Bank Churn Prediction

Turning Data into Retention Strategy

Presenter : Nirbhay Sharma
PG Applied Data Science and AI
E&ICT IIT Roorkee
Jan 25

Problem Statement:
Why Are Customers Leaving?

Despite stable account growth, a significant portion of customers have been **closing their accounts** leading to revenue leakage and reduced customer lifetime value.

Challenge:
The bank lacks a clear understanding of **why customers leave** — whether it's due to demographic, financial, or behavioral factors.

Key Factors Under Consideration:

- **Customer Profile:** Age, Gender, Geography, Tenure
- **Financial Health:** Credit Score, Balance, Estimated Salary
- **Engagement Indicators:** Number of Products, Card Ownership, Activity Level
- **Behavioral Trends:** Long tenure yet inactive members, high balance but low engagement

Goal:
Identify patterns and root causes behind churn to **enable early intervention** and design **targeted retention strategies** for at-risk customers.

Project Overview

Develop an AI-powered churn prediction model to identify customers at risk of leaving within the next 6 months.

Business Impact:

- Enables proactive retention strategies to reduce potential revenue loss.
- Strengthens customer loyalty and protects long-term profitability.
- Supports the bank's strategic objective of maximizing **Customer Lifetime Value (CLV)**.

Strategic Alignment:
This initiative directly contributes to improving **profitability**, **market share**, and **customer satisfaction** through data-driven decision-making.

Resources

Data:

The dataset contains **10,000 customer records** from a retail bank, capturing key demographic, financial, and behavioral attributes.

It serves as the foundation for understanding factors influencing customer churn.

Link

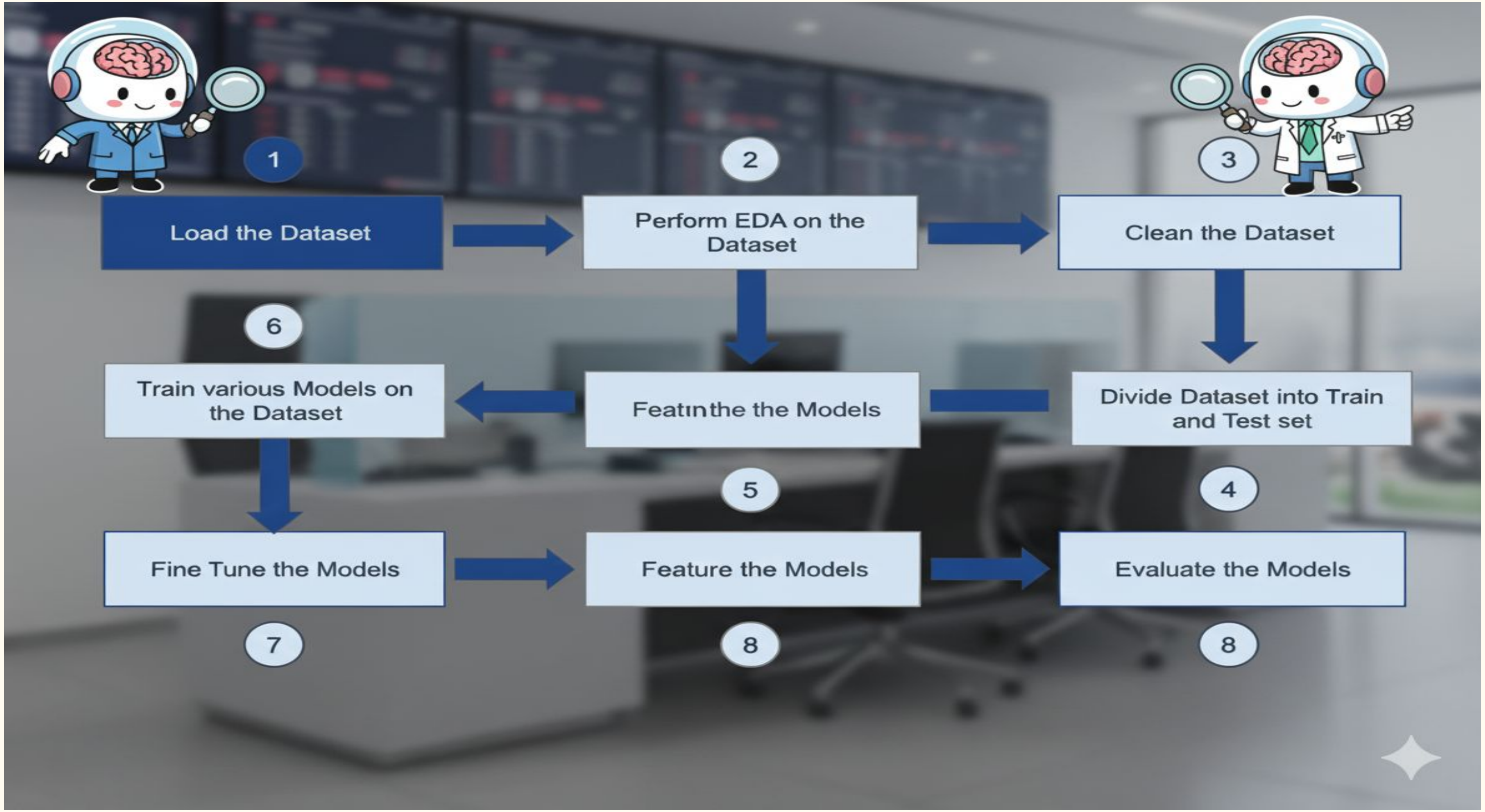
Code:

The code implements the complete **Bank Churn Prediction pipeline** — from data preprocessing and visualization to model training and evaluation.

It provides a reproducible workflow for building and analyzing the AI churn prediction model.

Link

Process


1
Load the Dataset
2
Perform EDA on the Dataset
3
Clean the Dataset
6
Train various Models on the Dataset
Featınthe the Models
Divide Dataset into Train and Test set
5
4
Fine Tune the Models
Feature the Models
Evaluate the Models
7
8
8

Data Insight

- CustomerId: Unique ID which is assigned to each customer
- Surname: Last name of the customer
- CreditScore: It defines the credit history of the customer.
- Geography: A customer's location
- Gender: It defines the Gender of the customer
- Age: Age of the customer
- Tenure: Number of years for which the customer has been with the bank
- NumOfProducts: refers to the number of products that a customer has purchased through the bank.
- Balance: Account balance
- HasCrCard: It is a categorical variable which decides whether the customer has credit card or not.
- EstimatedSalary: Estimated salary
- isActiveMember: Is is a categorical variable which decides whether the customer is active member of the bank or not (Active member in the sense, using bank products regularly, making transactions etc)
- Exited : whether or not the customer left the bank within six month. It can take two values ** 0=No (Customer did not leave the bank) ** 1=Yes (Customer left the bank)

```
df.head()
```

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

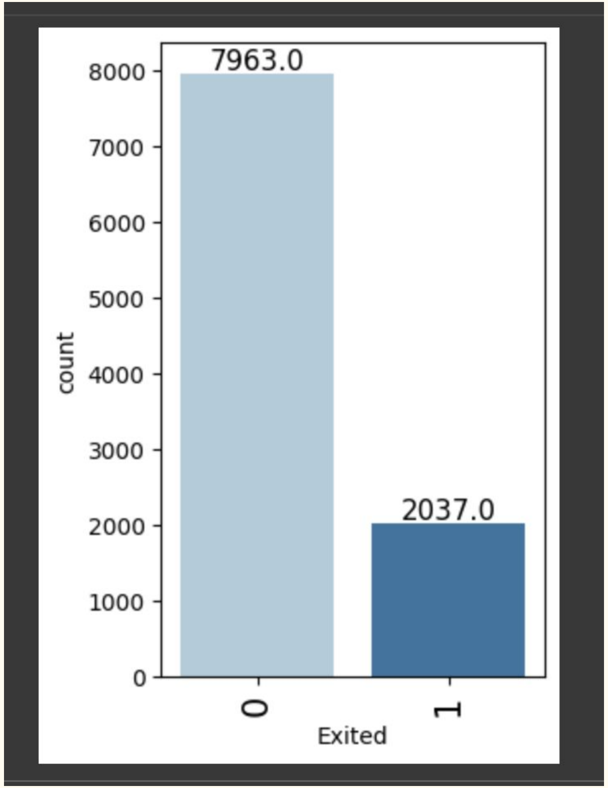
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

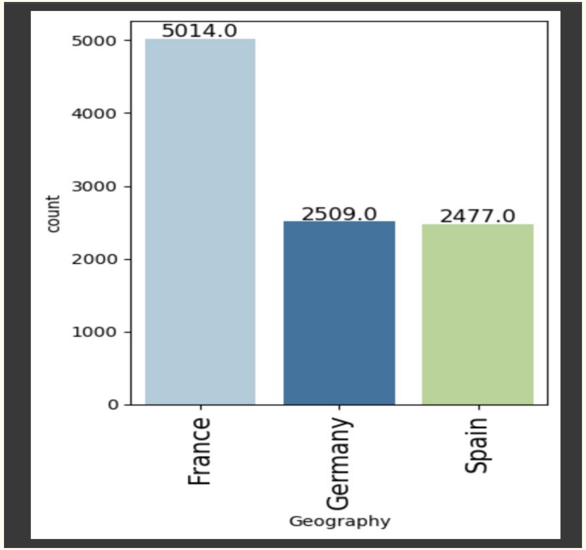
```
df.describe().T
```

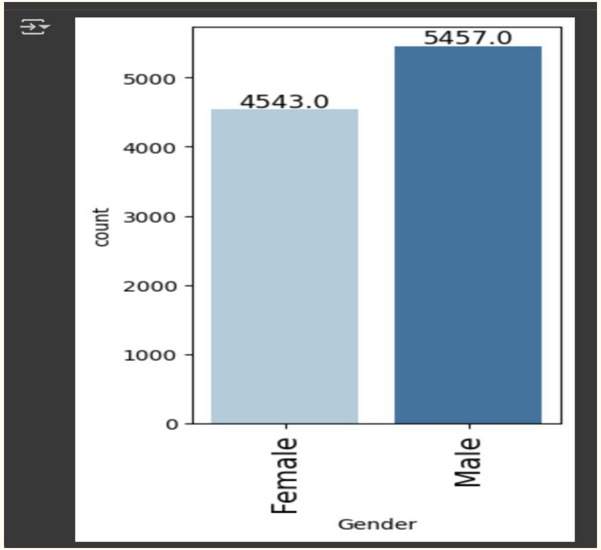
	count	mean	std	min	25%	50%	75%	max
RowNumber	10000.0	5.000500e+03	2886.895680	1.00	2500.75	5.000500e+03	7.500250e+03	10000.00
CustomerId	10000.0	1.569094e+07	71936.186123	15565701.00	15628528.25	1.569074e+07	1.575323e+07	15815690.00
CreditScore	10000.0	6.505288e+02	96.653299	350.00	584.00	6.520000e+02	7.180000e+02	850.00
Age	10000.0	3.892180e+01	10.487806	18.00	32.00	3.700000e+01	4.400000e+01	92.00
Tenure	10000.0	5.012800e+00	2.892174	0.00	3.00	5.000000e+00	7.000000e+00	10.00
Balance	10000.0	7.648589e+04	62397.405202	0.00	0.00	9.719854e+04	1.276442e+05	250898.09
NumOfProducts	10000.0	1.530200e+00	0.581654	1.00	1.00	1.000000e+00	2.000000e+00	4.00
HasCrCard	10000.0	7.055000e-01	0.455840	0.00	0.00	1.000000e+00	1.000000e+00	1.00
IsActiveMember	10000.0	5.151000e-01	0.499797	0.00	0.00	1.000000e+00	1.000000e+00	1.00
EstimatedSalary	10000.0	1.000902e+05	57510.492818	11.58	51002.11	1.001939e+05	1.493882e+05	199992.48
Exited	10000.0	2.037000e-01	0.402769	0.00	0.00	0.000000e+00	0.000000e+00	1.00

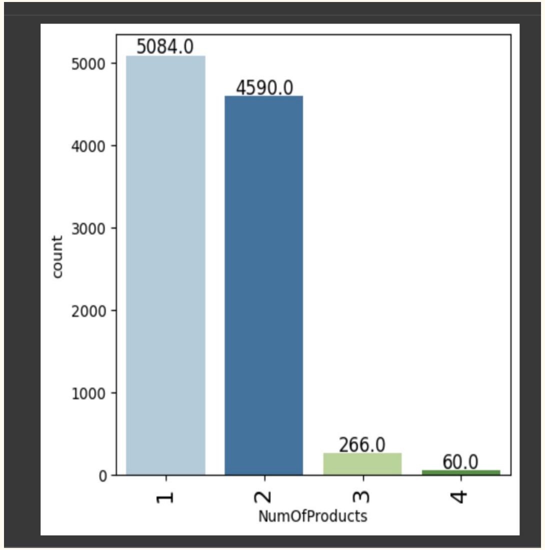
Data Findings

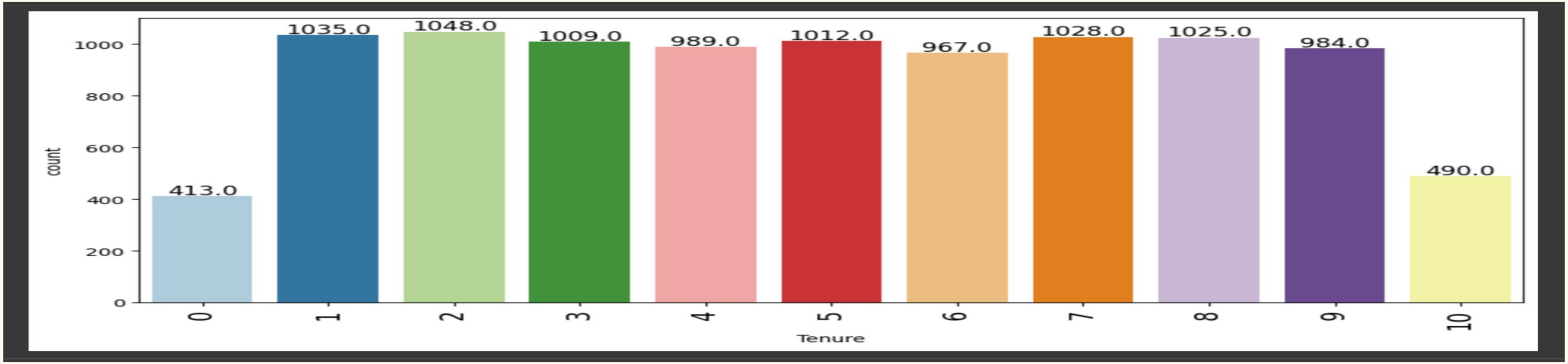


Exited	proportion
0	79.63
1	20.37

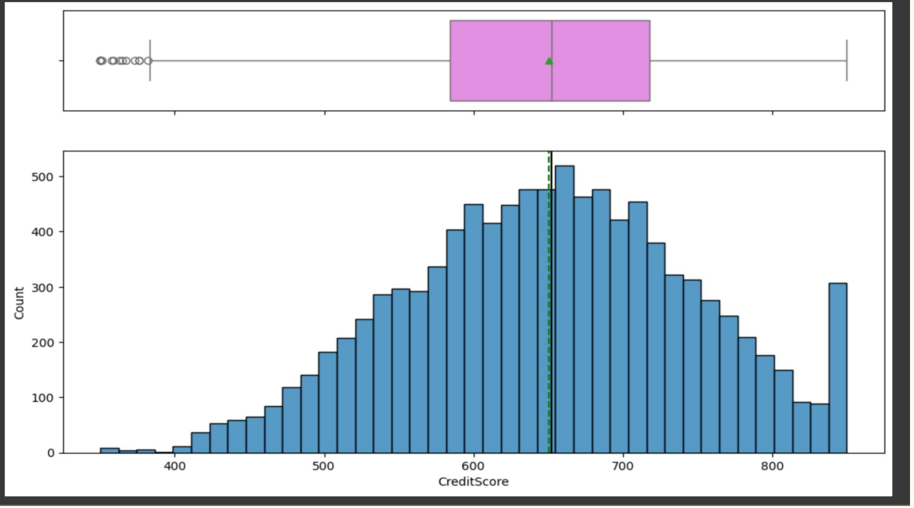

5014.0
2509.0
2477.0
count
France
Germany
Spain
Geography


4543.0
5457.0
count
Female
Male
Gender

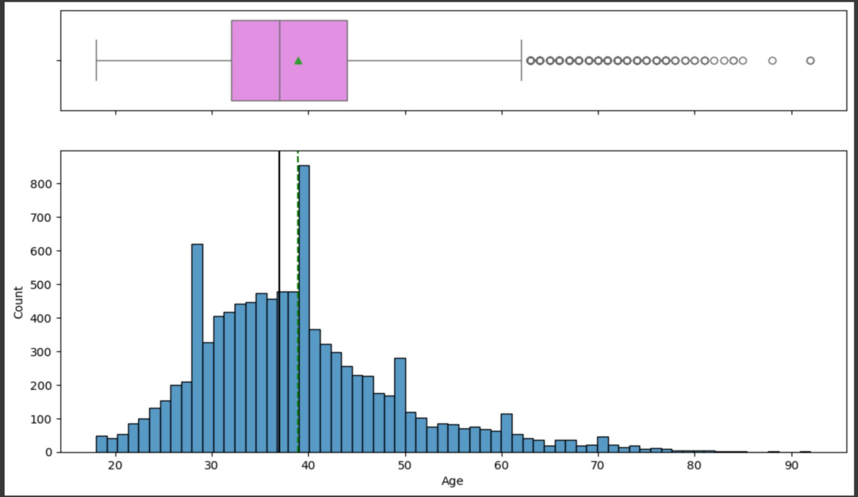

5084.0
4590.0
266.0
60.0
count
NumOfProducts


413.0
1035.0
1048.0
1009.0
989.0
1012.0
967.0
1028.0
1025.0
984.0
490.0
count
Tenure

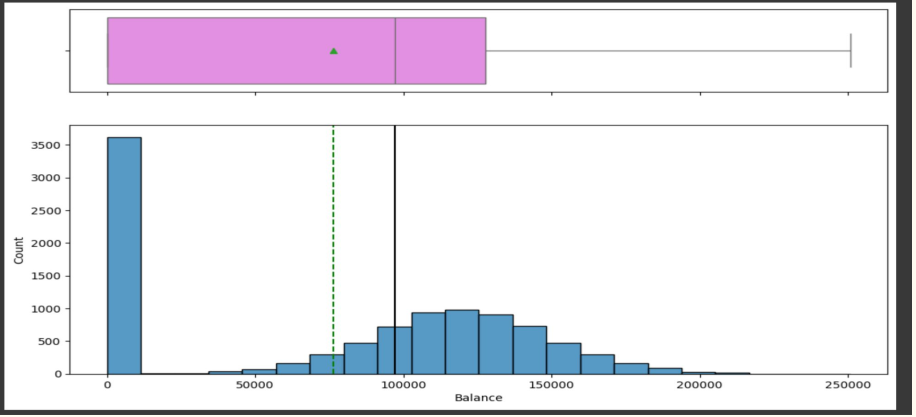
```
histogram_boxplot(df, "CreditScore")
```



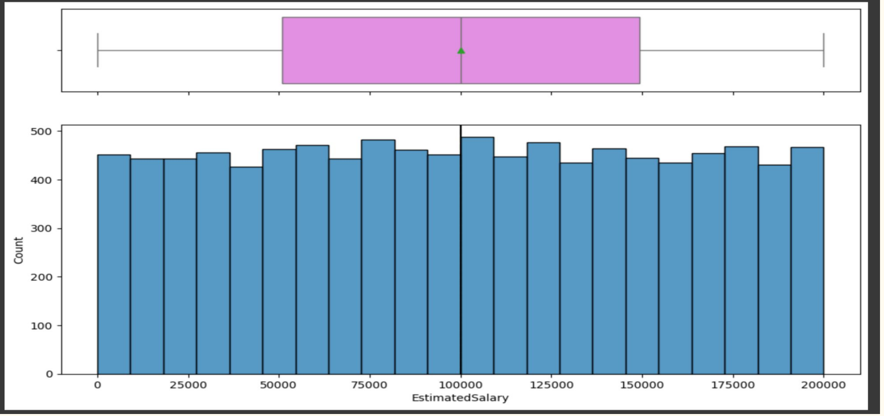

```
histogram_boxplot(df, "Age")
```

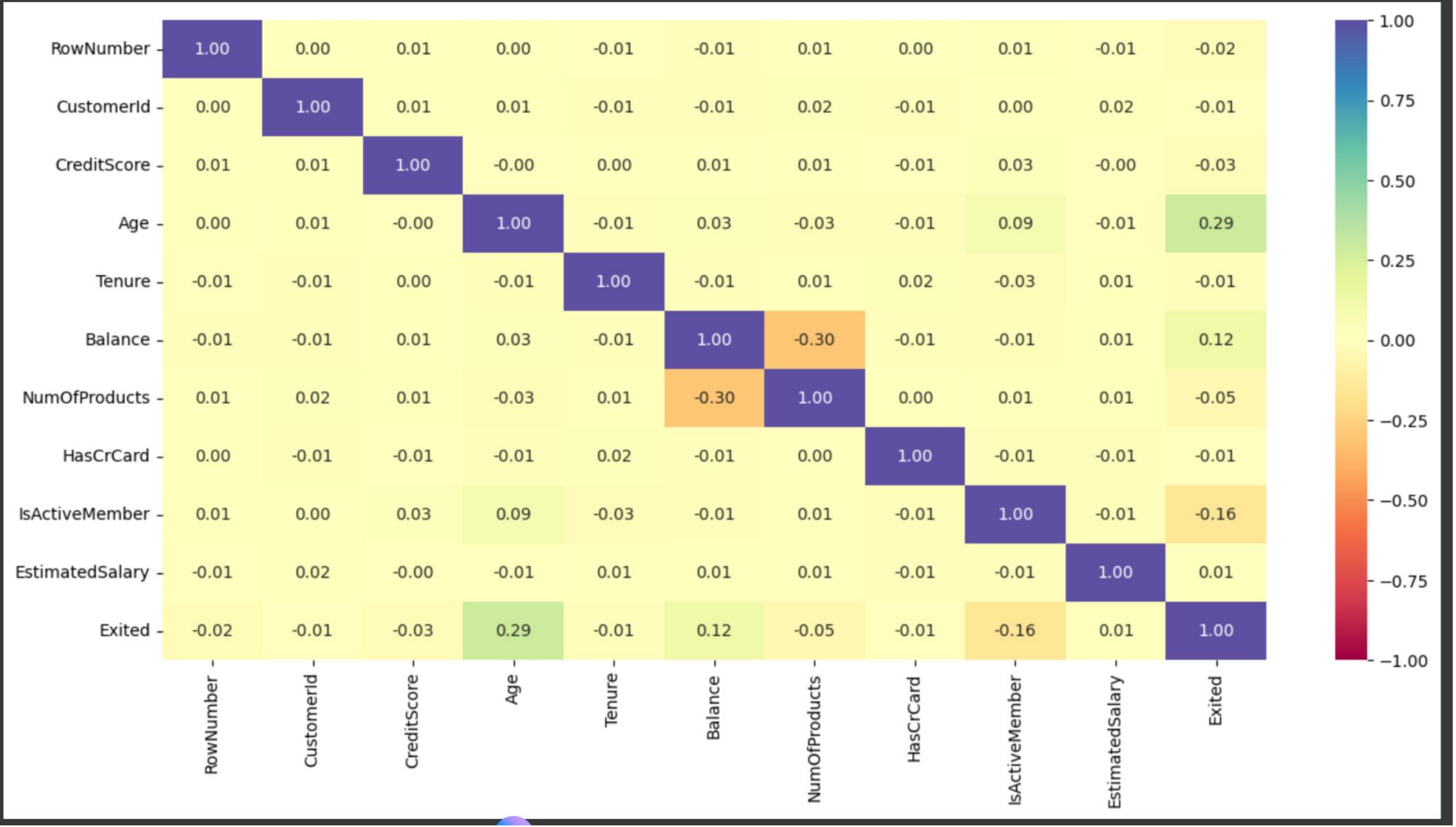



```
histogram_boxplot(df, "Balance")
```

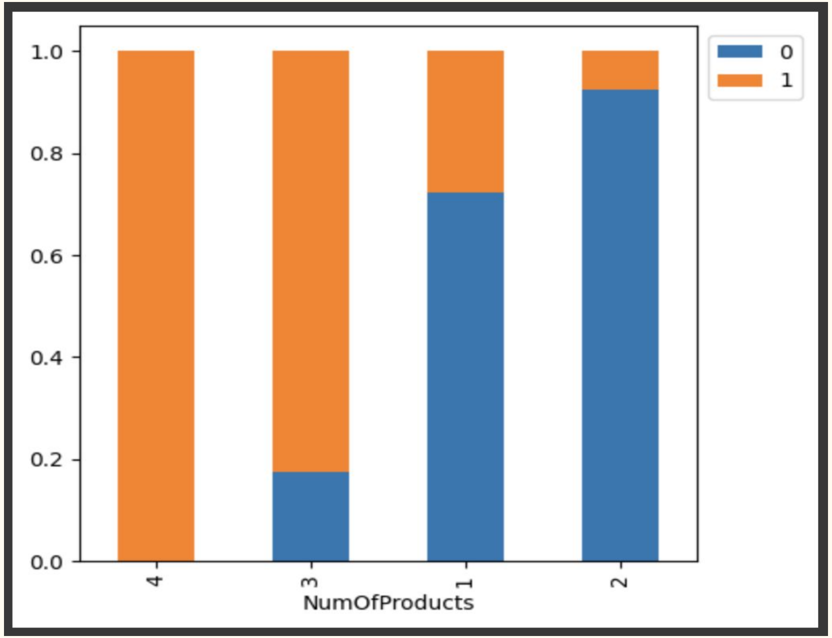



```
histogram_boxplot(df, 'EstimatedSalary')
```

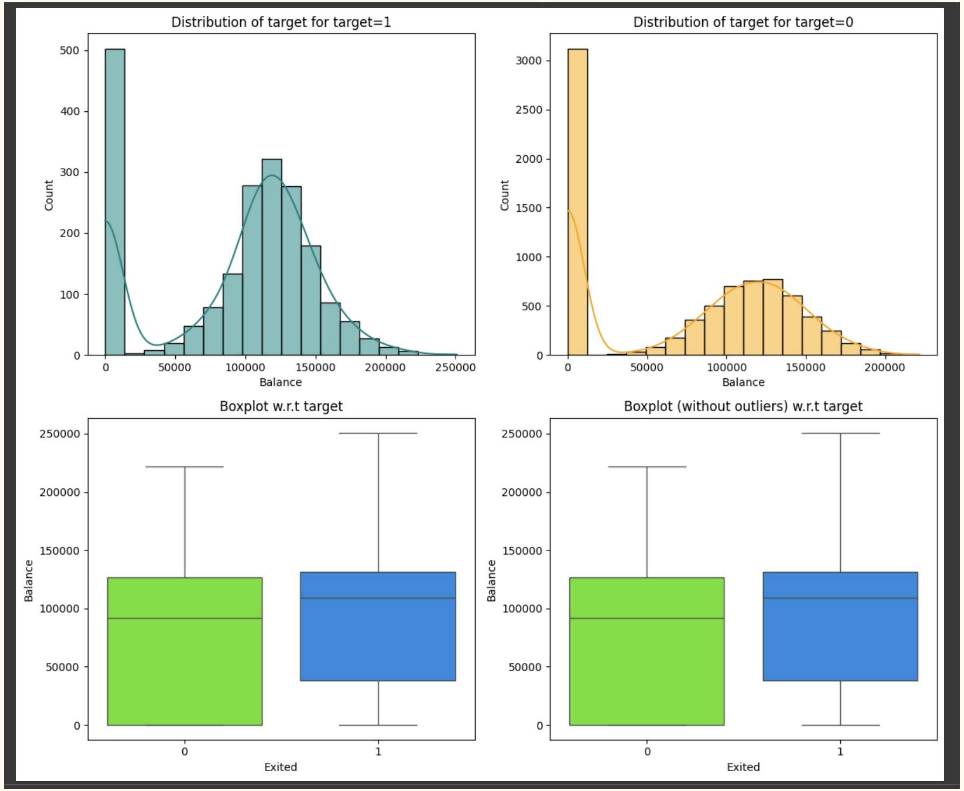




RowNumber 1.00 0.00 0.01 0.00 -0.01 -0.01 0.01 0.00 0.01 -0.01 -0.02
CustomerId 0.00 1.00 0.01 0.01 -0.01 -0.01 0.02 -0.01 0.00 0.02 -0.01
CreditScore 0.01 0.01 1.00 -0.00 0.00 0.01 0.01 -0.01 0.03 -0.00 -0.03
Age 0.00 0.01 -0.00 1.00 -0.01 0.03 -0.03 -0.01 0.09 -0.01 0.29
Tenure -0.01 -0.01 0.00 -0.01 1.00 -0.01 0.01 0.02 -0.03 0.01 -0.01
Balance -0.01 -0.01 0.01 0.03 -0.01 1.00 -0.30 -0.01 -0.01 0.01 0.12
NumOfProducts 0.01 0.02 0.01 -0.03 0.01 -0.30 1.00 0.00 0.01 0.01 -0.05
HasCrCard 0.00 -0.01 -0.01 -0.01 0.02 -0.01 0.00 1.00 -0.01 -0.01 -0.01
IsActiveMember 0.01 0.00 0.03 0.09 -0.03 -0.01 0.01 -0.01 1.00 -0.01 -0.16
EstimatedSalary -0.01 0.02 -0.00 -0.01 0.01 0.01 0.01 -0.01 -0.01 1.00 0.01
Exited -0.02 -0.01 -0.03 0.29 -0.01 0.12 -0.05 -0.01 -0.16 0.01 1.00
RowNumber
CustomerId
CreditScore
Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
Exited
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00



1.0
0.8
0.6
0.4
0.2
0.0
0
1
4
3
1
2
NumOfProducts



Distribution of target for target=1
Count
Balance
Distribution of target for target=0
Count
Balance
Boxplot w.r.t target
Balance
Exited
Boxplot (without outliers) w.r.t target
Balance
Exited

Data Preprocessing

- **Missing value treatment**
- **Feature engineering**
- **Removing Unnecessary data**
- **Outlier detection and treatment**
- **Preparing data for modeling**
- **Normalization and any other preprocessing steps**

Data Preprocessing

- **Missing value treatment**
- **Feature engineering**
- **Outlier detection and treatment**
- **Preparing data for modeling**
- **Any other preprocessing steps**

Model Building

Feedforward Neural Network

A type of artificial neural network where information moves **in one direction**: from input → hidden → output layers.

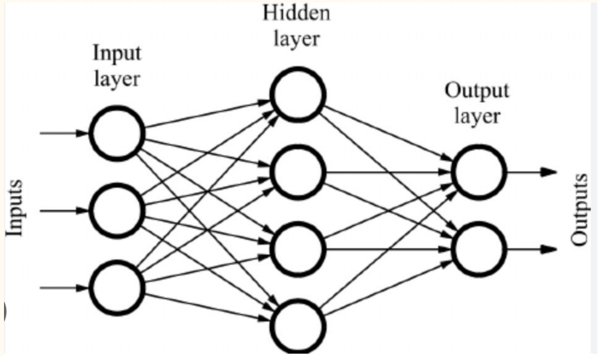
No cycles or loops; outputs of one layer feed only into the next layer.

Consists of:

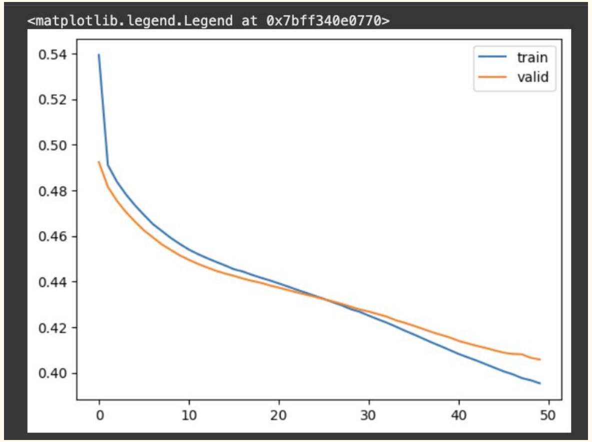
- **Input layer:** Receives features from the dataset.
- **Hidden layers:** Perform computations using activation functions (e.g., ReLU, Sigmoid).
- **Output layer:** Produces the final prediction or classification.

Learning: Uses **backpropagation** and gradient descent to update weights.

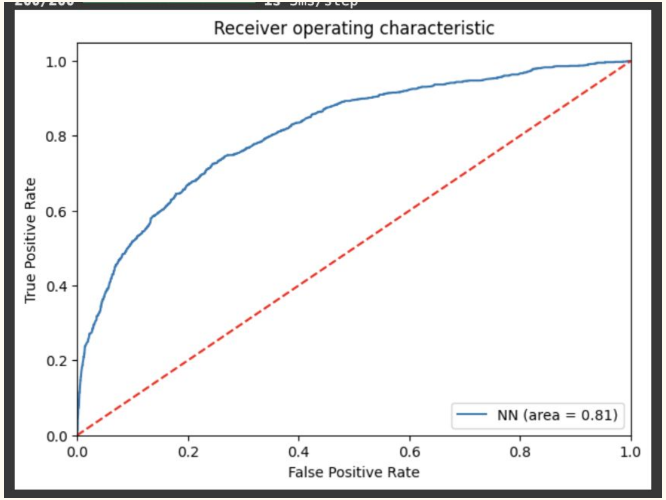
Commonly used for: regression, classification, and basic pattern recognition tasks.



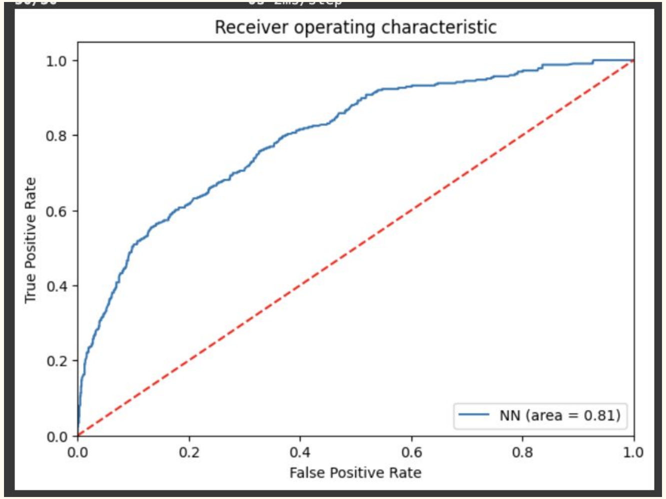
Neural Network with SGD Optimizer Accuracy



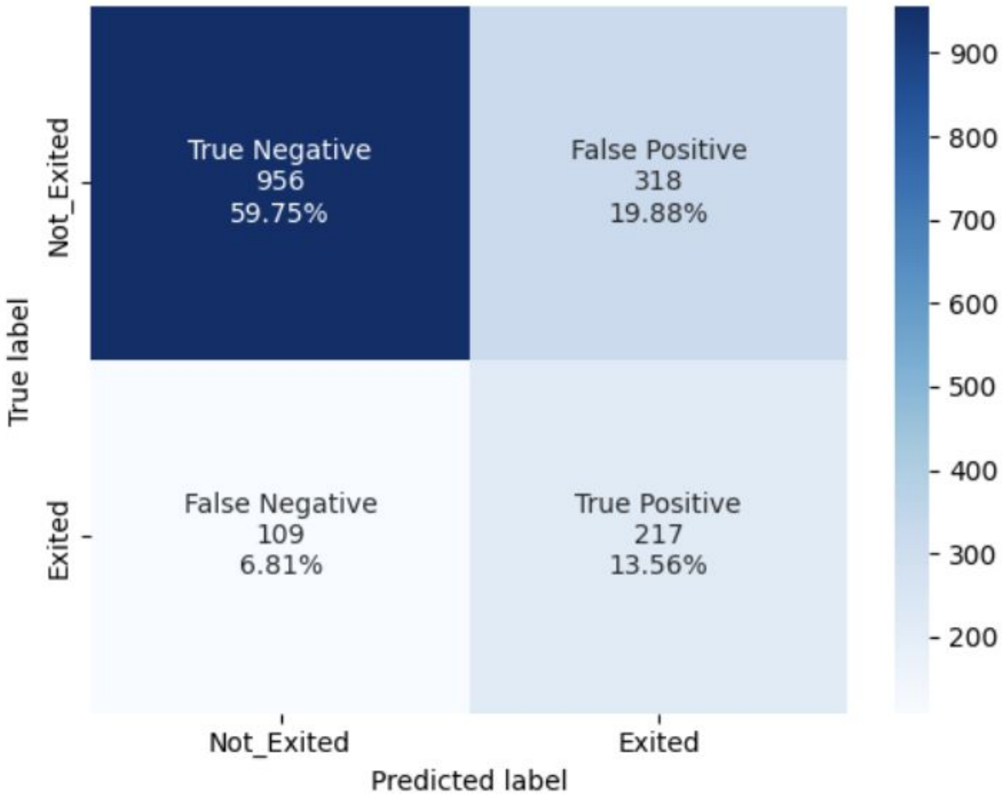
Train Data



Validation Data

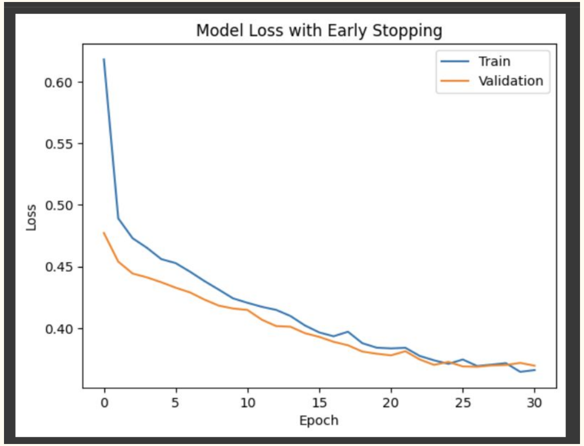


50/50 0s 1ms/step

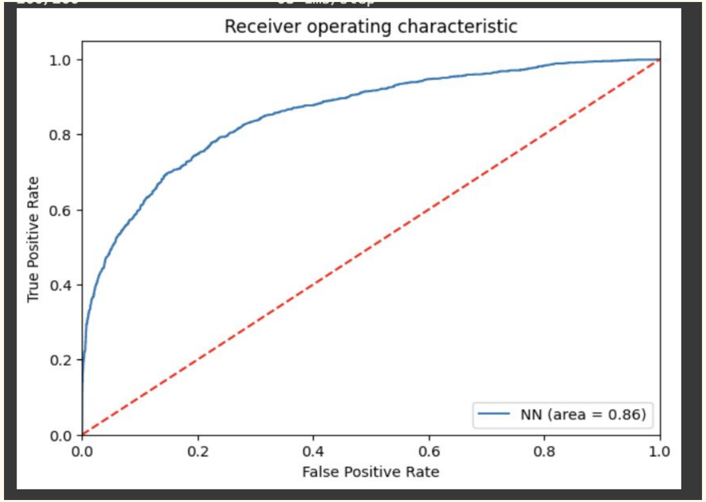


Accuracy=0.733
Precision=0.406
Recall=0.666
F1 Score=0.504

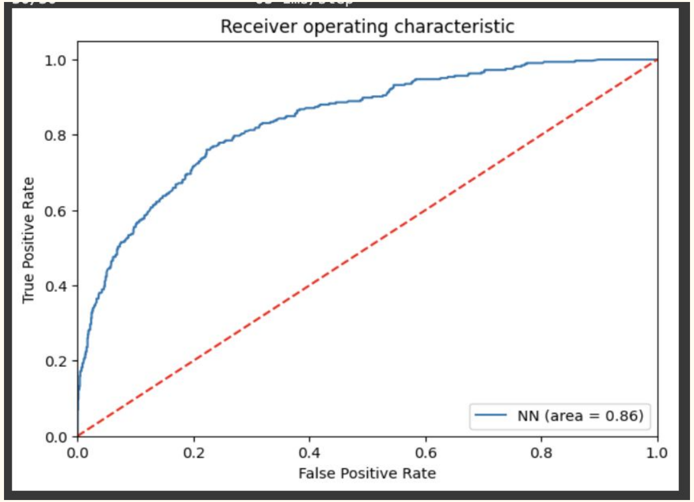
Neural Network with Adam optimizer and Dropout



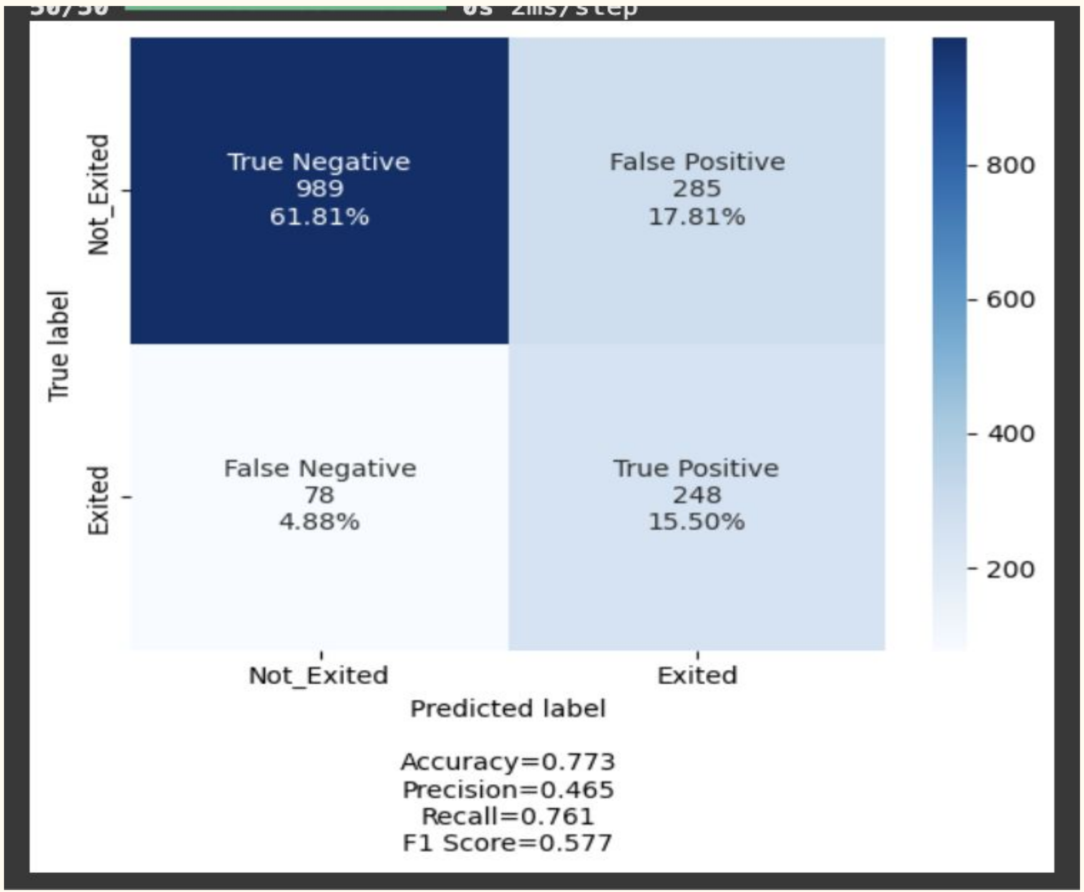
Train Data



Validation Data



Neural Network with Adam optimizer and Dropout



Model Performance Comparison and Final Model Selection

```
Final Metrics::

****************************************************************************************************
1. SGD Optimizer
   Accuracy=0.733       Precision=0.406 Recall=0.666    F1 Score=0.504

2. Adam Optimizer
  Accuracy=0.767        Precision=0.454 Recall=0.699    F1 Score=0.551

3. Adam Optimizer and DropOut
  Accuracy=0.773        Precision=0.465 Recall=0.761    F1 Score=0.577

4. SMOTE and SGD Optimizer
   Accuracy=0.757       Precision=0.438 Recall=0.687    F1 Score=0.535

5. SMOTE and Adam Optimizer
  Accuracy=0.757        Precision=0.435 Recall=0.644    F1 Score=0.519

6. SMOTE, Adam Optimizer and Dropout
  Accuracy=0.764        Precision=0.449 Recall=0.706    F1 Score=0.549

****************************************************************************************************
```

Findings & Recommendations

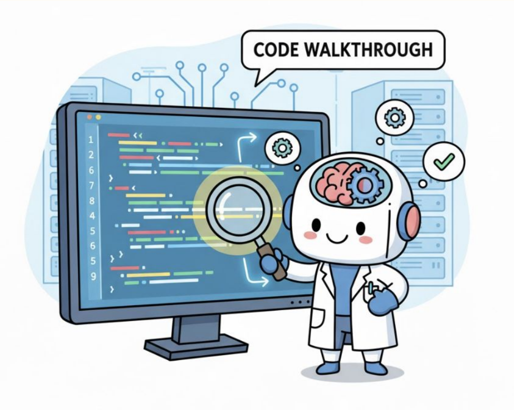
Engagement of dormant Members: The bank might launch a campaign to turn dormant members into active clients. To help these clients get the most out of their accounts, this can entail contacting them with exclusive deals or incentives or offering them individualized financial guidance.

Product Retention and Diversification: There is a chance to encourage customers to diversify their product holdings because the minority of customers who only own one product (51%). Create retention techniques to hold on to clients that possess numerous items, like incentives or packaged services.

Services That Consider Age: Given that quitting a bank is positively correlated with age, you should think about offering age-specific services or incentives to keep customers in particular age ranges. Customizing services to fit various stages of life could increase client retention.

Retention based on Tenure: Higher rates of customer churn are seen among those with shorter tenures—one year and zero years. Use promotions, individualized services, or onboarding programs to win over more customers throughout the early years of their bank relationship.

Code Walkthrough



Q & A

Thank You