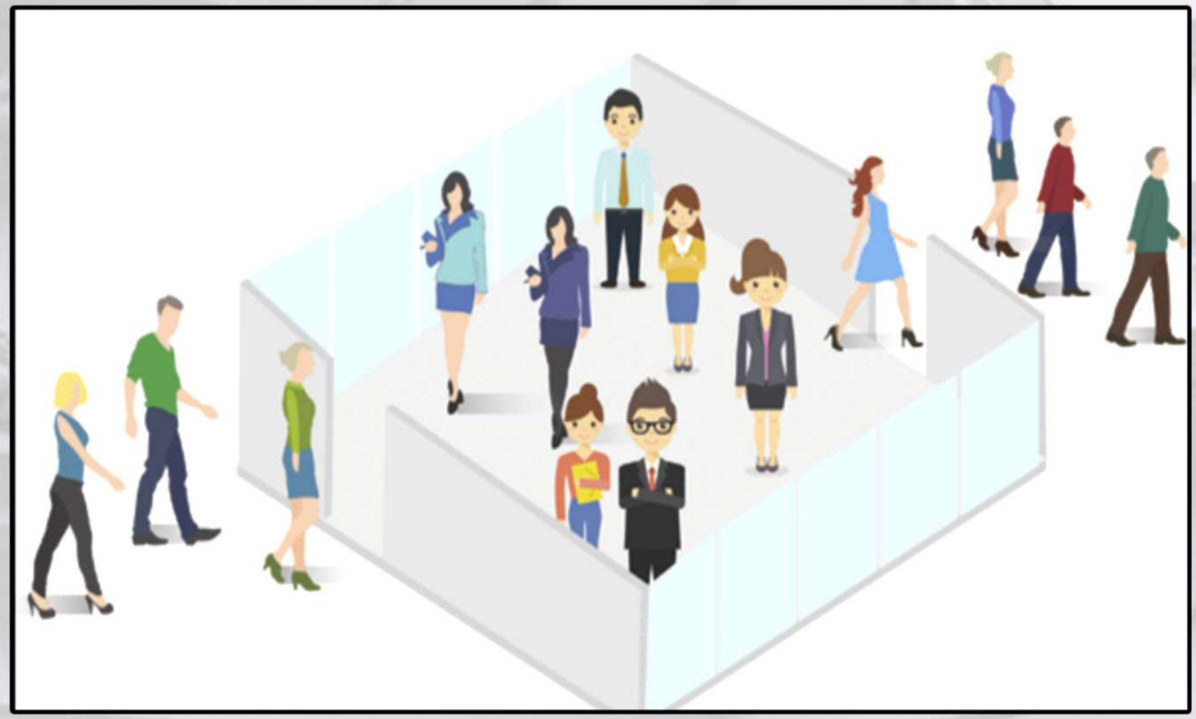
Customer Churn Predications - Logistic Regression & Artificial Neural Network

Padma Vasudevan

Agenda

- Introduction
- Problem Statement
- Solution Approach
- Dataset Overview
- Exploratory Data Analysis
- Modeling & Performance

Problem Statement



Customer attrition or churn, is a *critical phenomenon* in the banking industry that refers to the rate at which customers leave or discontinue their relationship with a particular bank.

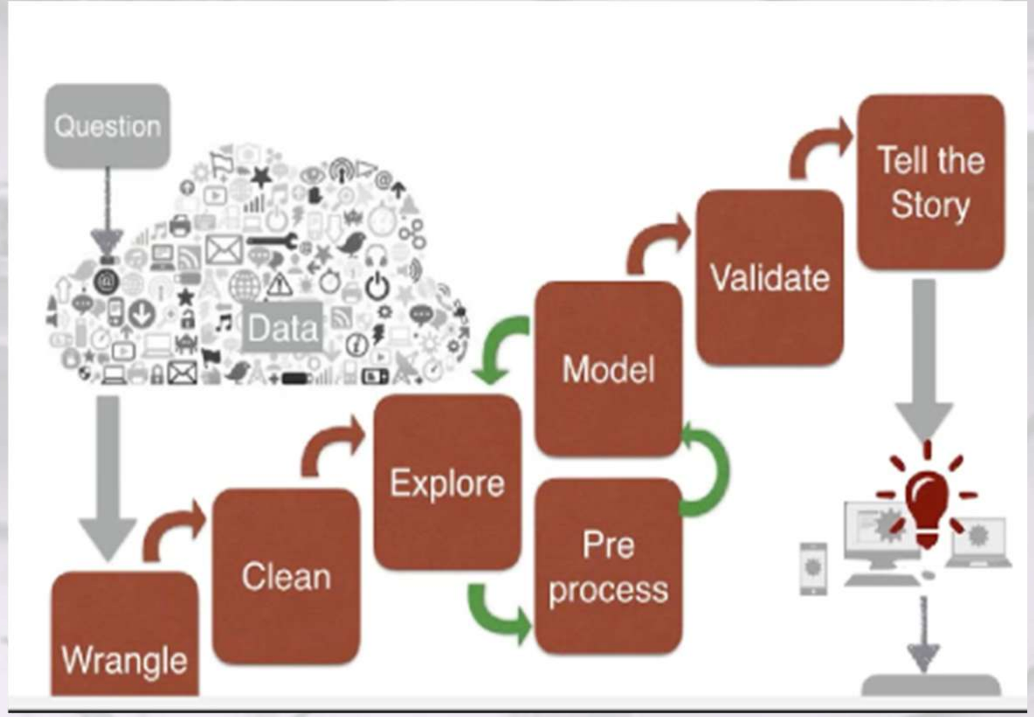
The objective of this project is to analyse bank customer data to identify key factors influencing the customer churn and help bank to develop retention strategies.

Solution Approach

- *Conducted data inspection*: Assessed data structure, handled missing values, and removed duplicates.

- Performed *EDA* using Python to analyse relationships between factors like *Credit score, Balance,* and *Exited*.

- Built a *logistic regression* model to predict customer churn and validated the model using *metrics* and *confusion matrix*.



Dataset Overview

Shape of the dataset: (*10000, 14*)

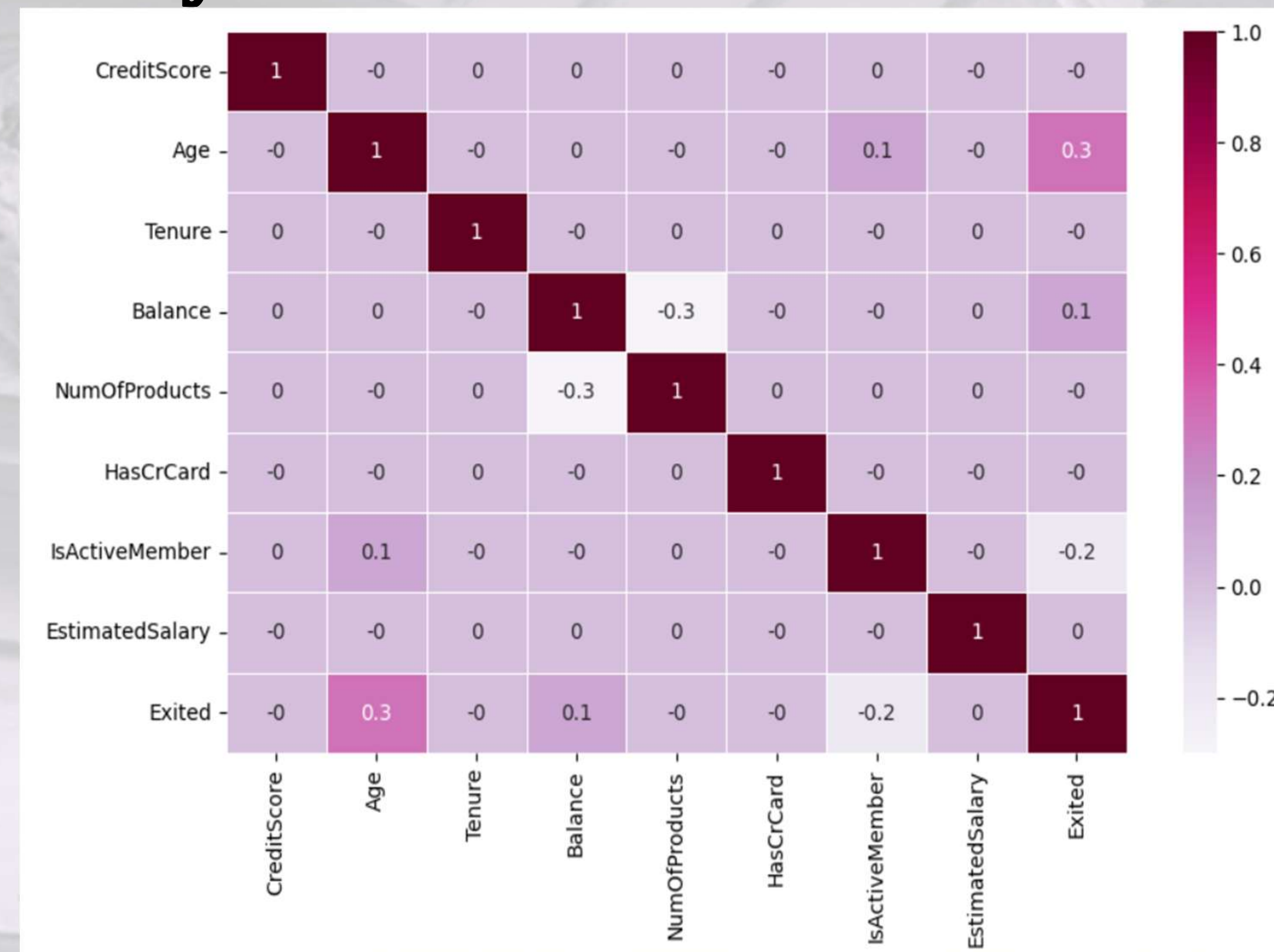
No Missing values
No Duplications
All the features are in proper datatypes

After Feature Engineering:
Shape: (*10000,11*)

Categorical Variables:
Surname, Gender, Geography

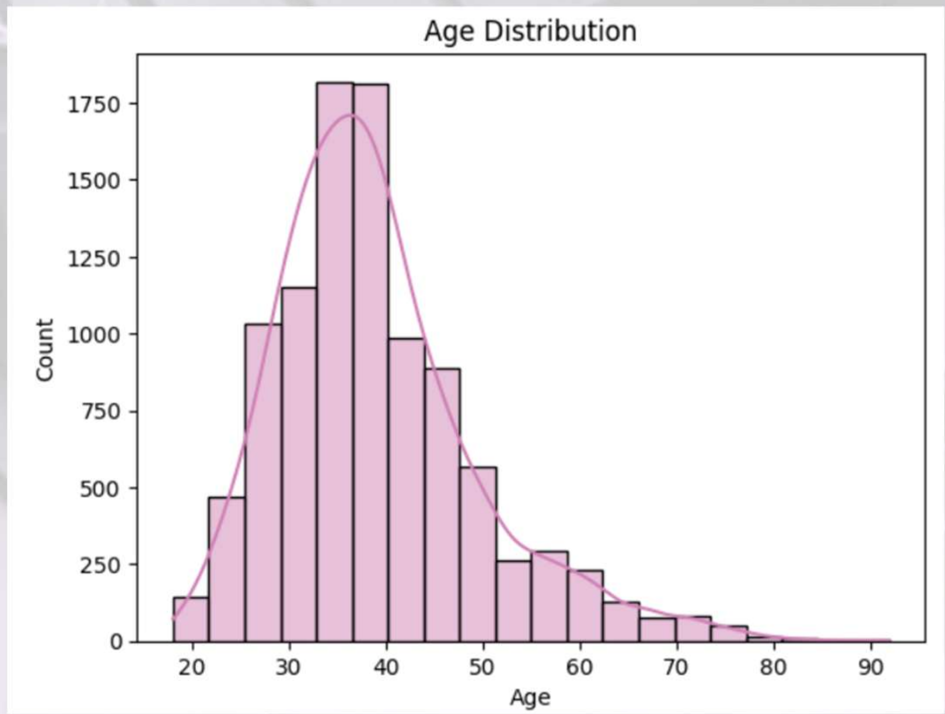
Numerical Variables:
'Age', 'Tenure', 'Balance', etc.,

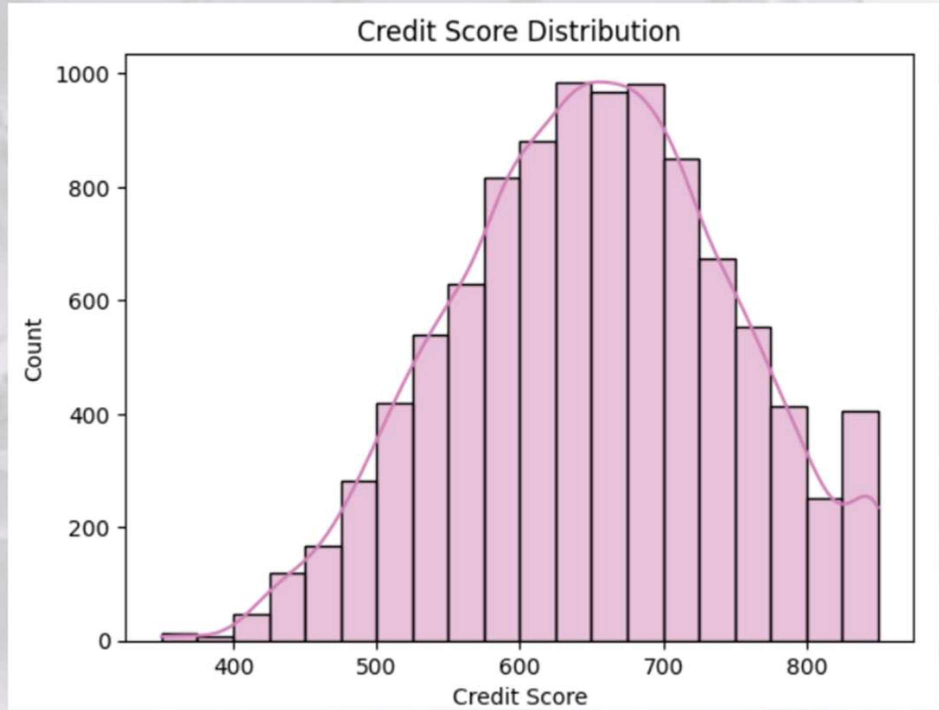
Exploratory Data Analysis



The correlation matrix indicates that most features have a weak correlation with the target variable, *Exited*.

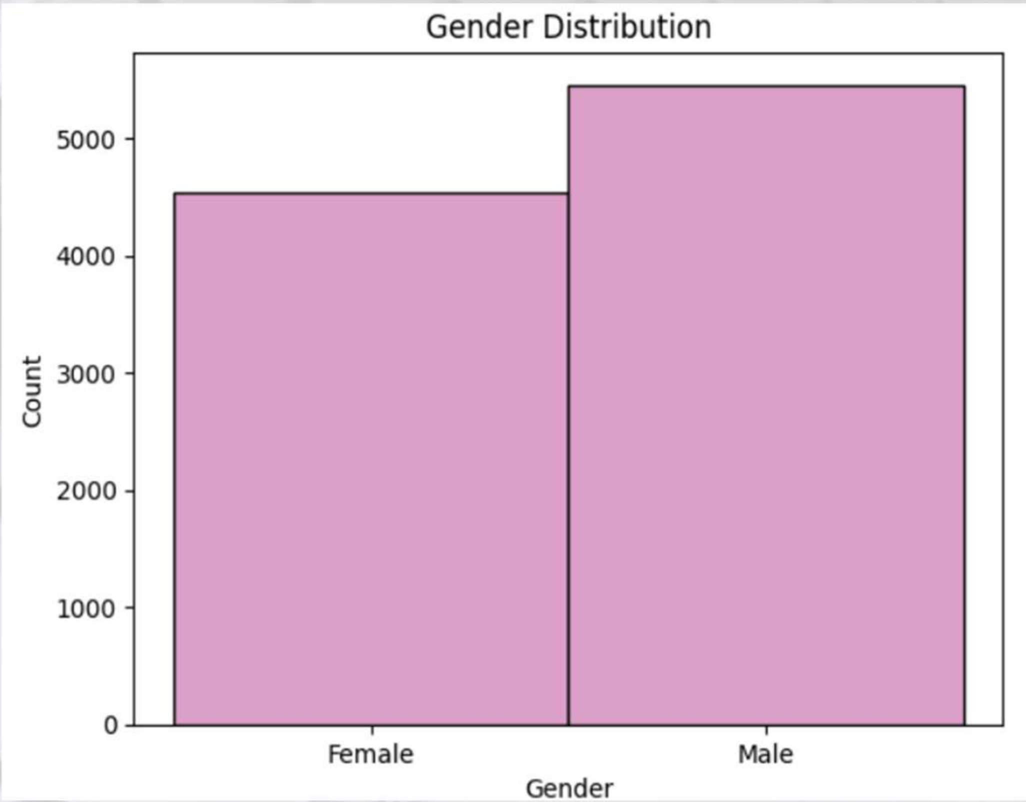
However, *Age* and *Balance* show a moderate correlation with the target, suggesting they may influence customer churn.

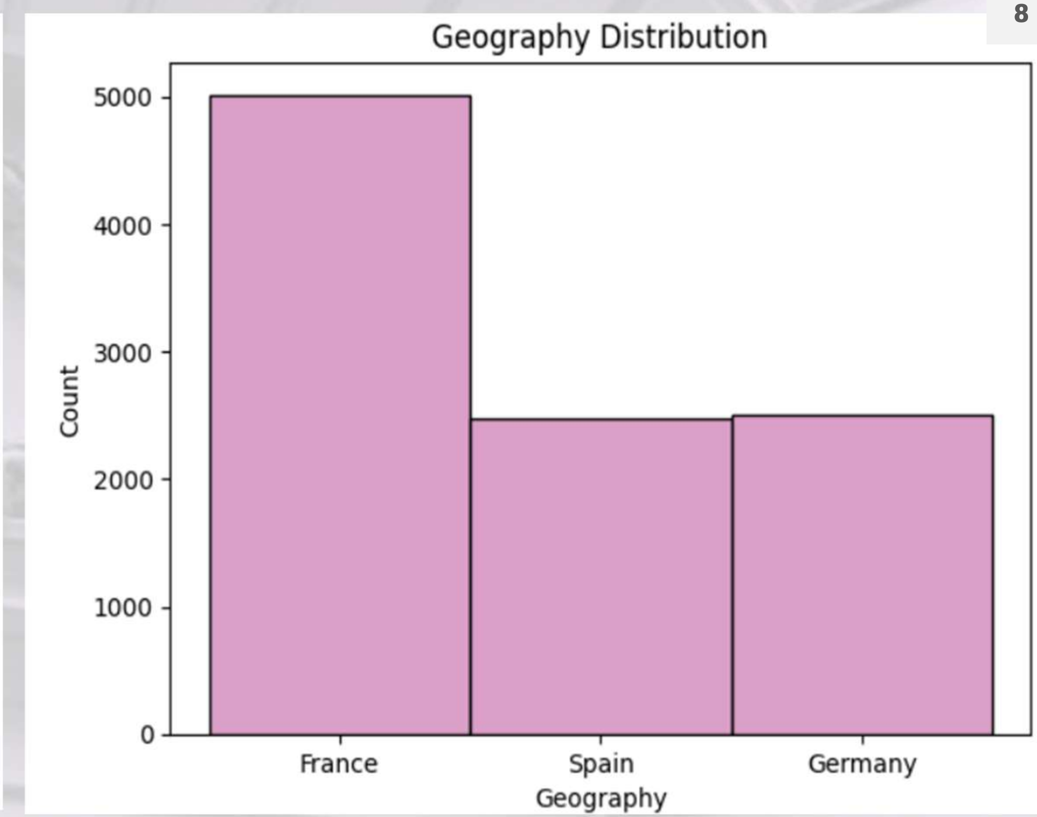




This plot shows the *distribution of ages* in our dataset.
From this we can observe that the majority of the customers in this dataset are in the age group between *30 to 45*.

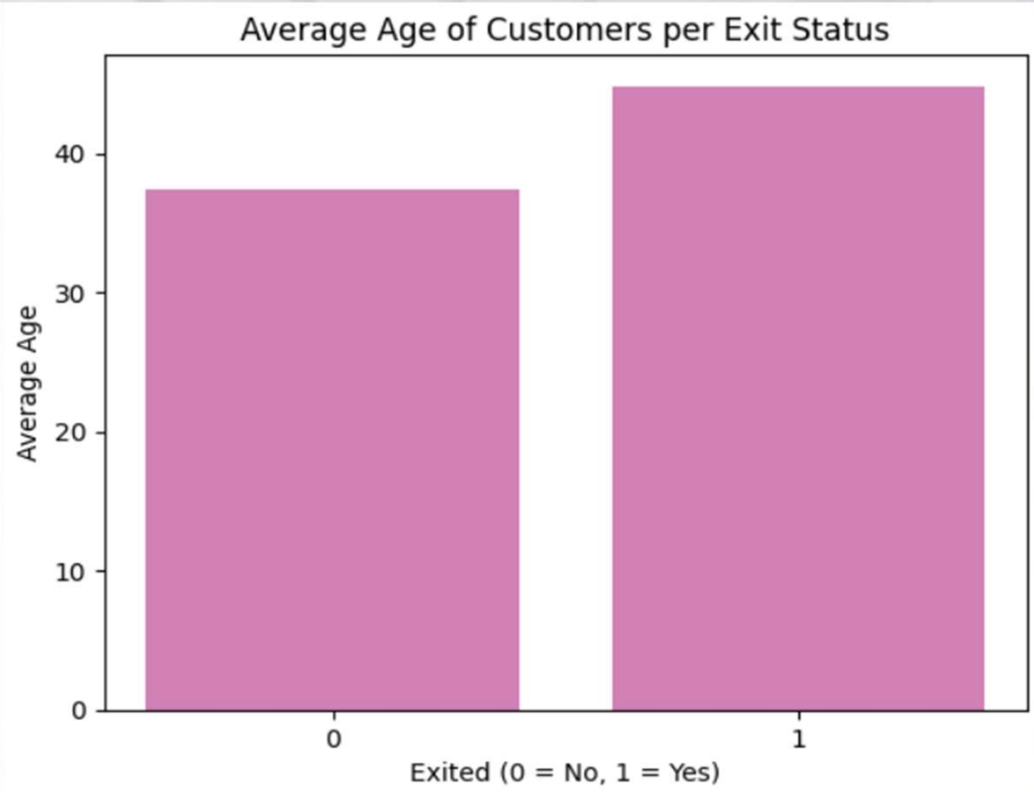
This plot shows the *distribution of credit score* data in our dataset.
From this we can observe that the majority of the customers in this dataset have credit score between *600 to 700*.

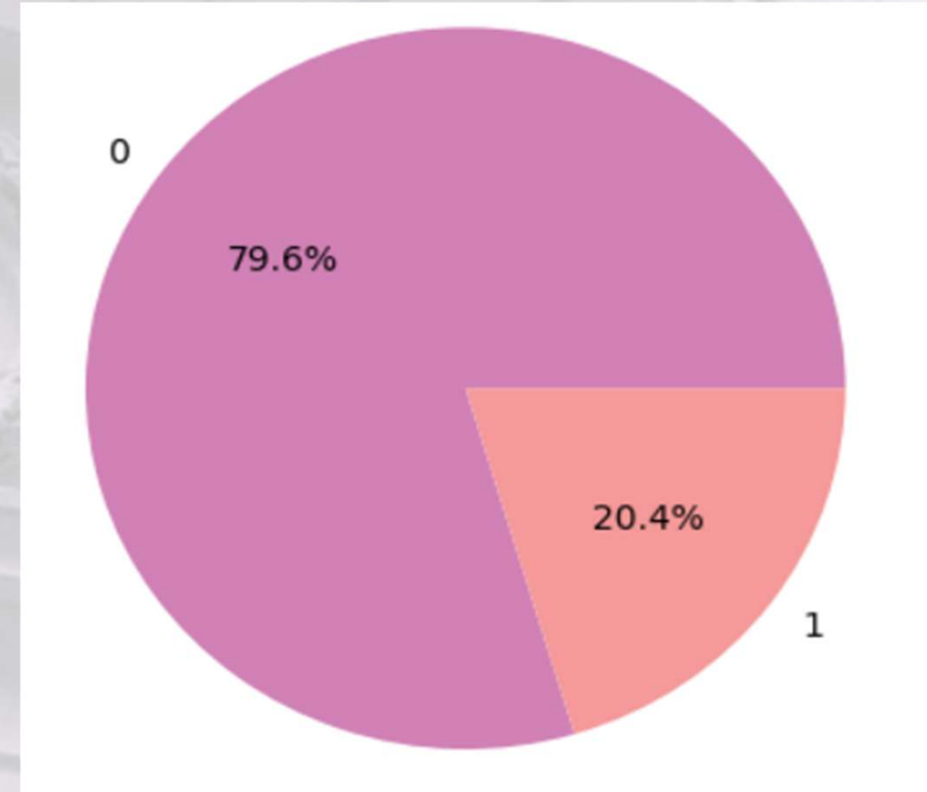




The *gender distribution* shows a slight male dominance, with *5,457 males and 4,543 females*, indicating a fairly balanced dataset for *gender-based analysis*.

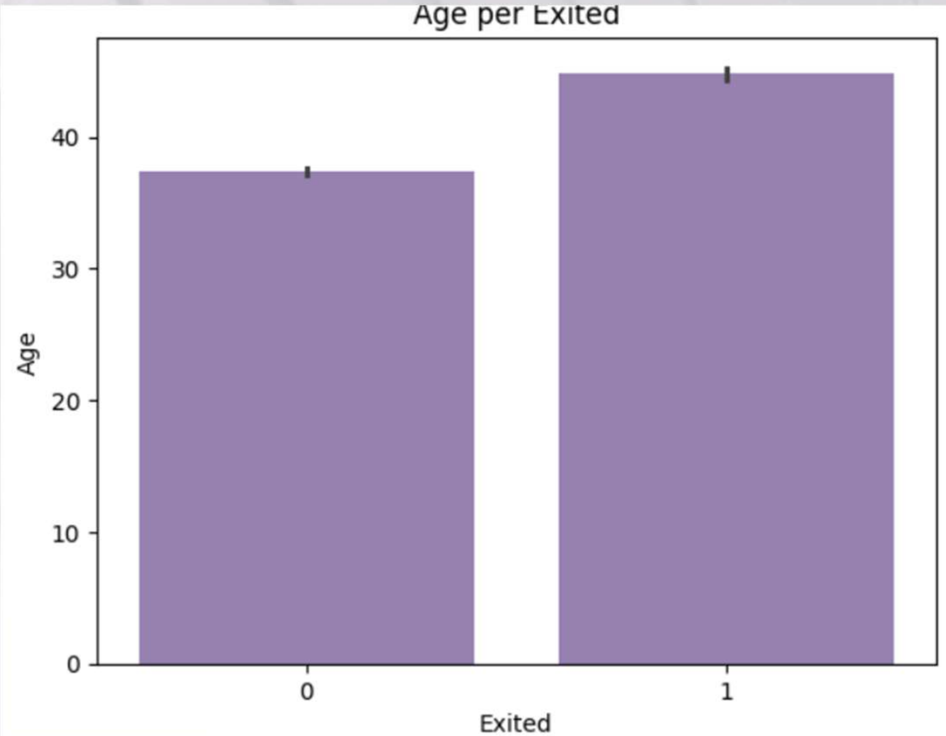
The data shows that the majority of entries are from *France (5,014)*, followed by *Germany (2,509) and Spain (2,477)*, highlighting France as the dominant segment for country-based analysis

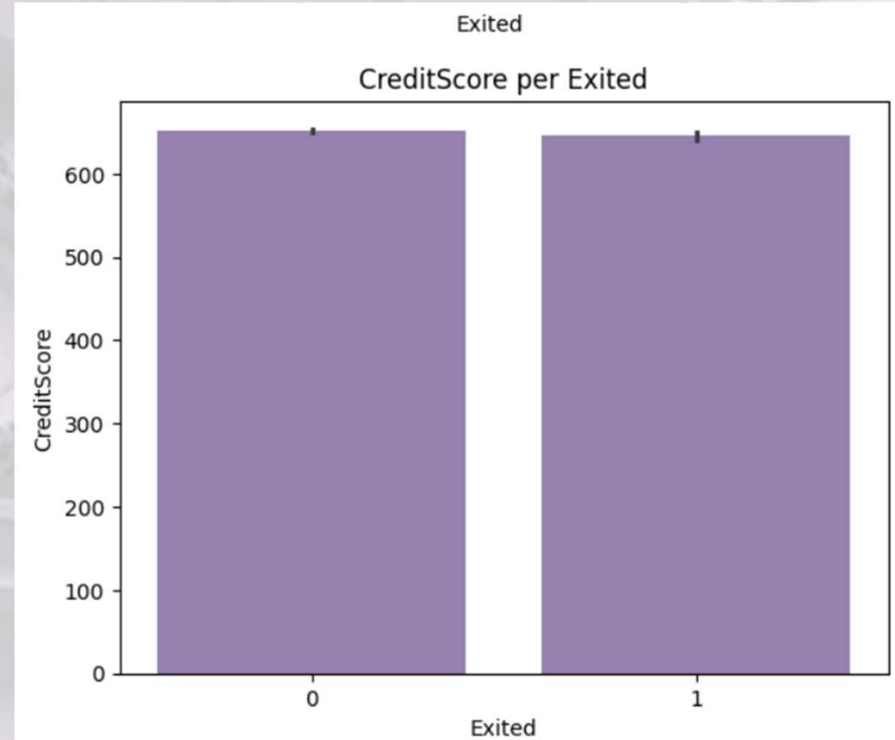




This plot suggests that older customers are more likely to churn, indicating *age could be a significant factor influencing customer retention.*

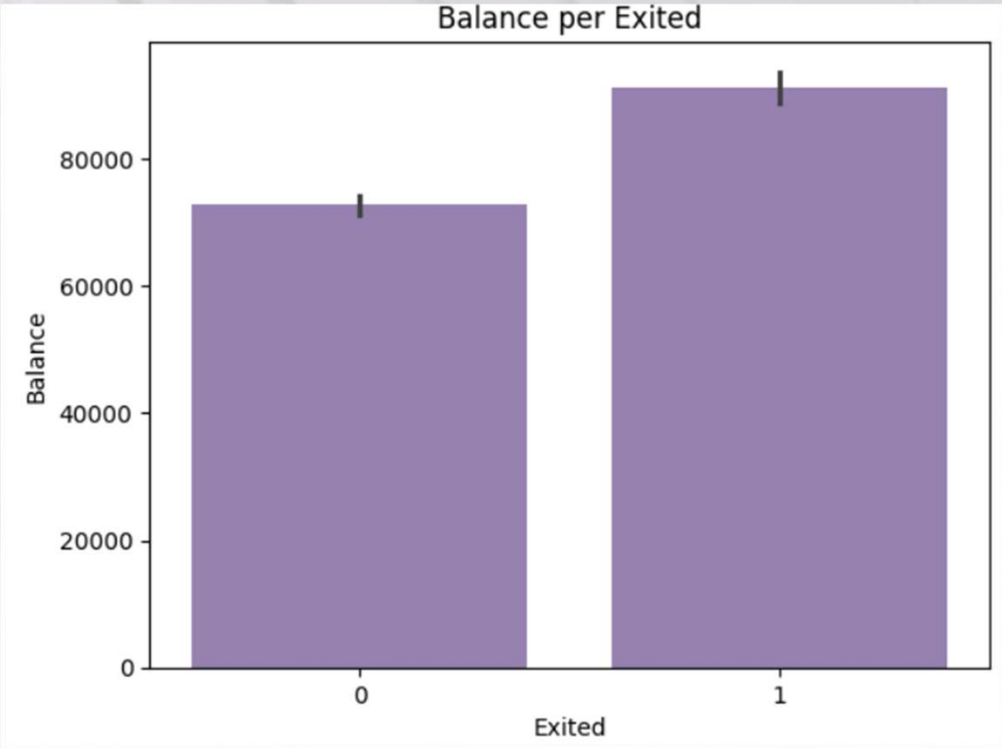
The pie chart shows that *79.6%* of customers stayed, while *20.4%* exited, indicating a relatively *low churn rate* in the dataset.

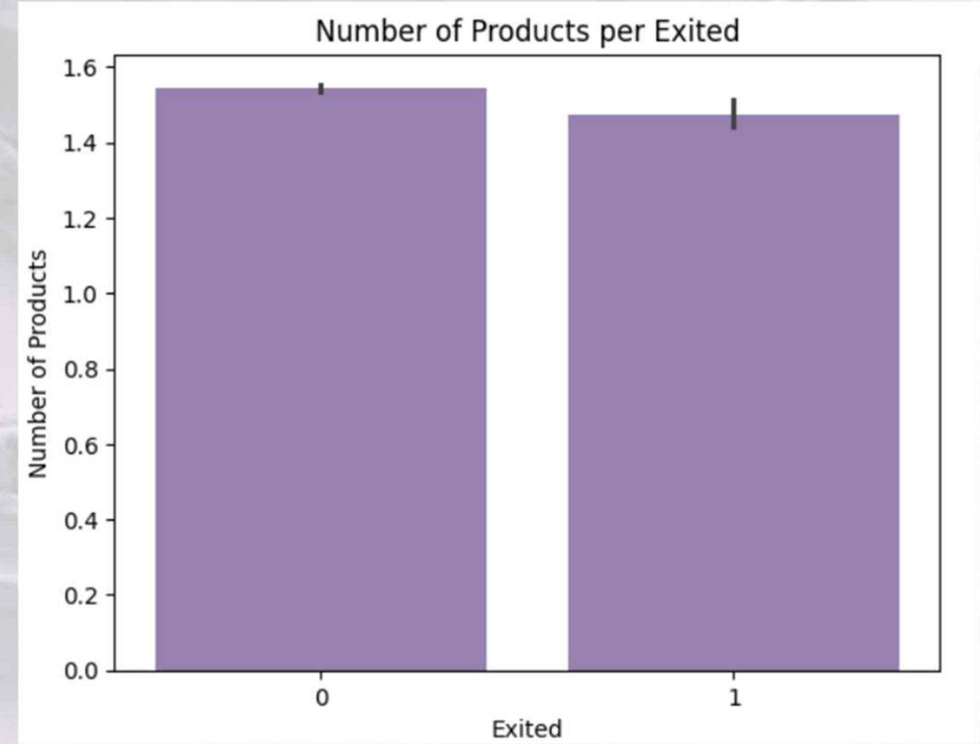




Customers who exited have an average age of *44.84,* higher than non-exited customers with an average age of *37.41*, indicating *age may influence churn.*

Exited customers have a slightly lower average credit score (*645.35*) compared to non-exited customers (*651.85*), suggesting a *minor correlation* between credit score and churn.





Exited customers have average account balance (*91,108.54*) compared to non-exited customers (*72,745.30*), indicating that customers with larger balances are more likely to churn.

Exited customers have a slightly lower average number of products (*1.48*) compared to non-exited customers (*1.54*), suggesting *that having more products may reduce the likelihood of churn*.

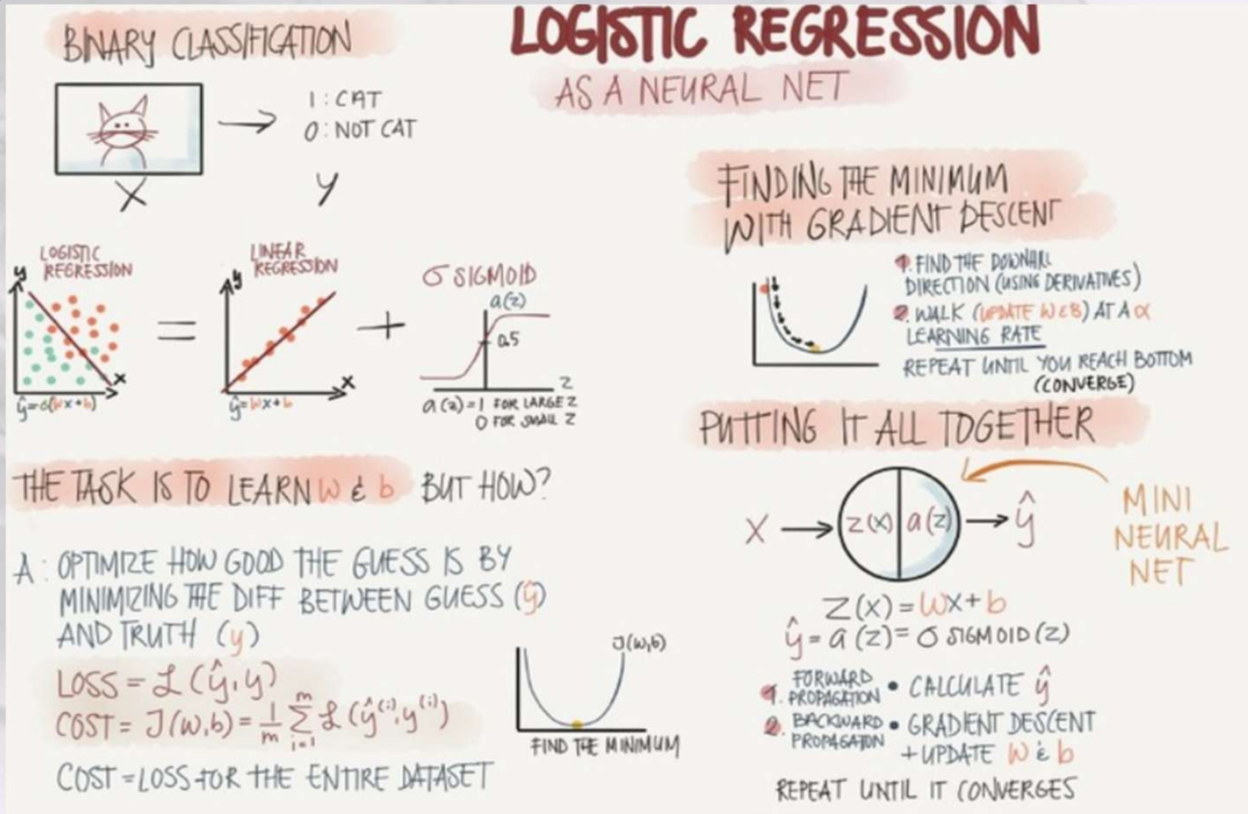
Modelling & Performance

Model Explanation

Logistic regression is a binary classification method.

It can be modelled as a function that can take in any number of inputs and constrain the output to be between *0 and 1*.

This means, we can think of *Logistic Regression* as a *one-layer neural network*.



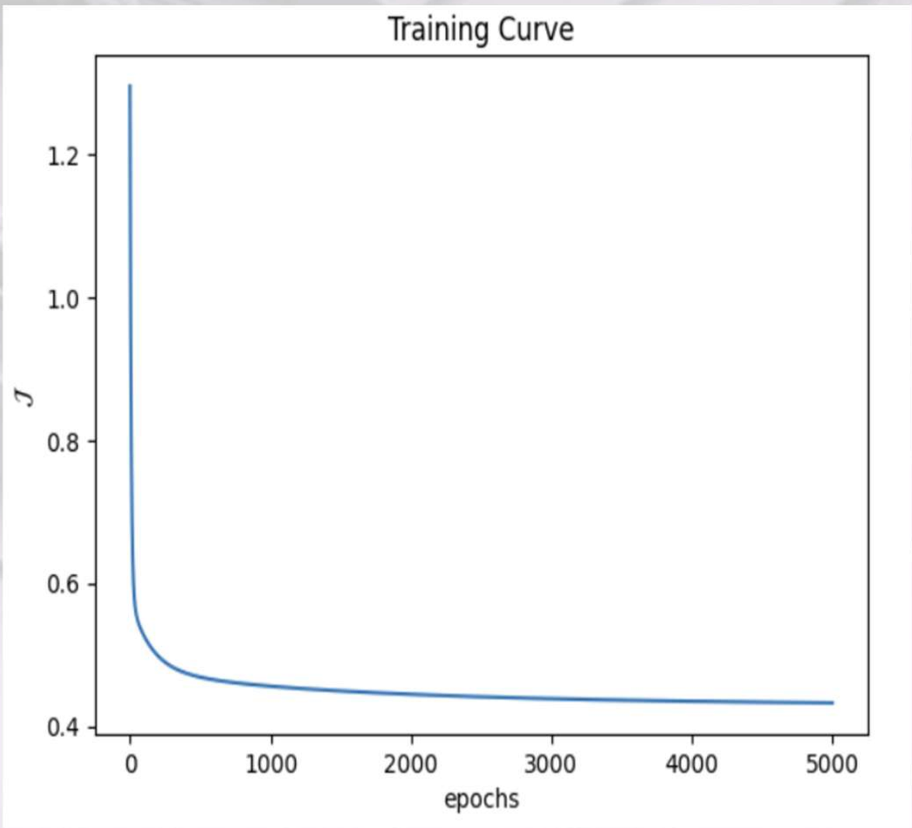
Modelling & Performance

Logistic Regression

Train Ratio: 80%

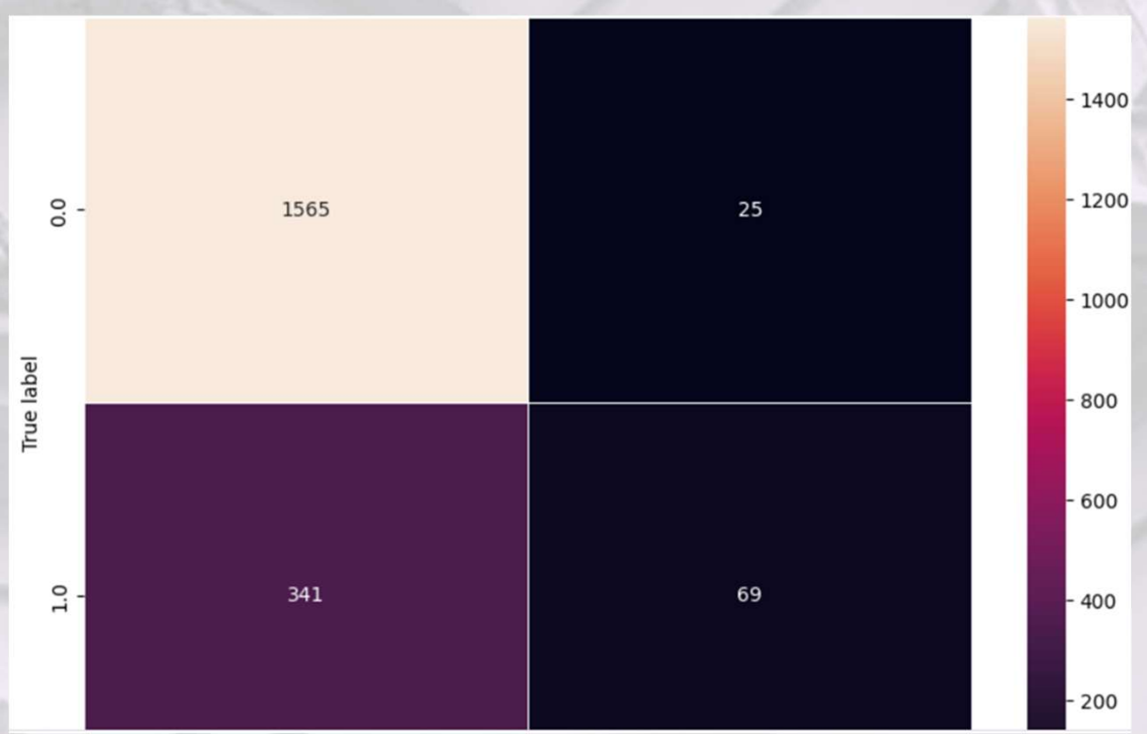
The *training curve* shows a *rapid decrease* in loss during the *initial epochs*, followed by a gradual convergence around *0.4*, indicating effective learning and model stability over time.

Training Accuracy: 0.81
Test Accuracy: 0.81



Logistic Regression

The confusion matrix shows high accuracy in predicting stayed customers (*1,565 true negatives*), but a significant number of exited customers (*341*) are misclassified as stayed.



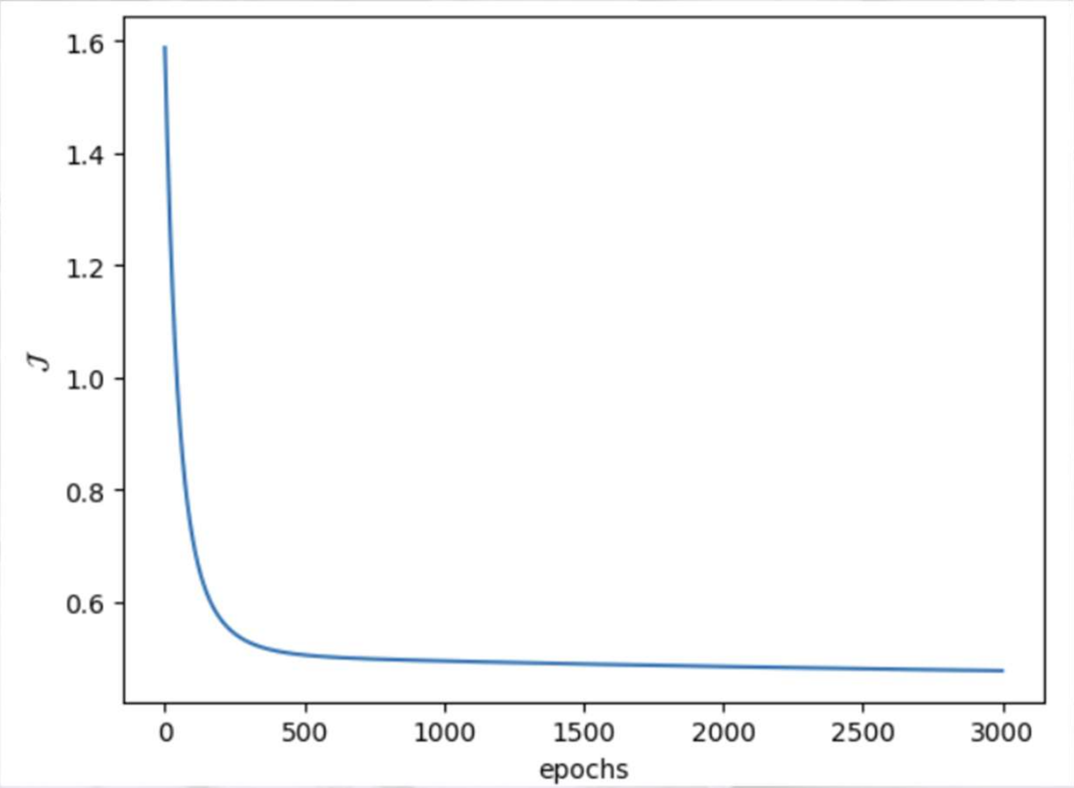
Training Accuracy: 0.81
Test Accuracy: 0.81

Artificial Neural Network

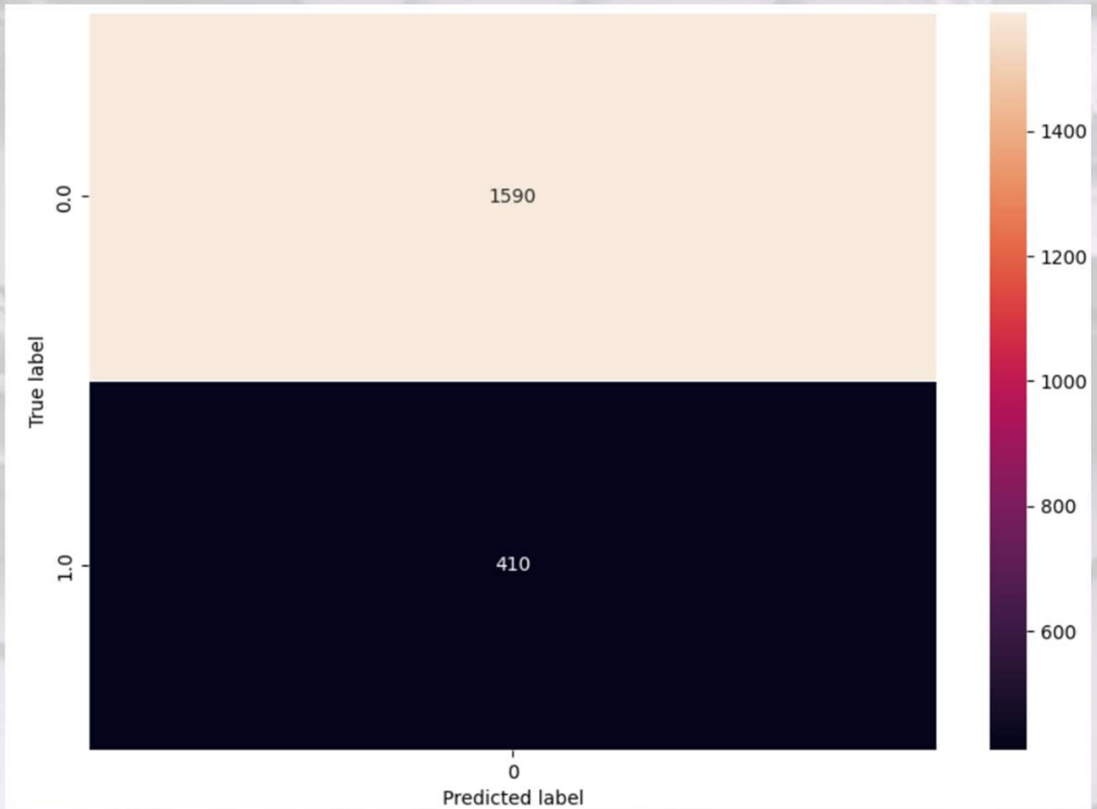
Train Ratio: 80%

The *training curve* shows a *rapid decrease* in loss during the *initial epochs*, followed by a gradual convergence around *0.6*, indicating effective learning and model stability over time.

Training Accuracy: 0.79
Test Accuracy: 0.79



Artificial Neural Network



The confusion matrix shows high accuracy in predicting stayed customers (*1,590 true negatives*), but a significant number of exited customers (*410*) are misclassified as stayed.

Training Accuracy: 0.79
Test Accuracy: 0.79

Summary

The Logistic Regression and ANN algorithms predict churn with an accuracy of 81% and 79%.

However, these are not sufficient for effective customer retention as it is very sensitive in the banking sector.

Therefore, I aim to develop advanced algorithms to build a model with higher accuracy.