Bank Churn Problem

Introduction to Neural Networks

PGAIML UT Austin
McCombs School of Business

October 13, 2023

Presented by Greg Wenzel

Contents / Agenda

Executive Summary

- Problem Statement
- Data
- Data Preprocessing
- Model Building and Hyperparameter Tuning
- Model Evaluation
- Conclusion

Executive Summary I

- Problem Statement

- The problem at hand is to address customer churn in a bank, where customers are leaving for other service providers. Understanding the factors influencing customer decisions is crucial for retaining them. The goal is to build a neural network-based classifier that predicts whether a customer will leave the bank in the next six months.

- Data

- The dataset consists of various features, including customer ID, credit score, geography, gender, age, tenure, balance, number of products, presence of a credit card, active member status, estimated salary, and the target variable, "Exited."

- Data Preprocessing

- The initial data exploration revealed no missing values or duplicated entries. Unique values were examined in each column, and the data was described statistically. Columns with unique or irrelevant information, such as CustomerId, Surname, and RowNumber, were dropped. Categorical features like Geography and Gender were one-hot encoded to prepare them for modeling. Numerical features were standardized to ensure that they all have the same scale.

Executive Summary II

- Model Building and Hyperparameter Tuning

A neural network model was constructed using TensorFlow and Keras. It included input, hidden, and output layers with dropout layers to prevent overfitting. The model's hyperparameters were optimized using GridSearchCV, considering batch size and learning rate.

- Model Improvement: Balancing Data

To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data, increasing the minority class's representation.

- Model Evaluation

The best model was selected based on hyperparameter tuning and was tested on the validation set. Receiver Operating Characteristic (ROC) analysis was conducted to determine the optimal threshold for classification. The results were evaluated using classification metrics, including precision, recall, F1-score, and confusion matrices.

- Conclusion

The final model achieved a recall of 0.770 for the "Exited" class, which indicates that it correctly identifies 77% of the customers who actually left the bank. Although the precision for this class is relatively low (0.457), the model's overall accuracy and F1-score are satisfactory. Further improvements in precision may be needed to reduce false positives. The model provides valuable insights into predicting customer churn for the bank. To improve performance further, additional feature engineering or trying different machine learning algorithms could be explored.

Business Problem Overview and Solution Approach

Problem Definition

The problem at hand is customer churn prediction for a bank. Customer churn refers to customers leaving the bank for alternative service providers. Understanding the factors that influence customer decisions to leave is essential for retaining them and maintaining business stability.

Solution Approach I

1. Data Collection: The first step is gathering the necessary data, which includes customer information such as credit scores, demographics, account balance, and transaction history. Additionally, historical data on whether customers have churned (1) or not (0) within the past six months is required.

2. Data Preprocessing: Once collected, the data should undergo preprocessing. This involves handling missing values, removing duplicates, and encoding categorical variables. Irrelevant columns, such as customer ID, can be dropped. It's essential to ensure that the data is in a clean and usable format for analysis.

3. Exploratory Data Analysis (EDA): EDA involves exploring the data to gain insights. Summary statistics, data distributions, and visualizations can help in understanding the data's characteristics. It's important to identify patterns, correlations, and potential features that might influence customer churn.

4. Feature Engineering: Feature engineering involves creating new features or transforming existing ones to improve the model's predictive power. For example, one could calculate customer tenure, create binary variables for active membership or credit card possession, and engineer interaction features if they are meaningful.

5. Data Splitting: The dataset should be divided into training, validation, and test sets. Typically, 70-80% of the data is used for training, 10-15% for validation, and the remaining 10-15% for testing.

Business Problem Overview and Solution Approach

Problem Definition

The problem at hand is customer churn prediction for a bank. Customer churn refers to customers leaving the bank for alternative service providers. Understanding the factors that influence customer decisions to leave is essential for retaining them and maintaining business stability.

Solution Approach II

6. Model Selection: Various machine learning models can be employed for customer churn prediction, including logistic regression, decision trees, random forests, support vector machines, and neural networks. The choice of model depends on the dataset's characteristics and the problem complexity. Neural networks, like the one mentioned in the previous conversation, can capture intricate patterns but may require more data and computational resources.

7. Model Training: The selected model is trained using the training dataset. During this phase, hyperparameters can be tuned to optimize the model's performance.

8. Model Evaluation: The model's performance is assessed using the validation dataset. Common evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC. These metrics help in understanding how well the model is at identifying customers who churn and customers who don't.

9. Model Fine-tuning: If the model's performance is unsatisfactory, further fine-tuning can be performed. This might involve adjusting hyperparameters, trying different algorithms, or engineering additional features.

Business Problem Overview and Solution Approach

Problem Definition

The problem at hand is customer churn prediction for a bank. Customer churn refers to customers leaving the bank for alternative service providers. Understanding the factors that influence customer decisions to leave is essential for retaining them and maintaining business stability.

Solution Approach III

10. Model Testing: Once the model performs well on the validation set, it's tested on the independent test set to ensure its generalization to unseen data.

11. Deployment: Once the model meets the desired performance criteria, it can be deployed in a production environment where it can be used to predict customer churn in real-time.

12. Continuous Monitoring: Customer churn is an evolving problem, and models need to be periodically re-evaluated and retrained as new data becomes available. Continuous monitoring allows for maintaining model accuracy and relevance over time.

In summary, the solution approach involves a series of steps, from data collection and preprocessing to model selection, training, and deployment. The key is to iteratively refine the model to achieve the best possible performance in predicting customer churn. Additionally, continuous monitoring ensures the model remains effective as customer behavior and data patterns evolve.

Exploratory Data Analysis

Univariate Analysis

Bivariate Analysis

Exploratory Data Analysis I

The EDA findings provide essential insights relevant to addressing the defined problem statement, guiding the solution approach, model building, and influencing the final conclusions. The dataset contains information about bank customers, including demographics, credit history, account balance, and churn status.

Dataset Summary:

Number of Rows: 10,000

Beginning Number of Columns : 14

Final Number of Columns: 11 (After dropping irrelevant columns)

Data Types: A combination of integer, float, and object (string) data types.

Description of Columns:

CreditScore: The credit score of the bank's customers. It defines their credit history.

Geography: The location (country) of the customer.

Exploratory Data Analysis I

Description of Columns II

Gender: The gender of the customer (male or female).

Age: The age of the customer.

Tenure: The number of years for which the customer has been with the bank.

Balance: The account balance of the customer.

NumOfProducts: Refers to the number of bank products that a customer has purchased.

HasCrCard: Binary categorical variable indicating whether the customer has a credit card (1 for yes, 0 for no).

IsActiveMember: Binary categorical variable indicating if customer is an active member of the bank (1=yes, 0= no).

EstimatedSalary: The estimated salary of the customer.

Exited: The target variable. It indicates whether the customer left the bank within six months (1 for yes, 0 for no).

Exploratory Data Analysis II – Dataset Description I

This dataset represents customer information for a bank. The goal is to predict customer churn, i.e., whether a customer will leave the bank within the next six months. Here are some key observations and insights:

Numerical Variables: The dataset contains various numerical variables such as "CreditScore," "Age," "Tenure," "Balance," "NumOfProducts," and "EstimatedSalary." These variables provide insights into the customer's financial profile, tenure with the bank, and age.

Categorical Variables: The dataset includes categorical variables like "Geography" and "Gender." These variables provide information about the customer's location and gender.

Binary Variables: "HasCrCard" and "IsActiveMember" are binary categorical variables indicating whether the customer has a credit card and whether they are an active member.

Target Variable: "Exited" is the target variable, which is binary (0 or 1), indicating whether the customer left the bank within six months.

Exploratory Data Analysis II – Dataset Description II

Data Quality: There are no missing values in the dataset, indicating that it is relatively clean.

Data Distribution: Exploratory data analysis (EDA) reveals the distribution of variables, allowing for insights into customer demographics and behaviors. For instance, you can observe the distribution of credit scores, ages, and account balances.

Class Imbalance: The dataset might exhibit class imbalance in the target variable "Exited," which could impact model training and evaluation. Techniques like resampling or using appropriate evaluation metrics may be necessary to address this imbalance.

Data Preprocessing: Data preprocessing steps, such as encoding categorical variables and standardizing numerical features, are essential before building predictive models.

Modeling: The dataset can be used to build predictive models to determine factors influencing customer churn and make predictions regarding whether a customer is likely to leave the bank.

In summary, this dataset provides valuable information for understanding and predicting customer churn in a bank. The goal is to leverage this data to build a predictive model that can help the bank proactively retain customers and improve customer satisfaction.

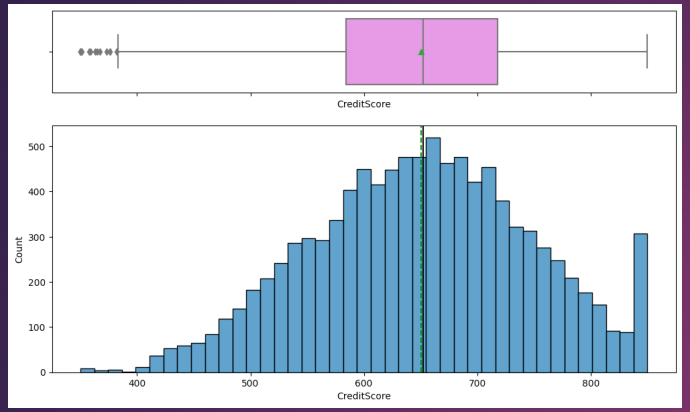
Univariate Analysis

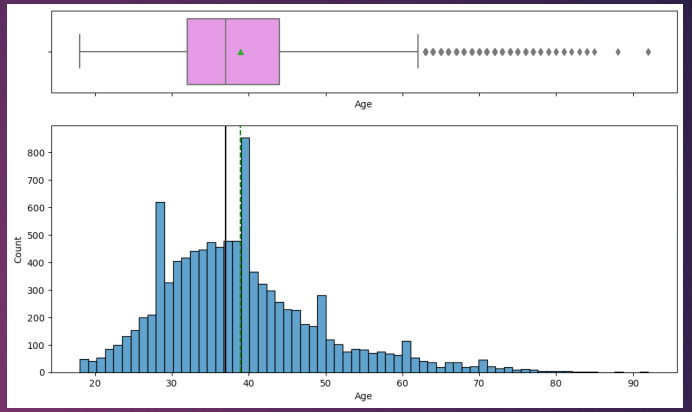
Univariate analysis focuses on examining a single variable at a time. It helps us understand the distribution, central tendency, spread, and characteristics of individual variables in the dataset. Common tools and techniques for univariate analysis include:

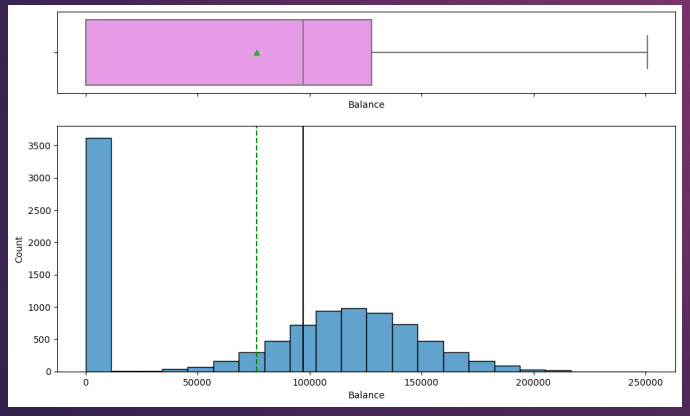
Numerical Variables (e.g., CreditScore, Age, Balance):
The CreditScore appears to follow a relatively normal distribution.
The Age variable shows a relatively even distribution.
The Balance variable exhibits a right-skewed distribution.

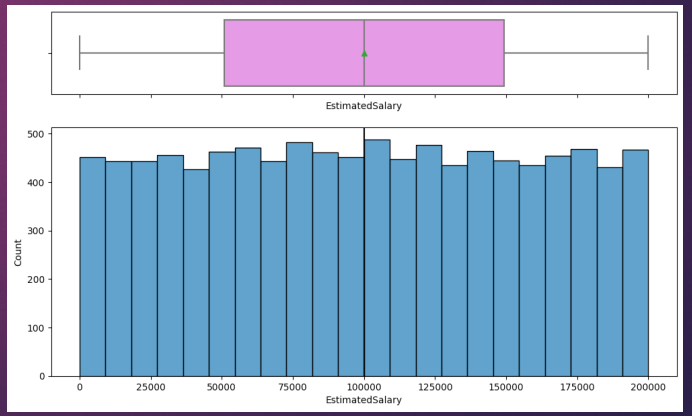
Categorical Variables (e.g., Geography, Gender):
The dataset contains customers from multiple geographical regions, with varying counts.
There's a fairly even distribution between male and female customers.

Univariate Visualization I

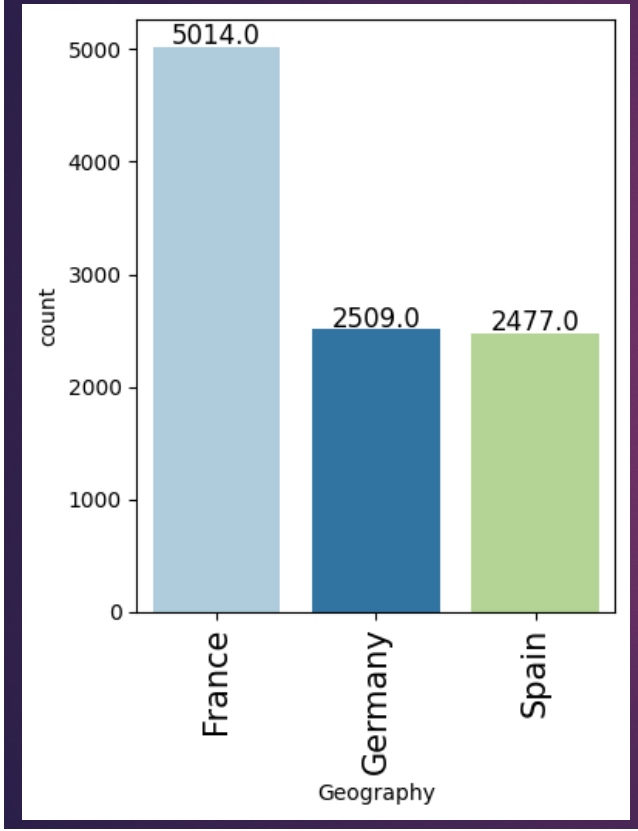


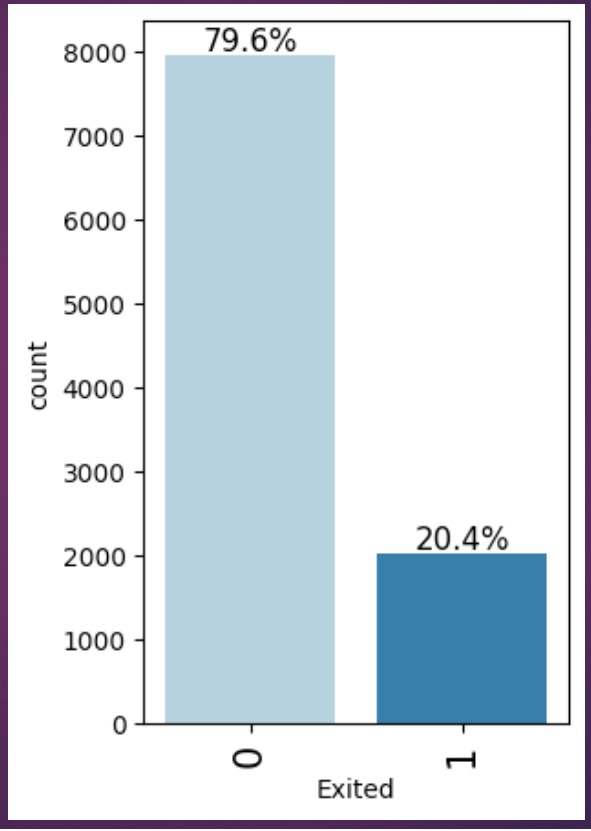


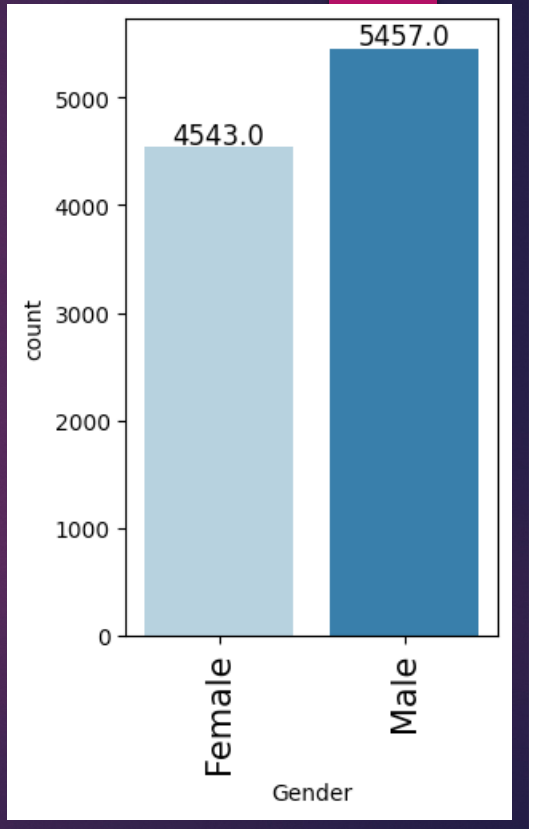




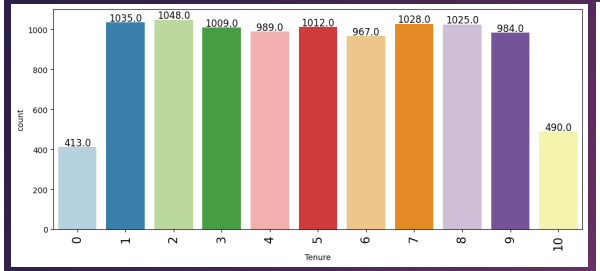
Univariate Visualization II

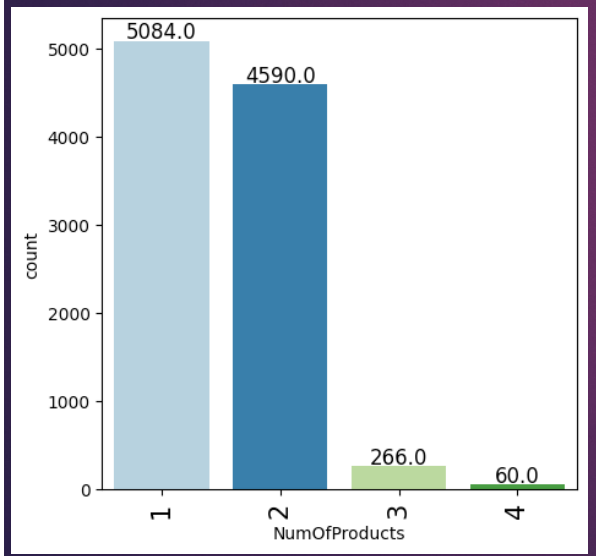


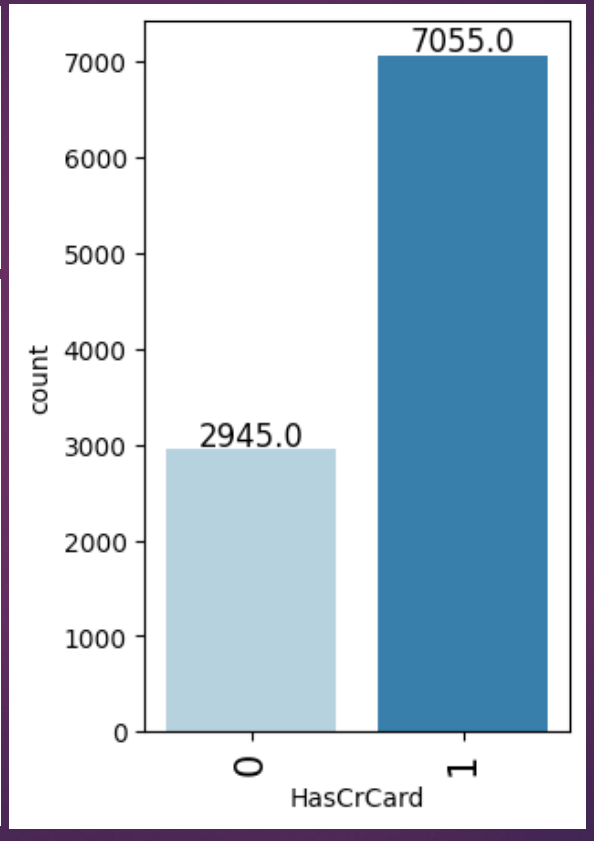


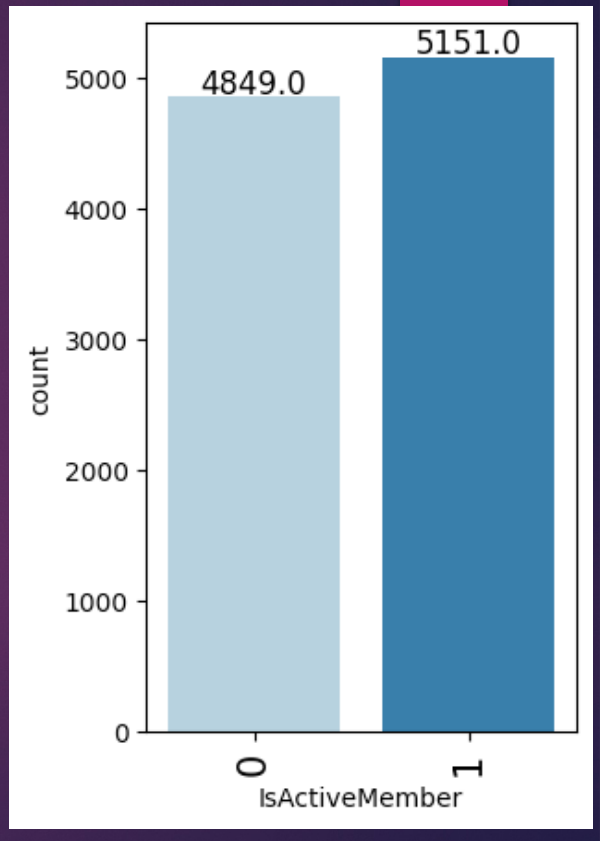


Univariate Visualization III









Bivariate Analysis I

Bivariate analysis explores the relationships between two variables simultaneously. It helps in understanding how one variable affects another and is particularly useful for finding correlations, patterns, and dependencies between variables. Common tools and techniques for bivariate analysis include:

Scatter Plots: Scatter plots are used to visualize the relationship between two numerical variables. They can help identify correlations, trends, or clusters.

Correlation Matrix: Compute correlation coefficients (e.g., Pearson's correlation) to quantify the strength and direction of the linear relationship between pairs of numerical variables. A correlation matrix summarizes these relationships.

Cross-Tabulation (Contingency Tables): Cross-tabulation is used for categorical variables. It shows the distribution of one variable concerning the categories of another variable. It's helpful in understanding associations between categorical variables.

Heatmaps: Heatmaps of correlation matrices provide a visual representation of relationships between numerical variables. They use color coding to indicate the strength and direction of correlations.

Bivariate Analysis II

Scatter plots and correlation coefficients revealed weak or no significant linear correlations between numerical variables (CreditScore, Age, Balance) and churn.

Cross-tabulations showed how churn rates (Exited) varied across different geographical regions (Geography) and between genders (Gender).

CreditScore vs. Age: A scatter plot shows no strong linear relationship between CreditScore and Age. The correlation coefficient suggests a weak, if any, linear correlation.

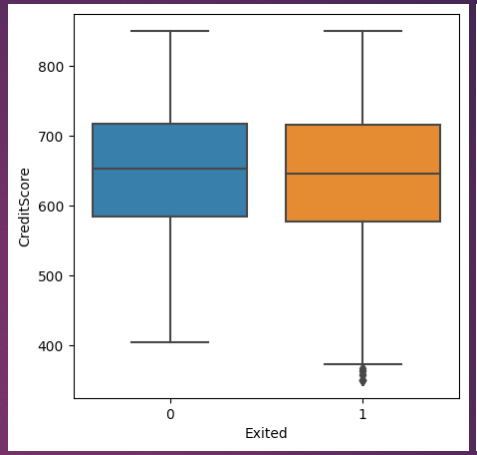
Balance vs. Age: A scatter plot shows no clear linear relationship between Balance and Age. The correlation coefficient indicates a weak correlation.

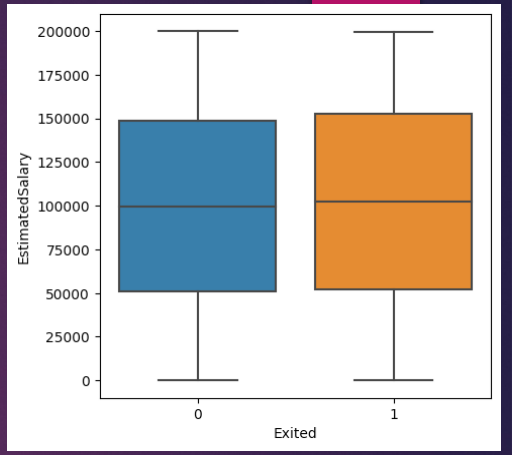
Geography vs. Exited: Cross-tabulation reveals how customer churn (Exited) varies across different geographical regions. It provides insights into the influence of geography on churn rates.

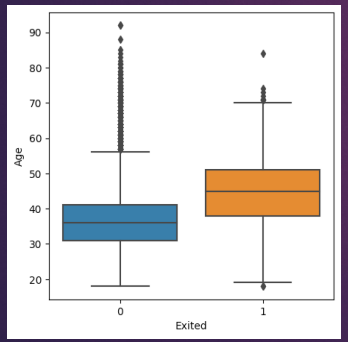
Gender vs. Exited: Cross-tabulation illustrates how churn rates (Exited) differ between genders. It helps understand if gender has an impact on customer churn.

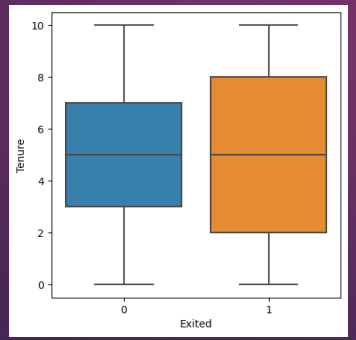
These analyses provide an initial understanding of the dataset's variables and relationships. Further, more advanced analyses and modeling can be performed to uncover deeper insights and build predictive models for customer churn. Relationships between variables were examined, especially concerning customer churn (Exited).

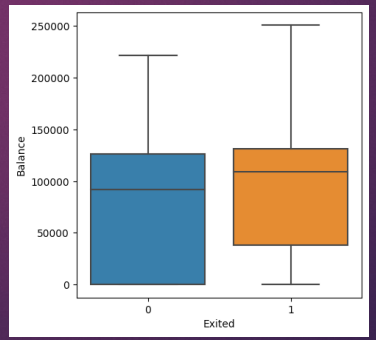
Bivariate Visualization I

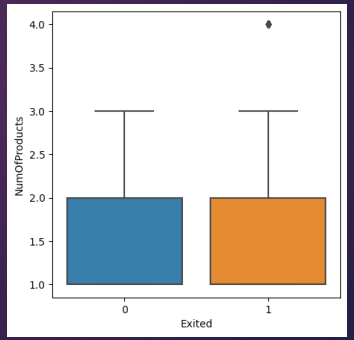






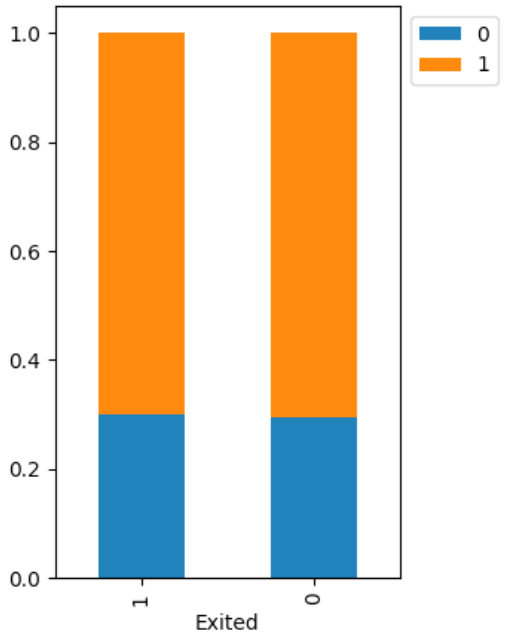




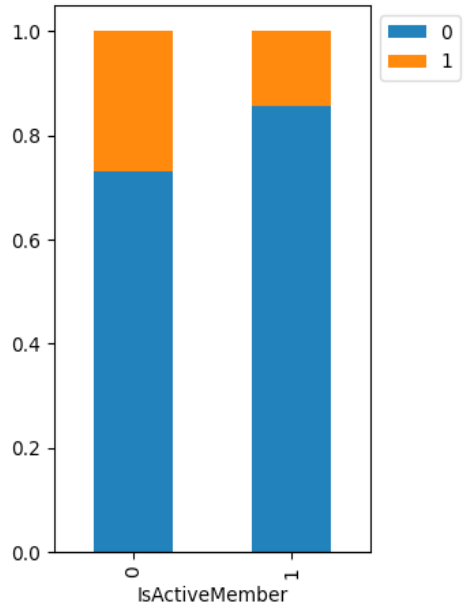


Bivariate Visualization II

HasCrCard	0	1	All
Exited			
All	2945	7055	10000
0	2332	5631	7963
1	613	1424	2037



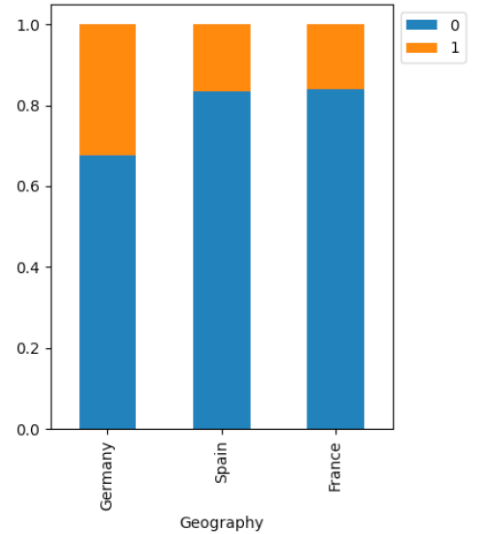
Exited	0	1	All
IsActiveMember			
All	7963	2037	10000
0	3547	1302	4849
1	4416	735	5151



Exited Vs Geography

```
stacked_barplot(ds, "Geography", "Exited" )
```

Exited	0	1	All
Geography			
All	7963	2037	10000
Germany	1695	814	2509
France	4204	810	5014
Spain	2064	413	2477



Data Preprocessing

Duplicate value check

Missing value treatment

Outlier check (treatment if needed)

Feature engineering

Data preparation for modeling

Data Preprocessing Tasks I

Duplicate Values : 0

```
#Correct code to check for duplicate values
duplicates = ds[ds.duplicated()]
num_duplicates = len(duplicates)
print(f'Number of duplicate rows: {num_duplicates}')
```

```
Number of duplicate rows: 0
```

No Missing Value Treatment

```
num_values = ds.isnull().sum()
print(num_values)
```

```
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

```
ds.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
```

Outlier Treatment → No treatment is necessary

Data Preprocessing Tasks II

Feature Engineering and Data Preparation for Modeling

- 'RowNumber', 'CustomerID' & 'Surname' features were dropped from data
- 'Exited' was dropped as 'X ' variable & created as 'y'
- Testing, training and validation sets were created
- Dummy Variables created for categorical values of 'Geography' & 'Gender'
- StandardScaler is used for 'Normalizing' numeric feature values
- 'SMOTE' was applied to balance the dataset with hyperparameter tuning
- Threshold selections were chosen after model tuning based on G-mean / ROC
- Classification Reports were generated to evaluate model performance on testing sets
- Confusion Matrix visualizations were created to assess predictions and performance

```
columns_to_drop = ['CustomerId', 'Surname', 'RowNumber']
ds = ds.drop(columns=columns_to_drop)
```

```
duplicates = ds[ds.duplicated()]
num_duplicates = len(duplicates)
print(f'Number of duplicate rows: {num_duplicates}')
```

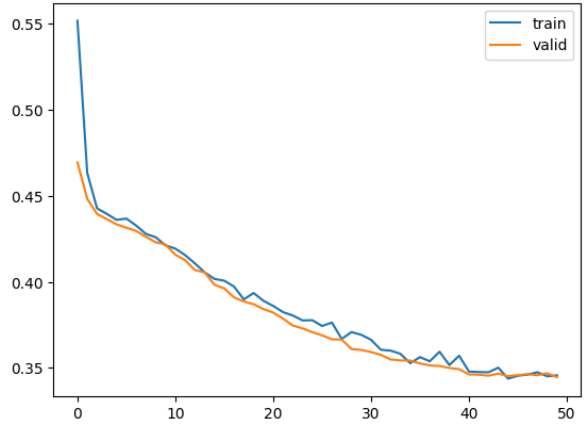
```
Number of duplicate rows: 0
```

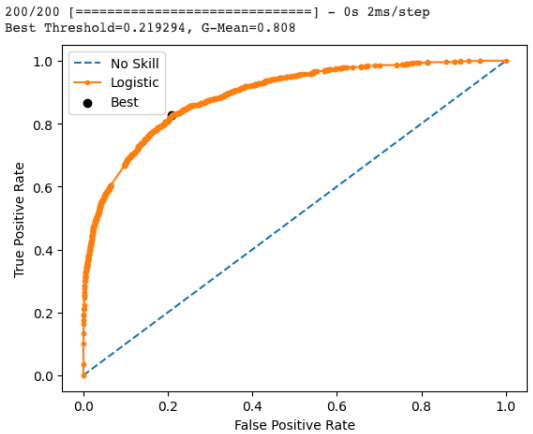
Model Performance Summary

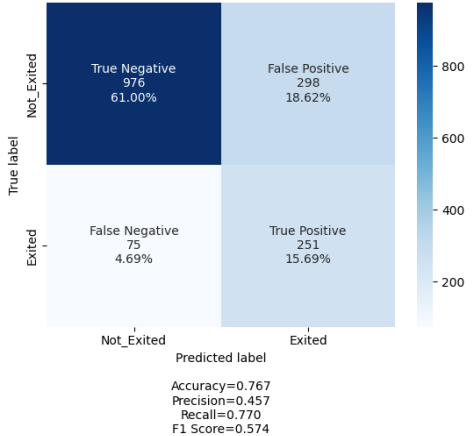
Overview of model and its parameters

Summary of the final model for prediction

Summary of key performance metrics for training and test data in tabular format for comparison







Model Performance Summary I

Overview of model and its parameters I

Model Overview (Before Hyperparameter Tuning):
Type: Artificial Neural Network (ANN)
Purpose: Classification
Input Layer: Dense layer with ReLU activation function : Number of Neurons: 64
Hidden Layer: Dropout layer (to prevent overfitting) : Dropout Rate: 0.5
Hidden Layer: Dense layer with ReLU activation function : Number of Neurons: 32
Output Layer: Dense layer with Sigmoid activation function : Number of Neurons: 1 (Binary classification)

Model Parameters (Before Hyperparameter Tuning):
Loss Function: Binary Cross-Entropy
Optimizer: Adam
Learning Rate: 0.001
Number of Epochs: 100

Grid Search for Hyperparameter Tuning:
Grid search was used to optimize two hyperparameters: batch size and learning rate.
The hyperparameter tuning process used the KerasClassifier with a custom function called “create_model_v2”.
Grid search considered different values for batch size (40, 64, 128) and learning rate (0.01, 0.001, 0.1).
The best combination of hyperparameters was selected based on the highest validation accuracy.

Model Performance Summary I

Overview of model and its parameters II

Model After Applying SMOTE for Data Balancing - Model 3:
Type: Artificial Neural Network (ANN)
Purpose: Classification
Input Layer: Dense layer with ReLU activation function : Number of Neurons: 32
Hidden Layer: Dropout layer (to prevent overfitting) : Dropout Rate: 0.2
Hidden Layer: Dense layer with ReLU activation function : Number of Neurons: 16
Hidden Layer: Dropout layer (to prevent overfitting) : Dropout Rate: 0.1
Hidden Layer: Dense layer with ReLU activation function : Number of Neurons: 8
Output Layer: Dense layer with Sigmoid activation function
Number of Neurons: 1

Model Parameters After SMOTE (Model 3):
Loss Function: Binary Cross-Entropy
Optimizer: Adam : Learning Rate: 0.001
Early Stopping: Patience of 5 epochs for early stopping
Batch Size: 64

Model Performance Summary II

Summary of the final model for prediction I

1. Model Exploration and Selection:
Multiple models were explored in the dataset. These models varied in their architecture, hyperparameters, and training strategies. The goal was to find a model that could predict customer churn with high accuracy, especially focusing on the recall metric, as identifying potential churners is crucial for business retention strategies.

2. Hyperparameter Tuning:
Hyperparameter tuning is the process of systematically searching for the best combination of hyperparameters that optimize a model's performance. In the case of estimator_v2, hyperparameters like learning rate (lr) and batch size were tuned using techniques like grid search. The best parameters were selected based on their performance on the validation set.

3. Model Architecture:
The final model, estimator_v2, is a neural network with four layers:
A dense layer with 64 units.
A dropout layer for regularization to prevent overfitting.
Another dense layer with 32 units.
A final dense layer with a single unit for binary classification.

4. Model Evaluation:
After training, the model was evaluated on a validation set. Key performance metrics like precision, recall, f1-score, and accuracy were considered. The recall metric was given priority because it's essential to identify as many actual churners as possible, even if it means flagging some non-churners incorrectly.

Model Performance Summary II

Summary of the final model for prediction II

4. II : For estimator_v2, the recall for predicting churn (labelled as '1') was found to be 0.77, which means the model correctly identified 77% of the actual churners in the validation set. This is a significant metric for businesses as it helps in proactively addressing potential churn and implementing retention strategies.

5. Threshold Optimization:
To further enhance the model's performance, the threshold for classifying a customer as a churner was optimized. Instead of the default 0.5, a threshold that maximized the geometric mean was chosen. This optimization ensures a balance between false positives and false negatives, especially crucial when one class (churners) might be more costly for the business than the other.

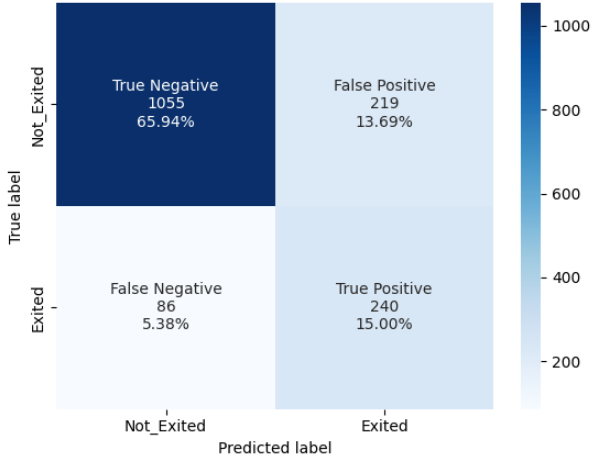
6. Final Model Testing:
After all the optimizations, the model was tested on a separate test dataset to evaluate its real-world performance. The results from this test set were in line with the validation set, indicating that the model is robust and generalizes well to new, unseen data.

Conclusion:
The estimator_v2 model was concluded as the final model based on its performance metrics, especially the recall, and its ability to generalize well to new data. The process of iterative model training, hyperparameter tuning, threshold optimization, and rigorous evaluation ensured that the model is both accurate and reliable for predicting customer churn.

Model Performance Summary III

Final Model Comparison of Testing and Training Data for Performance Metrics

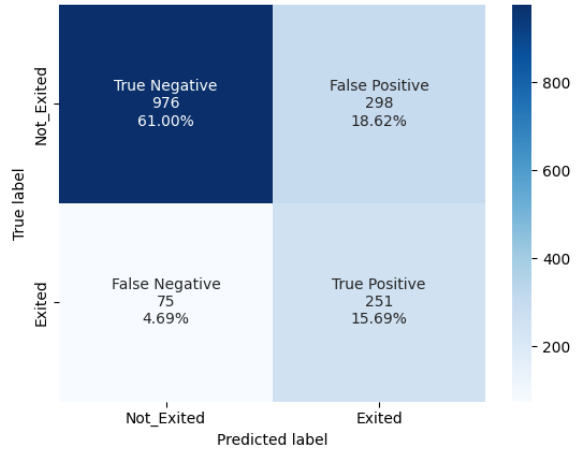
Train



Accuracy=0.809
Precision=0.523
Recall=0.736
F1 Score=0.611

	precision	recall	f1-score	support
0	0.92	0.83	0.87	1274
1	0.52	0.74	0.61	326
accuracy			0.81	1600
macro avg	0.72	0.78	0.74	1600
weighted avg	0.84	0.81	0.82	1600

Test



Accuracy=0.767
Precision=0.457
Recall=0.770
F1 Score=0.574

	precision	recall	f1-score	support
0	0.93	0.77	0.84	1593
1	0.47	0.78	0.58	407
accuracy			0.77	2000
macro avg	0.70	0.78	0.71	2000
weighted avg	0.84	0.77	0.79	2000

Appendix

Data Background and Contents I

DATA OVERVIEW:

THE DATASET IS RELATED TO A BANK'S CUSTOMER CHURN DATA. THE PRIMARY GOAL IS TO PREDICT WHETHER A CUSTOMER WILL CHURN (LEAVE THE BANK) OR NOT BASED ON VARIOUS FEATURES.

FEATURES:

ROWNUMBER: A UNIQUE IDENTIFIER FOR EACH ROW/RECORD

CUSTOMERID: A UNIQUE IDENTIFIER FOR EACH CUSTOMER.

SURNAME: THE SURNAME OF THE CUSTOMER.

CREDITSCORE: THE CREDIT SCORE OF THE CUSTOMER.

GEOGRAPHY: THE COUNTRY OR REGION WHERE THE CUSTOMER RESIDES.

GENDER: THE GENDER OF THE CUSTOMER (E.G., MALE, FEMALE).

AGE: THE AGE OF THE CUSTOMER.

TENURE: THE NUMBER OF YEARS THE CUSTOMER HAS BEEN WITH THE BANK.

BALANCE: THE CURRENT BALANCE IN THE CUSTOMER'S ACCOUNT.

NUMOFPRODUCTS: THE NUMBER OF PRODUCTS THE CUSTOMER HAS WITH THE BANK.

Data Background and Contents II

DATA OVERVIEW II:

HASCRCARD: INDICATES WHETHER THE CUSTOMER HAS A CREDIT CARD (LIKELY BINARY: 1 FOR YES, 0 FOR NO).

ISACTIVEMEMBER: INDICATES WHETHER THE CUSTOMER IS AN ACTIVE MEMBER (LIKELY BINARY: 1 FOR YES, 0 FOR NO).

ESTIMATEDSALARY: THE ESTIMATED SALARY OF THE CUSTOMER.

EXITED: THE TARGET VARIABLE INDICATING WHETHER A CUSTOMER EXITED (CHURNED) OR NOT (LIKELY BINARY: 1 FOR YES, 0 FOR NO).

DATA INSIGHTS:

THE DATASET CONTAINS 10,000 RECORDS.

THERE ARE NO MISSING VALUES IN THE DATASET, AS INDICATED BY THE NON-NULL COUNTS.

THE DATASET CONTAINS A MIX OF NUMERICAL (BOTH CONTINUOUS AND DISCRETE) AND CATEGORICAL FEATURES.

INITIAL EXPLORATORY DATA ANALYSIS INCLUDES UNIVARIATE ANALYSIS, BIVARIATE ANALYSIS, AND CORRELATION PLOTS TO UNDERSTAND THE RELATIONSHIPS BETWEEN VARIABLES AND THEIR IMPACT ON THE TARGET VARIABLE.

Data Background and Contents III

VISUALIZATIONS:

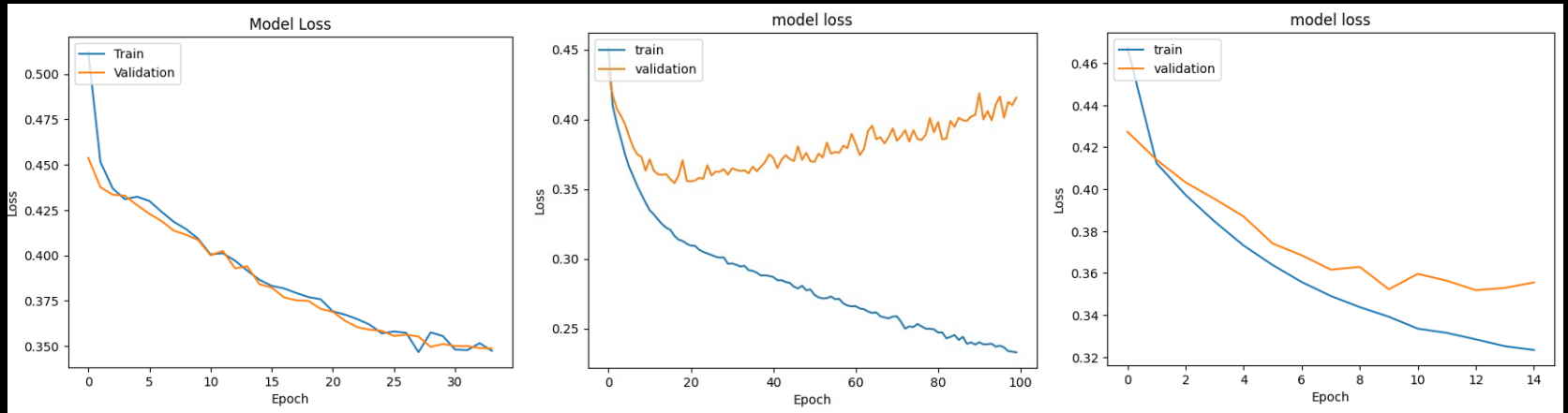
THE NOTEBOOK CONTAINS VARIOUS VISUALIZATIONS TO UNDERSTAND THE DISTRIBUTION AND RELATIONSHIPS OF FEATURES. THESE INCLUDE:

HISTOGRAMS AND BOX PLOTS FOR CONTINUOUS FEATURES LIKE CREDITSCORE, AGE, BALANCE, AND ESTIMATEDSALARY.

BAR PLOTS FOR CATEGORICAL FEATURES LIKE EXITED AND GEOGRAPHY.

A CORRELATION HEATMAP TO UNDERSTAND THE LINEAR RELATIONSHIPS BETWEEN FEATURES.

IN SUMMARY, THE DATASET PROVIDES COMPREHENSIVE INFORMATION ABOUT THE BANK'S CUSTOMERS, AND THE PROBLEM FOCUSES ON ANALYZING THIE DATA TO PREDICT CUSTOMER CHURN. THE COMBINATION OF FEATURES OFFERS A HOLISTIC VIEW OF THE CUSTOMER'S RELATIONSHIP WITH THE BANK, WHICH CAN BE INSTRUMENTAL IN PREDICTING THEIR LIKELIHOOD TO CHURN.



Description of Data

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

Additional Observations

General Observations Worth Mentioning

There are no outstanding correlations among features based on heatmap review

3 notable correlations exist with 'Exited' & 'Age'= 0.29, 'Exited' & 'Balance'= 0.12 and an interesting correlation between 'Exited' & IsActiveMember= 0.16

The largest correlation among the features exists between 'Balance' and 'NumOfProducts' = Negative 0.30

France has the highest number of current customers and the lowest ratio of 'Exited' customers : 83.8% are current customers / 16.2% 'Exited'

Germany has the lowest number of current customers and the highest ration of 'Exited' customers : 67.7% are current customers / 32.4% 'Exited'

Number of customers by region:
France : 5014 Current = 4204 : Exited = 810
Germany : 2509 Current = 1695 : Exited = 814
Spain : 2477 Current = 2064 : Exited = 413

There may be an opportunity to identify features within the region of Germany that may lead to further understanding of 'Existed ' customer tendencies

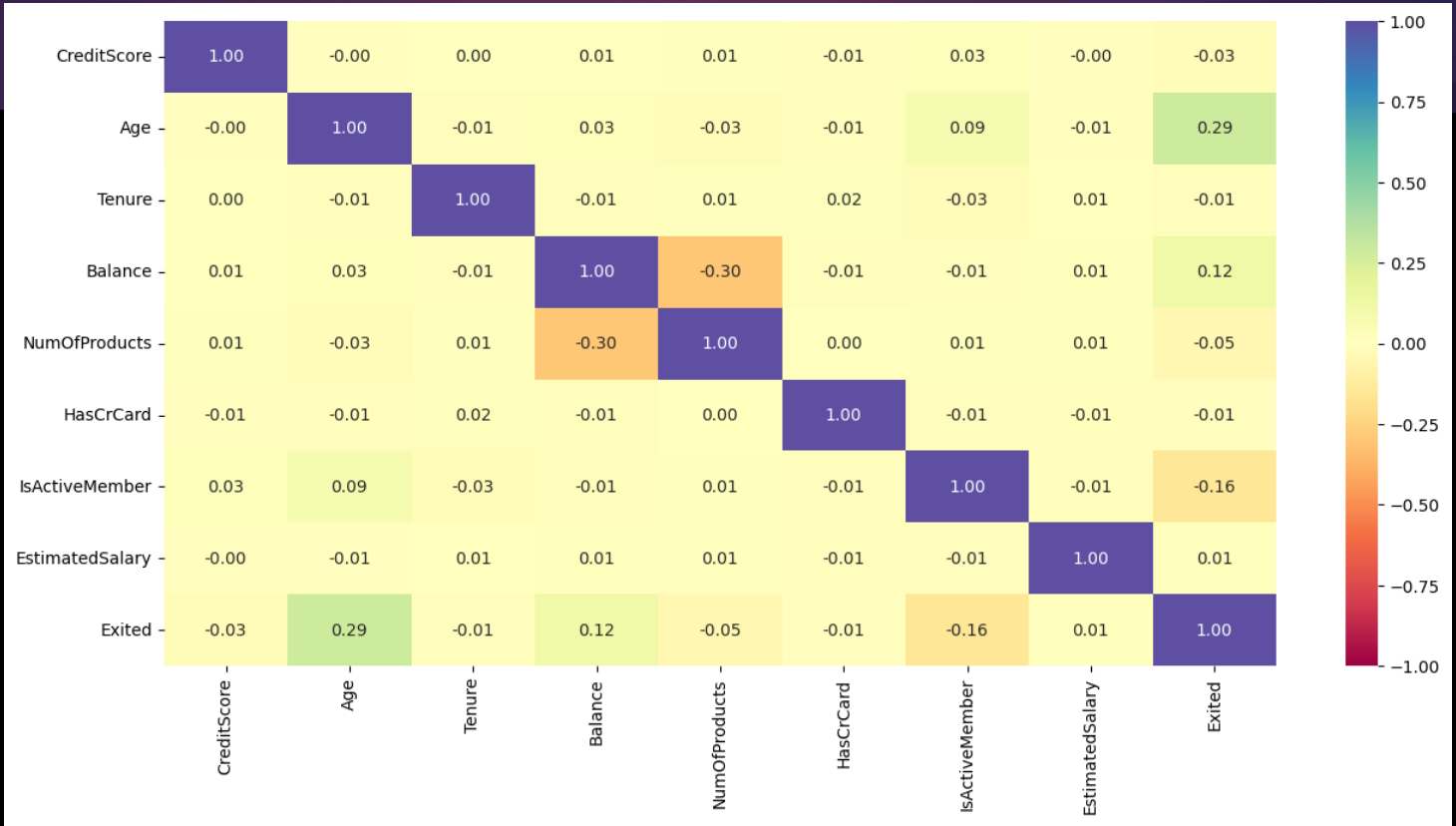
Youngest Customer = 18
Oldest Customer = 92

'EstimatedSalary', 'CreditScore', 'HasCrCard' and 'Tenure' indicate virtually no influence on any of the features in the dataset

The models offered similar results after hyperparameter tuning & the final model was chosen due to consistent testing/training results, recall score & accuracy

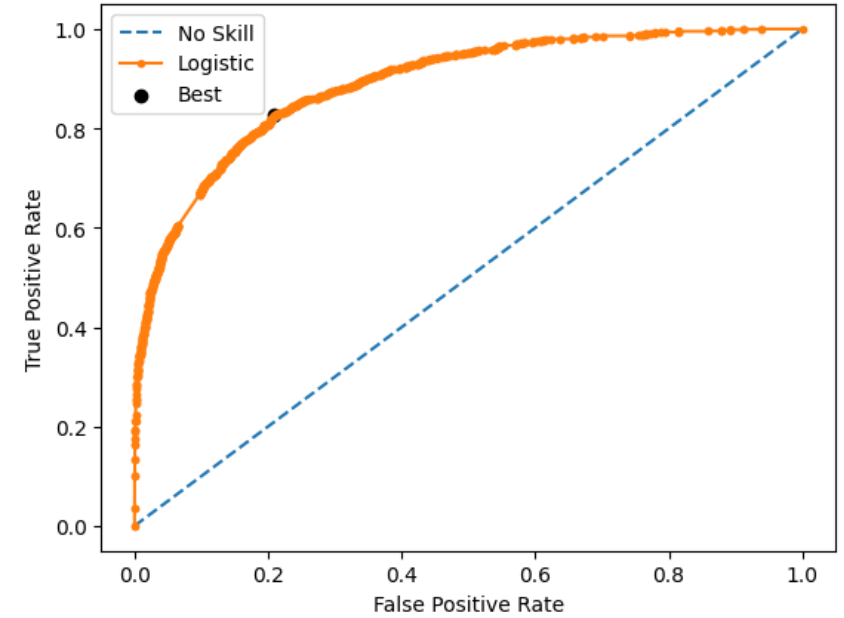
It is possible that higher Recall scores can be attained though the expense and effort may prove greater than the value of the increase in result.

Feature Heatmap

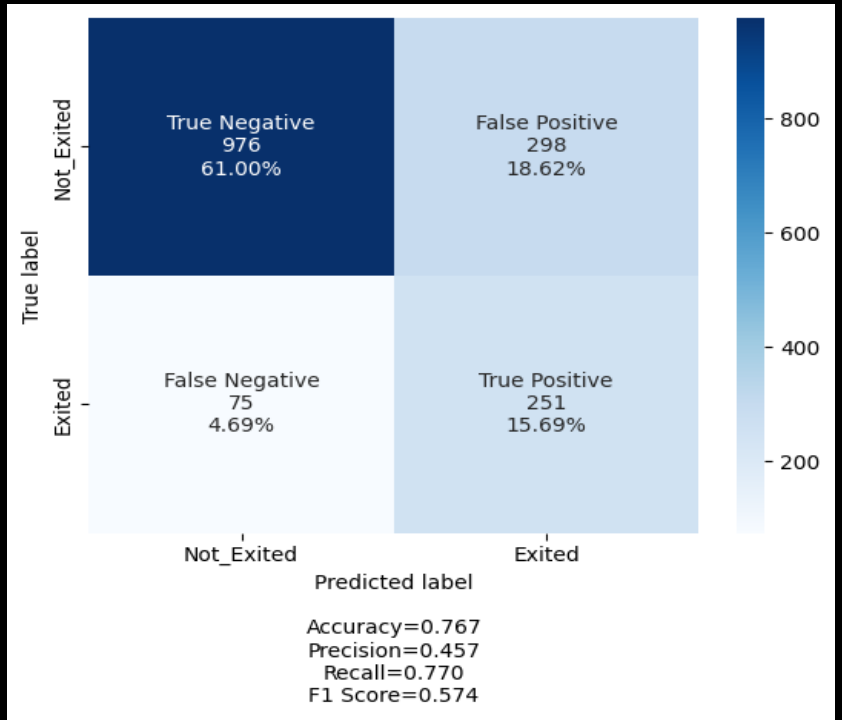


Final Model
ROC Curve

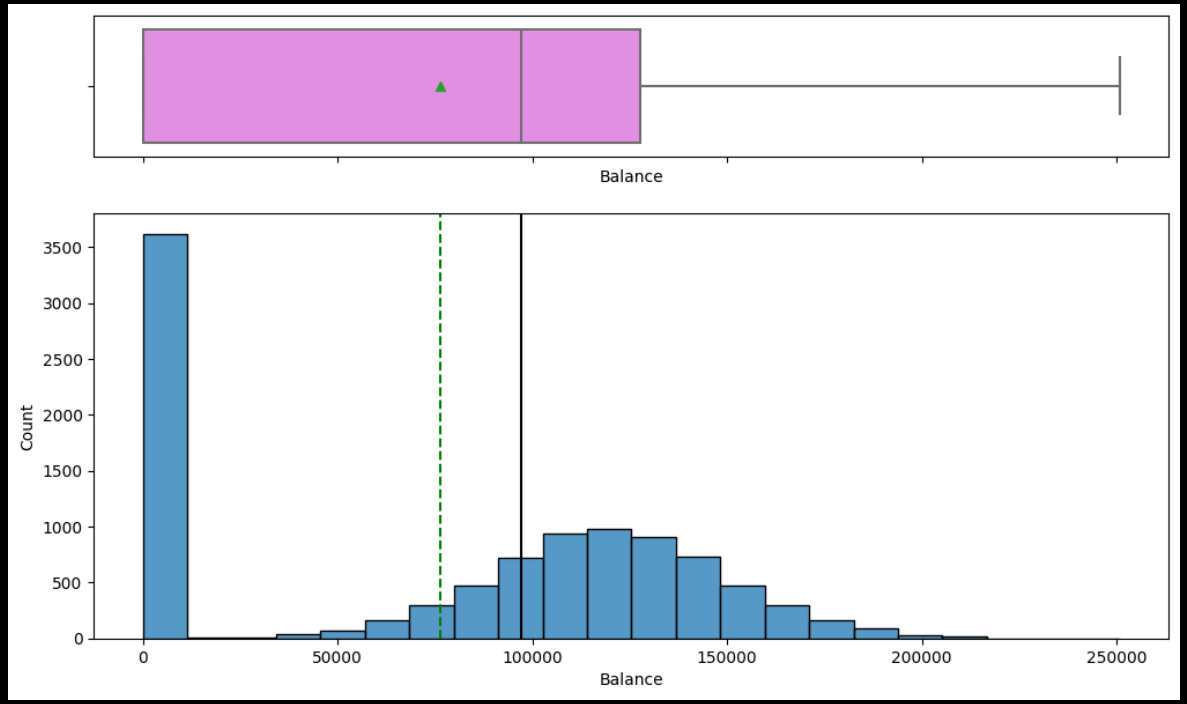
```
200/200 [==============================] - 0s 2ms/step
Best Threshold=0.219294, G-Mean=0.808
```



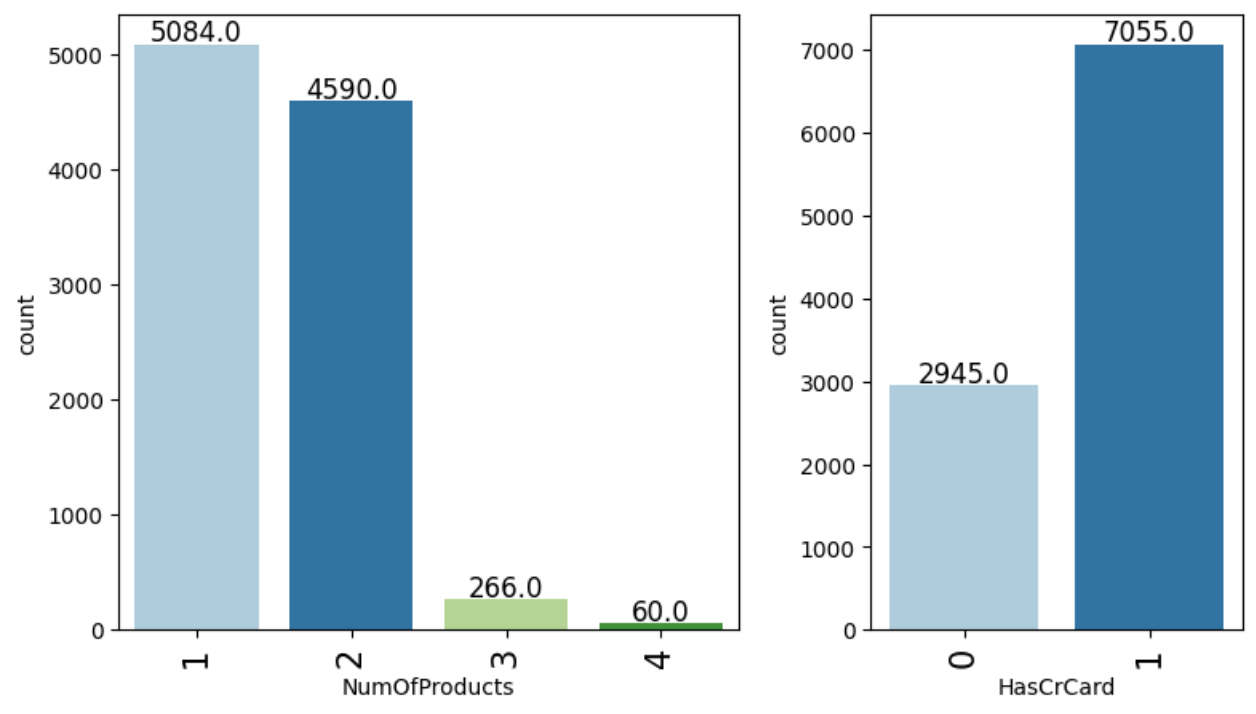
Final Model
Confusion Matrix



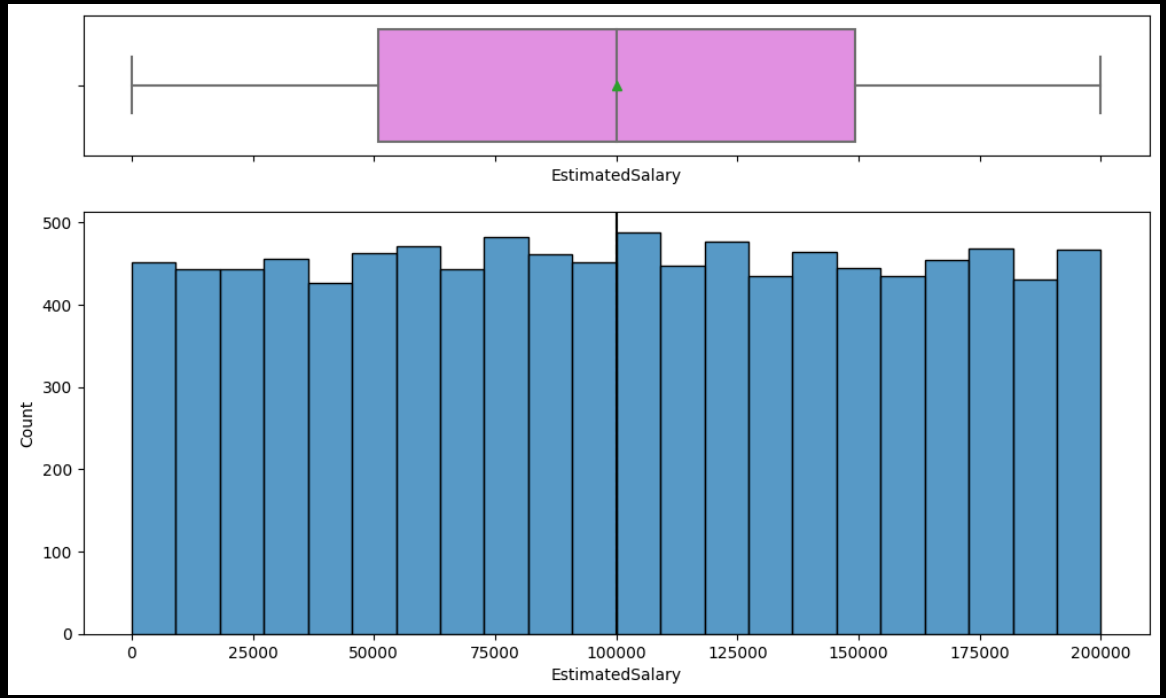
‘Balance’ Histogram_Boxplot



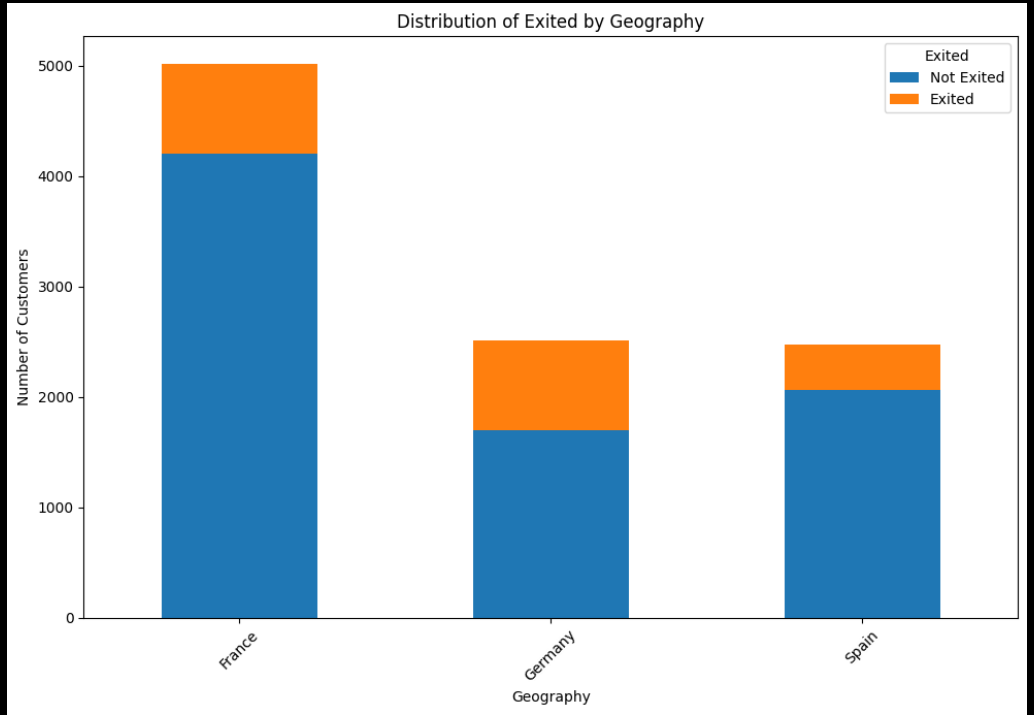
‘HasCrCard’
‘NumOfProducts’
Barplot



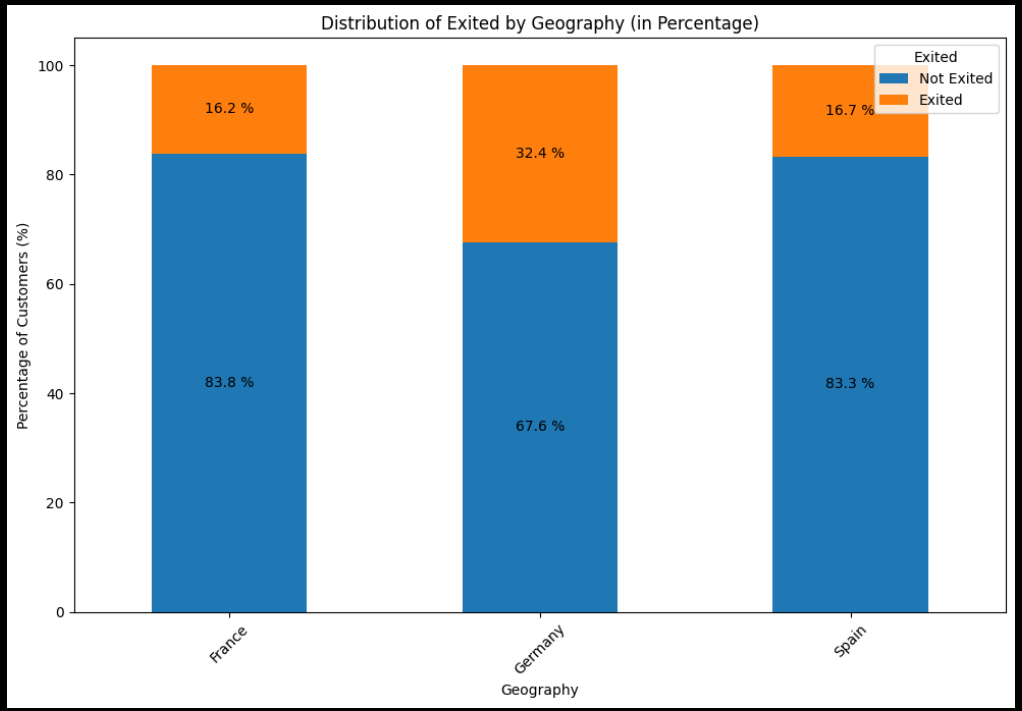
‘EstimatedSalary’ Histogram_barplot



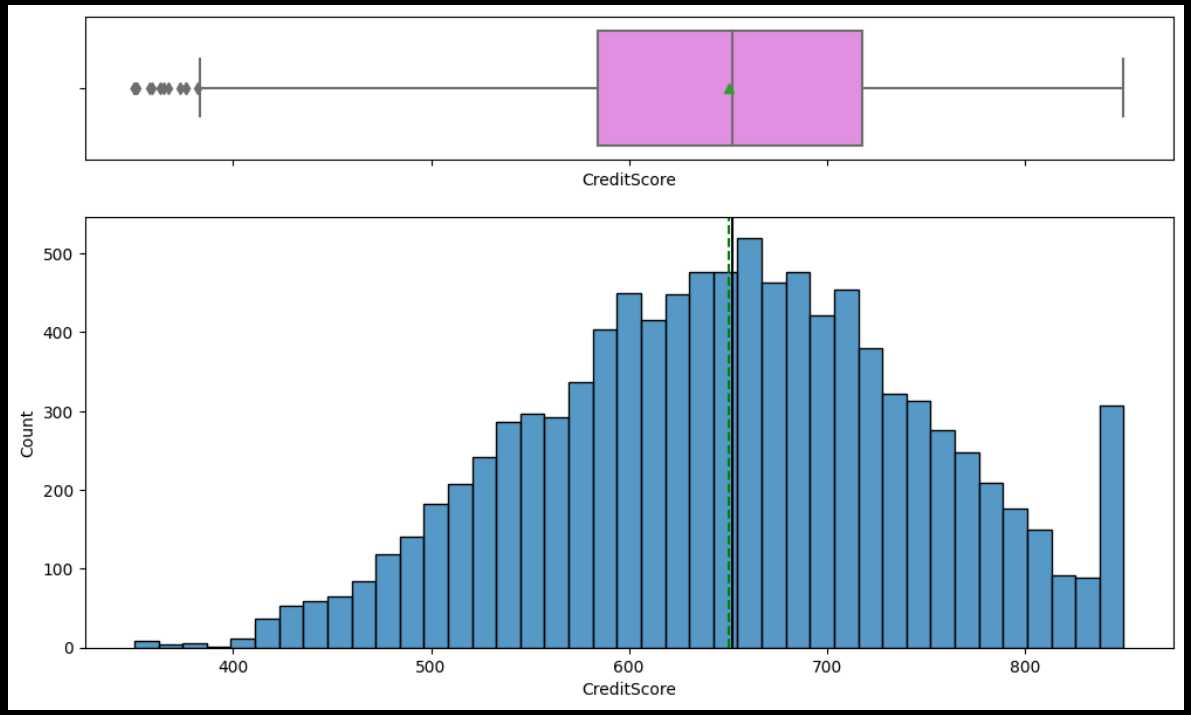
‘Exited’ vs ‘Geography’ Stacked Barplot



‘Exited’ vs ‘Geography’ in Percentages



‘CreditScore’ Histogram_boxplot



Loss Function
‘Model Loss’