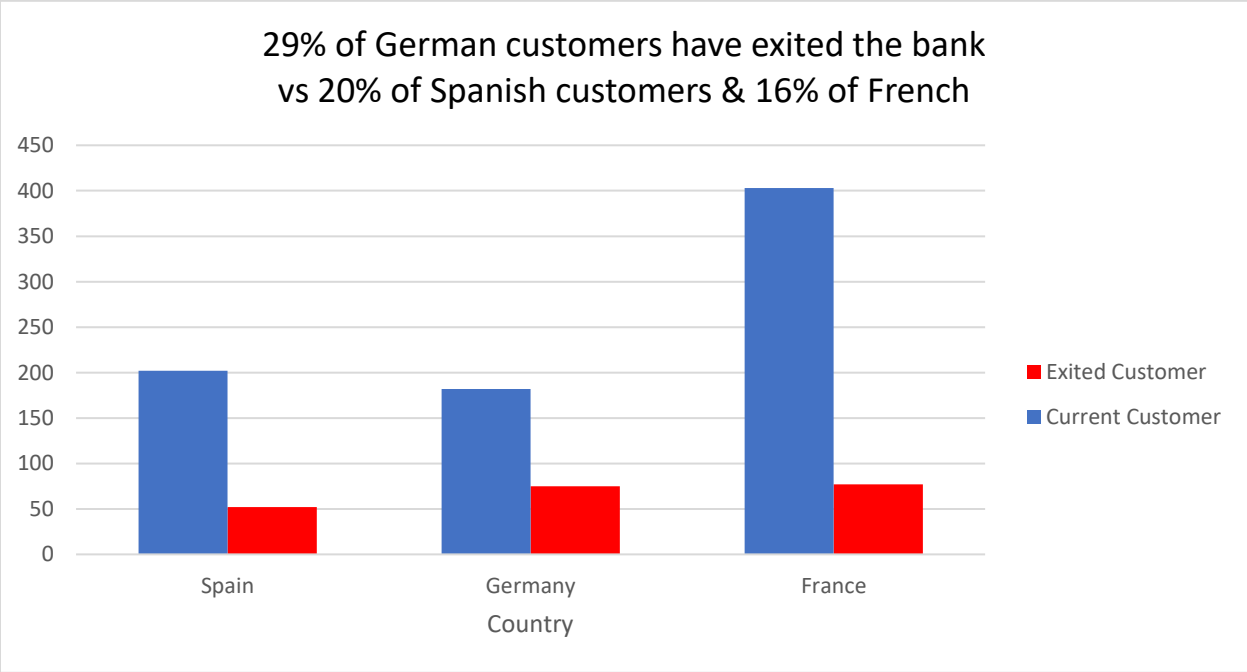





The average age of Exited Customers [45] is 21% higher than Current Customers [37]
50
45
40
35
30
25
20
15
10
5
0
Age
Current
Exited
ExitedFromBank?


Only 43% of Current Customers are Female vs 59% of Exited Customers
500
450
400
350
300
250
200
150
100
50
0
Male
Female
Gender
Exited Customer
Current Customer





Methodology

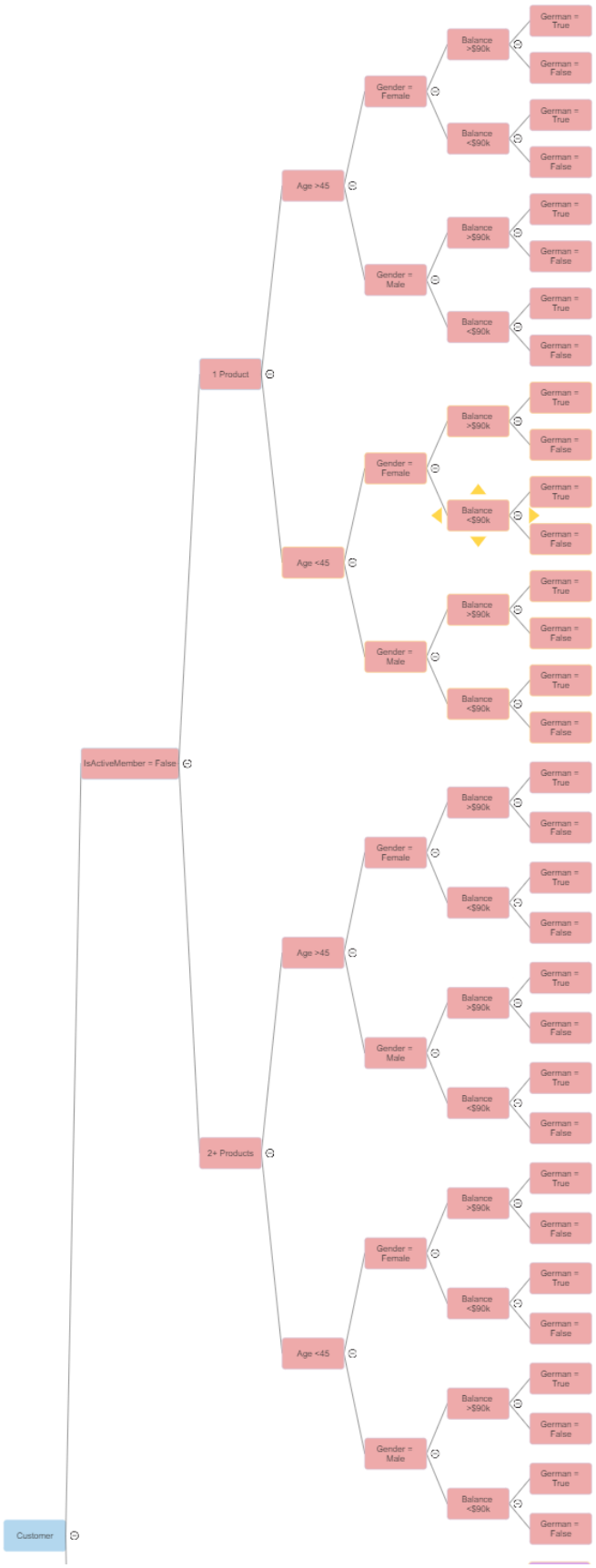
These insights were reached by using Excel to analyze statistical information about individual variables. I then used pivot tables to group, filter, transform, and further explore both individual variables and the relationship between variables, such as variable distributions and comparisons.

Summary of Results

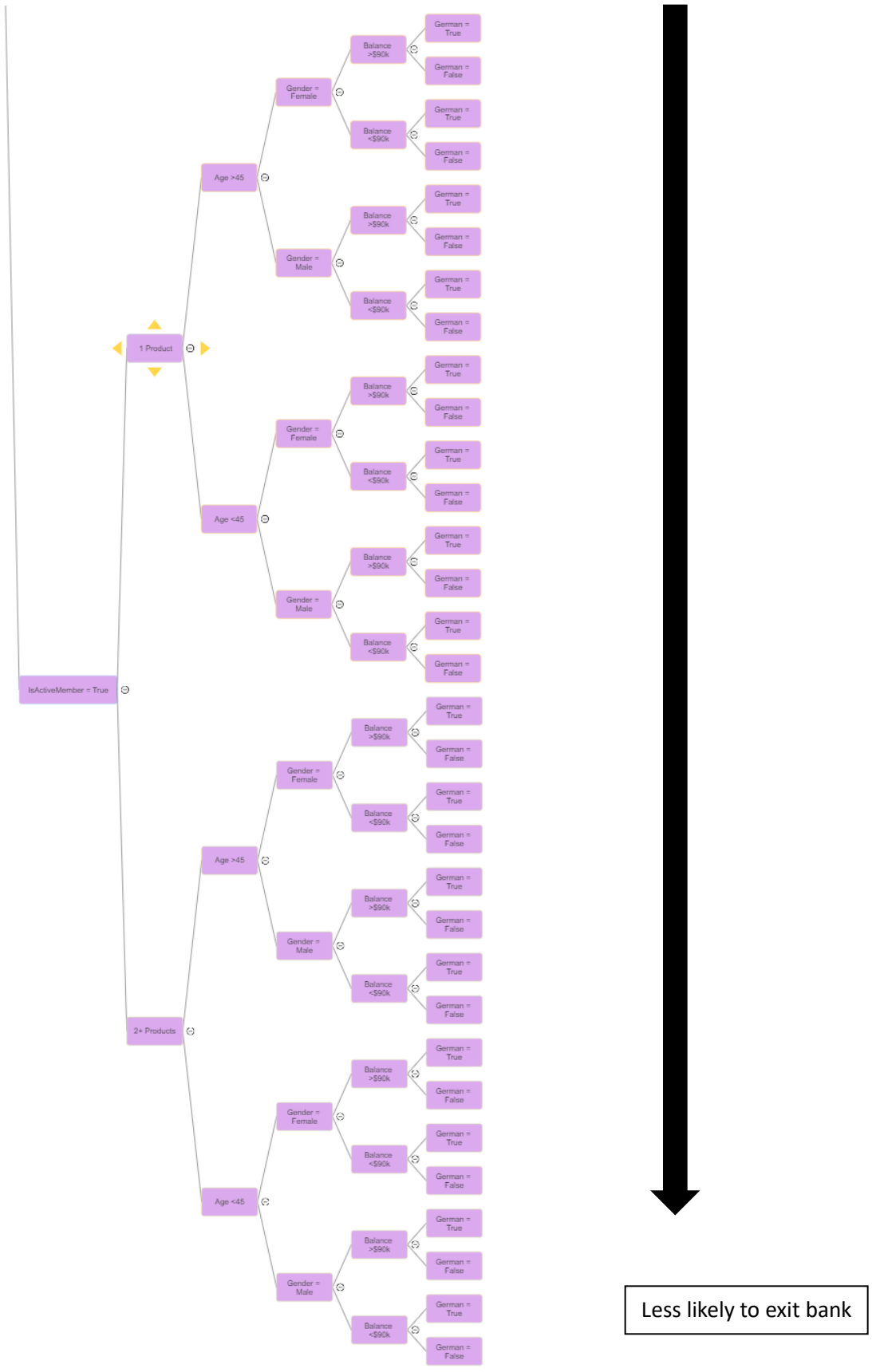
The primary factors that contribute to client loss are as follows in order of greatest to lesser impact:

- Active status
- 2 or more products
- Age >45
- Gender = Female
- Balance >$90k
- Nationality = German

4. Decision Tree – see below



More likely to exit bank


IsActiveMember = True
1 Product
2+ Products
Age >45
Age <45
Gender = Female
Gender = Male
Balance >$90k
Balance <$90k
German = True
German = False
Less likely to exit bank