Bank Customer Churn Analysis & Prediction: (Feature Engineering)

Prepared by: Hassan Waked – Team Leader, DEPI Graduation Project.

Date: 20th Jan 2024.

---

1. Label Encoding:

Label Encoding was applied to the Column **gender**. The categories were mapped as follows:

- '**female**' was encoded as 0
- '**male**' was encoded as 1

A new column named **genderlabel** was created to store the encoded values.

---

2. One-Hot Encoding:

One-hot encoding was applied to the categorical column **geography**.
This resulted in three new binary columns:

- **geography_france**
- **geography_germany**
- **geography_spain**

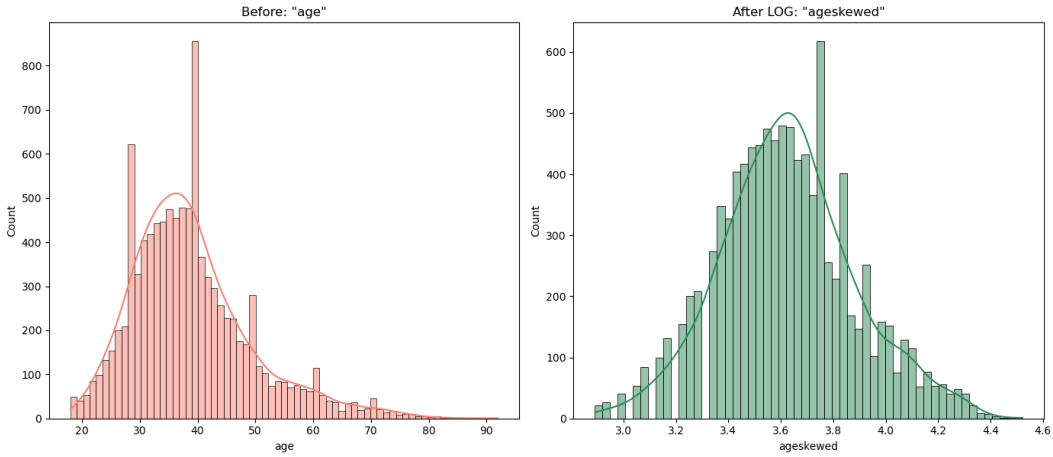
The original geography column was removed after encoding.

---

3. Log Transformation:

The Column **age** was transformed using logarithmic transformation due to its right-skewed distribution.

- The original skewness was **1.0118**, which indicates a positive skew.
- After log transformation, the skewness became **0.1824**, making it approximately symmetric.



The transformed column was renamed to **ageskewed**, and the original age column was removed from the dataset.

---

4. p-value Statistical Testing:

Each feature was statistically tested for significance in relation to the target column exited. The features that showed **significant relationships** (with p-value < 0.05) include:

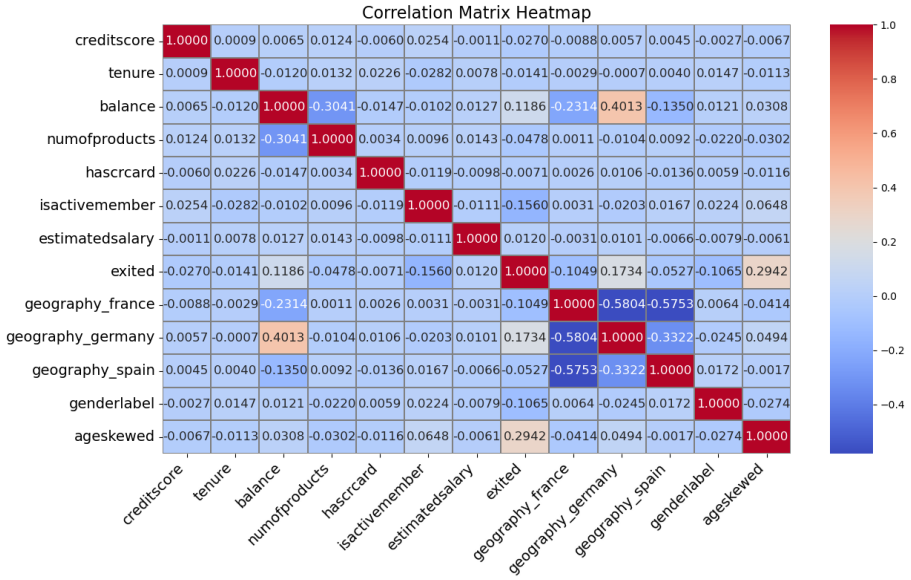
- **creditscore**
- **balance**
- **numofproducts**
- **isactivemember**
- **geography_france**
- **geography_germany**
- **geography_spain**
- **genderlabel**
- **ageskewed**

The following features were **not statistically significant**:

- **tenure**
- **hascrcard**
- **estimatedsalary**

---

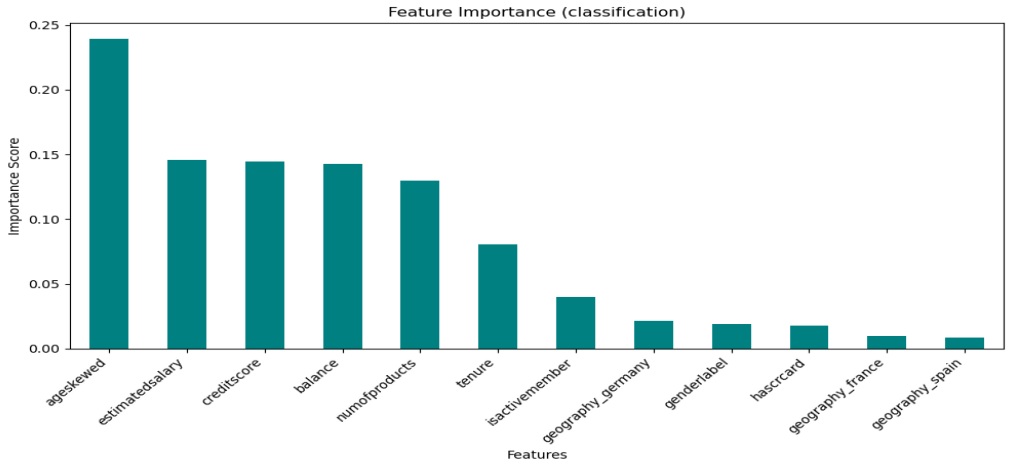
5. Correlation Heatmap:



6. Feature Importance Analysis:

Feature importance was computed using a classification model to identify the most influential variables in predicting the target.
The top contributing features were:

- **ageskewed (most important)**
- **estimatedsalary**
- **creditscore**
- **balance**
- **numofproducts**
- **tenure**
- **isactivemember**
- **geography_germany**
- **genderlabel**
- **hascrcard**
- **geography_france**
- **geography_spain**



Conclusion:

This report documented the complete preprocessing and feature engineering pipeline for the churn prediction project.
Transformations included encoding, log transformation, and statistical evaluation to select the most relevant variables.

The original dataset shape was (9996, 13) after feature engineering.

The dataset is now ready for splitting, model training and evaluation.