Intro to Data Mining

Data Assessment and Cleaning

The data was taken from Pig E. Bank's data and is the most reliable dataset possible for this analysis. The original dataset contains PII and will be treated securely. The original dataset was retained and stored separately; a working copy was created and will be cleaned and manipulated as necessary.

- The analysis requested does not require customer names.
    - The customer field (column) was deleted.
- The Row_Number field is not necessary for Excel review.
    - The Row_Number field was deleted.
- There were no duplicate records in the remaining dataset.
- It appears that the reported credit scores are all using US standards and are between 300 and 850.
- There were three missing values for credit score.
    - Since the records with missing values represent only 0.3% of the dataset, and the balances were greater than the mean of all records, the mean of the reported values for credit score (649) is considered conservative and will be used for this analysis.
- There were inconsistencies in the country field where both the full name of the country and a two-letter abbreviated version were used.
    - The two-letter abbreviated versions were converted to full country names for this analysis.
- There were inconsistencies in the gender field where gender was reported as either F, Female, M, or Male.
    - For consistency, gender reported as Female was converted to F and gender reported as Male was converted to M.
- There was a missing value for gender.
    - Since the record with the missing value represented 0.1% of the dataset, the NULL value was converted to "Not Reported".
- There were inconsistencies in the age field where 11 records contained reported age of 2.
    - It is of note that 8 of the 11 have credit cards, all 11 reported gender as female and country as Spain. It is also of note that 2 is the only age reported under 18. It appears that the reported age of 2 is an error. As the total number of records believed to be in error represents 1.1% of the dataset, the average all ages reported (39) will be used for this analysis.
- There was one missing value in the age field.
    - Since the record with the missing value represented 0.1% of the dataset, the NULL value was converted to the average age of the dataset (39).

- The Tenure field did not show any obvious data inconsistencies or missing values.
- The Balance field showed that over 35% of the records had a zero balance.
    - It is noted that approximately 5.7% of the records in the dataset show both zero balance and that the customer has exited from the bank.
    - No action will be taken with respect to the balance field at this time. The zero balances do not appear to be obvious errors.
- There are no obvious data inconsistencies or missing values for the NumOfProducts, HasCrCard?, or IsActiveMember fields.
- The Estimated Salary field contains values that appear to be too small for annualized numbers and too large for weekly, bi-weekly, or monthly numbers.
    - Since the range of numbers varies too greatly to be considered reliable for this analysis, this field was deleted.
- There are no obvious data inconsistencies or missing values for the ExitedFromBank? field.

Analysis

Limitations of dataset:

- No information was given with respect to the period or purpose for the Estimated Salary field, rendering it useless. It is possible that this field was mislabeled and could indicate estimated revenue from all accounts for the bank. I would have tried to get more information about this field; however, that was not possible for this exercise.
- No information was given with respect to products included in the balance field. It is unclear how much of each reported balance is associated with which product. It is also unclear whether any of the reported balances included bad debt. This data could have provided clarity regarding departing customers.
- No information was given with respect to the products offered in NumOfProducts. There is no way to determine if the products are saving, checking, investment, credit, or loan accounts.
- No information was given with respect to what constitutes an active vs. inactive member.
- The countries listed were all in the EU; however, the currency for the balance and estimated salary figures was in USD. Also, the credit scores appeared to conform to US FICO scores, but could also be US Vantage scores.
    - The assumption made for analysis was that these were US FICO scores.

Review of dataset:

The original dataset had 991 records, 204 of which (approximately 21%) were designated as having exited from the bank. Almost half of the records represented customers from France, with the other half broken up almost evenly between Germany and Spain. Basic statistical breakdowns are as follows:

All Records	
Averages	
Credit Score	648.51
Age	39.10
Balance	$ 78,002.72

Exited from Bank	
Averages	
Credit Score	636.57
Age	45.22
Balance	$ 90,239.22

Stayed with Bank	
Averages	
Credit Score	651.61
Age	37.51
Balance	$ 74,830.87

Most Frequent Value	
Gender	M
Tenure	2
Number of Products	1
Credit Card	1
Active Member	1

Most Frequent Value	
Gender	F
Tenure	1
Number of Products	1
Credit Card	1
Active Member	0

Most Frequent Value	
Gender	M
Tenure	8
Number of Products	1
Credit Card	1
Active Member	1

Count of Records by Country	
France	480
Germany	257
Spain	254
Total:	**991**

Count of Records by Country	
France	77
Germany	75
Spain	52
Total:	**204**

Count of Records by Country	
France	403
Germany	182
Spain	202
Total:	**787**

% of Records by Country	
France	48.4%
Germany	25.9%
Spain	25.6%
Total:	**100%**

% of Records by Country	
France	37.7%
Germany	36.8%
Spain	25.5%
Total:	**100%**

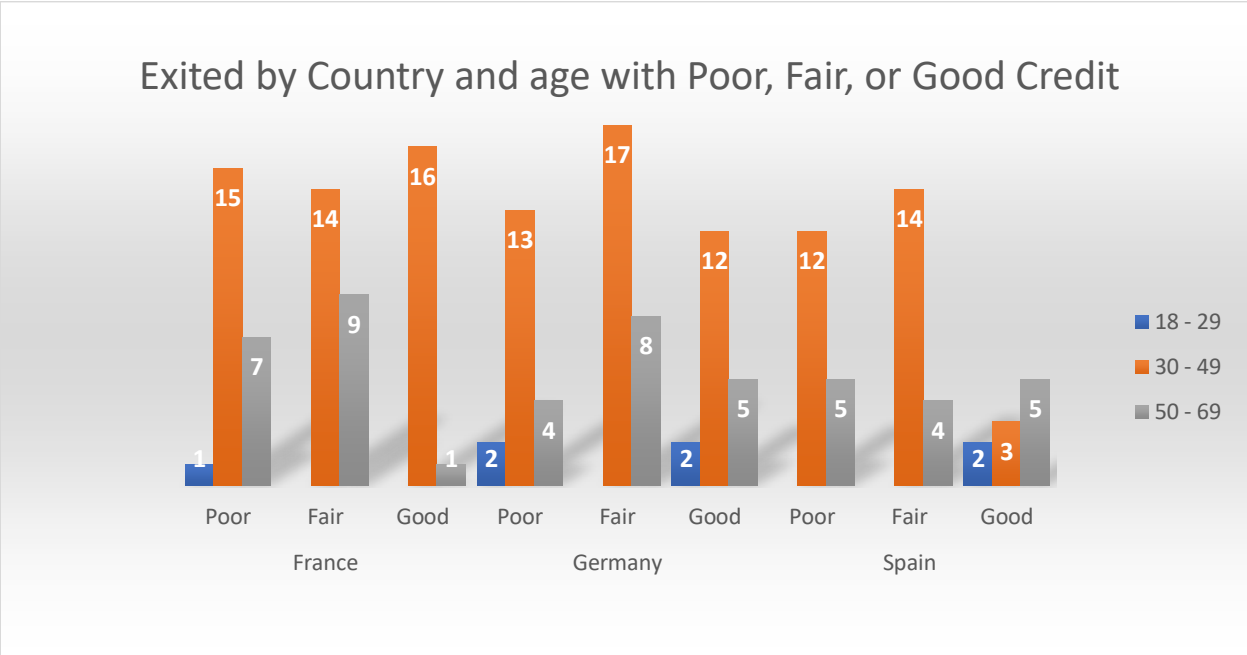
% of Records by Country	
France	51.2%
Germany	23.1%
Spain	25.7%
Total:	**100%**

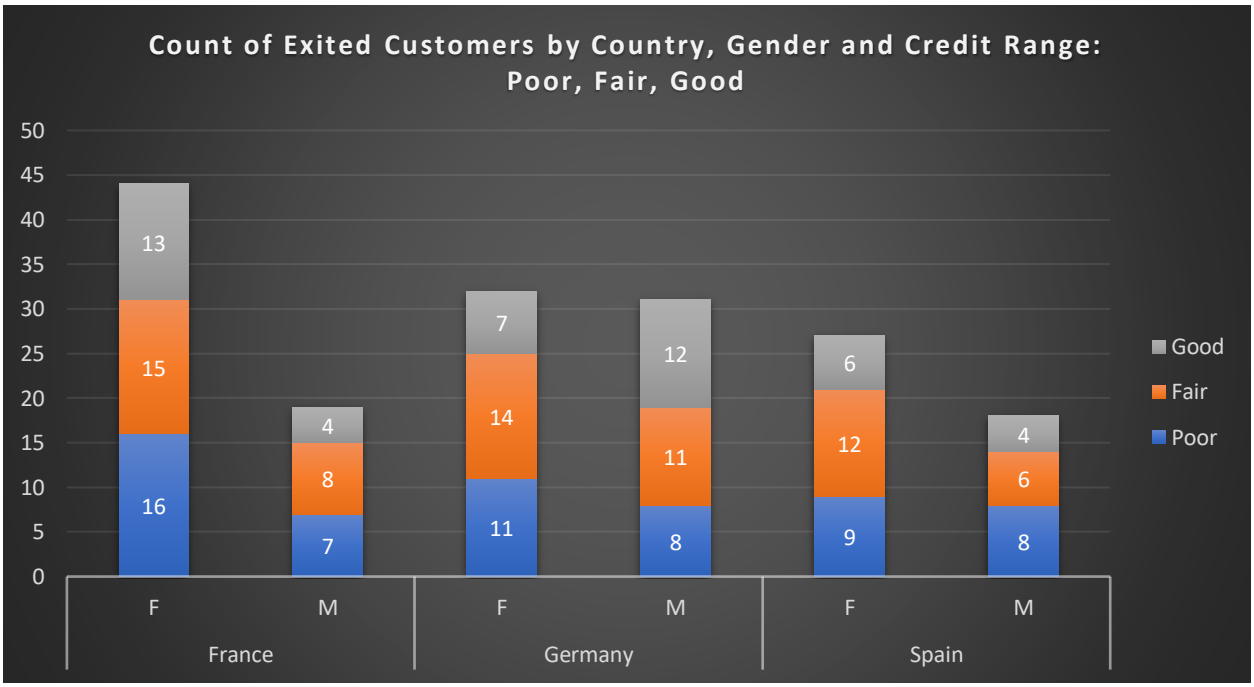
It was noted that there were commonalities in those that exited the bank that were different from the original dataset and from those that stayed. It appears that the credit scores were lower, and the age was slightly higher in the group that exited than in the other two groups. Further, it appears that the group of clients who exited were less likely to be active members.

More females exited than males. Based on tenure, it appears that clients who had been with the bank a shorter amount of time were more likely to exit than longer tenured clients. Finally, more clients from Germany exited that would be expected based on the distribution based on the original dataset and those that stayed with the bank.

To make the analysis of these differences easier, flags were added to the dataset to break down the Credit Score (based on FICO ranges), Balance, and age. The initial theory was that a specific age range was more likely to have exited, notably aged between 30 and 49. However, the analysis showed that no customers aged 70 or older exited the bank. This was determined to be the primary criterion for the decision tree.

Although other factors, such as tenure and balance may have played a part, the analysis showed that credit rating and gender were more definitive. Specifically, credit ratings under 740, which FICO uses for ratings of Poor, Fair, and Good were more closely related to those who exited. More females than males exited in all instances.





These findings were used to complete the decision tree.

Decision Tree