Bank Churn Prediction

by Sarah Choi

Date: 11/15/23

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary - Conclusions

	Conclusion
Predictive Model	• Analyzed the dataset from a bank for their customers and applied five different type of techniques to build the neural network models and selected the best Classification predictive model. • The bank can deploy the final model to i) predict the customers who will leave the bank in next six months. ii) to find the drivers of attrition. iii) based on which the bank can take appropriate actions to improve service to retain their them.
Model performance	• 5 models were built and validated; 4 models were fine tuned to get the best performance for final selection.
Outliers	• There are outliers in the variable (Age, Exited), Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results.
Customer attrition Observations	• Our data analysis from the data collection shows that there are ~20% customers left the bank in the last 6 months. • These customers who left the bank are in the older age group, i.e. older than 40. They have been with the bank for longer tenure. • The customers who left the bank were mostly from Germany and there were more female customers than male customers.

Executive Summary - Recommendations

	Recommendations
Marketing Strategy	• Bank can use the final model from this exercise to identify with a reasonable degree of accuracy whether a customer is likely to leave in the next 6 months and should monitor the customer account activities and provide action to prevent them from leaving. • Banks rely on a one-size-fits-all approach. Because of the lack of personal touch, banks tend to use the same approach for all customers, regardless of their financial situation or personality. • Bank should hold marketing campaign gear toward age 40 and older customers and probably tailor to female customers to ensure great customer service. Poor service and poor financial advice emerged as top reasons why customers leave their banks.
Outliers	• It is observed that there are large quantities of outliers in many variables (Age, Exited). • Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results. • Recommended approach: analyze how the initial machine learning or data science problem was framed, how the target population and sample were chosen, and how so many outliers made into the dataset. • The stakeholders (people who directly benefit or lose from the project) should be informed of the decision.

Business Problem Overview and Solution Approach

Problem Statement

Businesses like banks which provide service must worry about problem of 'Customer Churn' i.e. customers leaving and joining another service provider. It is important to understand which aspects of the service influence a customer's decision in this regard. Management can concentrate efforts on improvement of service, keeping in mind these priorities.

The objective of this analysis is You to build a neural network-based classifier that can determine whether a customer will leave the bank or not in the next 6 months.

Solution Approach

Identity focus area

To extract actionable insights from data collection.

Perform Data Analysis & Build Predictive Model

- Perform EDA on Variables that influence the customer churn.
- Build a neural network classification model to predict if the customer is going to churn or not.
- Optimize the model using various techniques to determine the best predication model.

Summarize Recommendation

Summarize the insights and recommendations that will help the bank's business.

Data Overview

Variable	Description
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	It refers to the number of products that a customer has purchased through the bank
Balance	Account balance
HasCrCard	It is a categorical variable that decides whether the customer has a credit card or not
EstimatedSalary	Estimated salary
isActiveMember	It is a categorical variable that decides whether the customer is an active member of the bank or not (Active member in the sense, using bank products regularly, making transactions, etc
Exited	It is a categorical variable that decides whether the customer left the bank within six months or not. It can take two values 0=No (Customer did not leave the bank) 1=Yes (Customer left the bank)

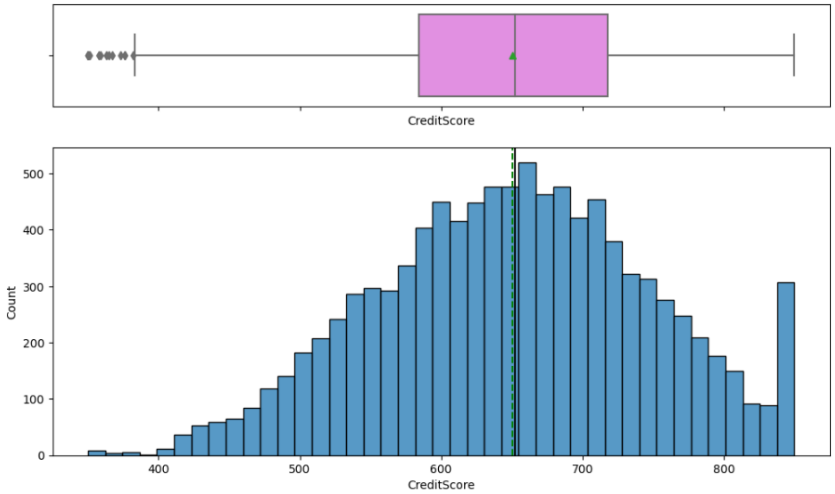
Observations	Variables
10,000	14

Note:

- The case study is from an open-source dataset from Kaggle. The dataset contains 10,000 sample points with 14 distinct features
- There is no missing value in the data

EDA Results - Univariate Data Analysis

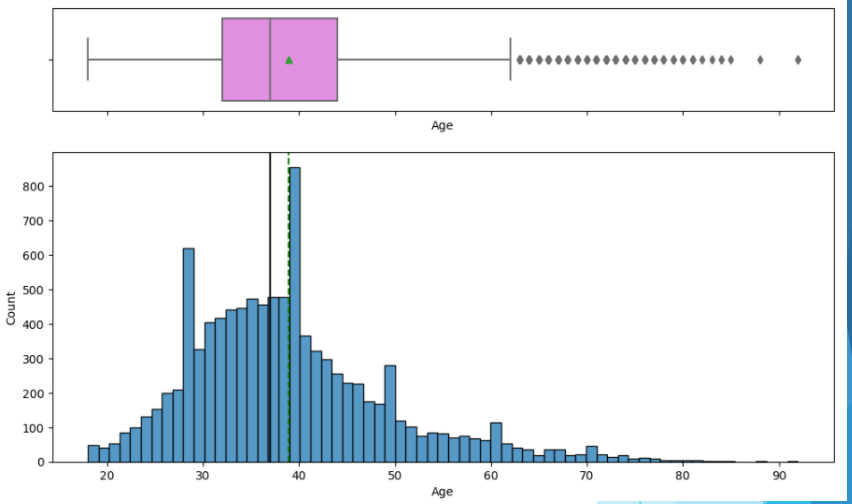
Customer's Credit Score



Observations:

- Credit score has normal distribution, the number of customers who have credit score above the median of 650 are slightly higher than those customers who have below 650 score.

Customer's Age

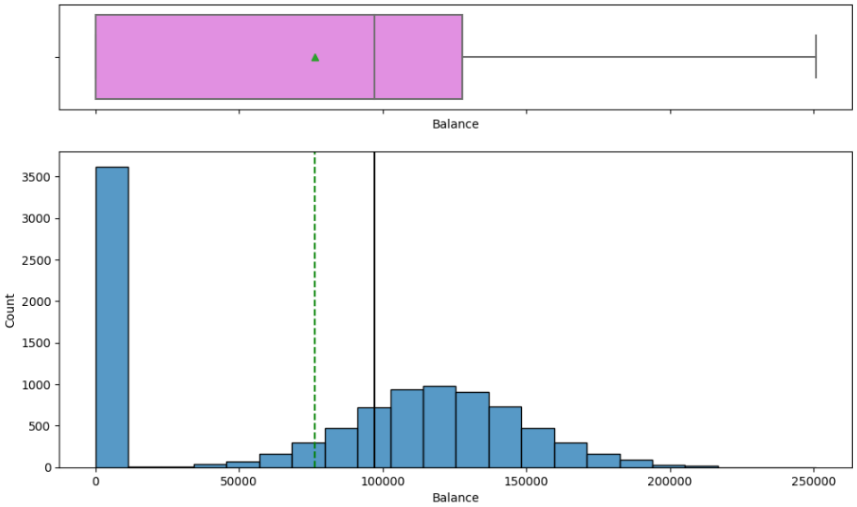


Observations:

- The distribution of the Age is skewed towards the right.
- The highest number of customer is at age 40
- There are many outliers in this variable above age 60 and they are being represented as outliers by the boxplot.

EDA Results - Univariate Data Analysis

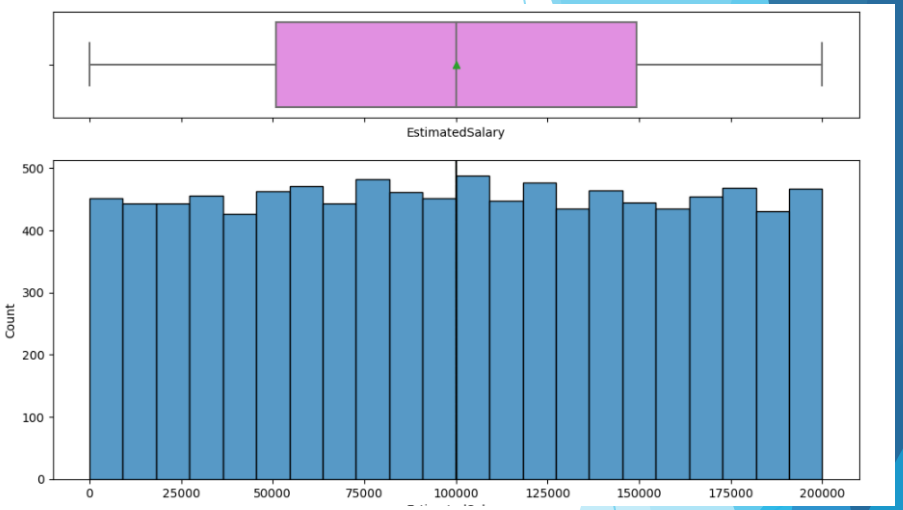
Customer's Balance



Observations:

- The distribution of Balance is skewed towards the right the right due to majority of the customer ~3,600 (36% of total) have zero balance

Customer's Estimated Salary

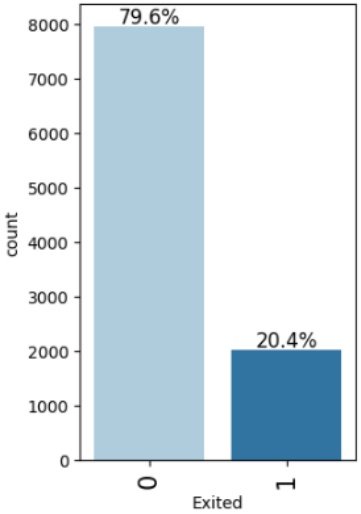


Observations:

- Estimated salary has symmetrical and normal distribution

EDA Results - Univariate Data Analysis

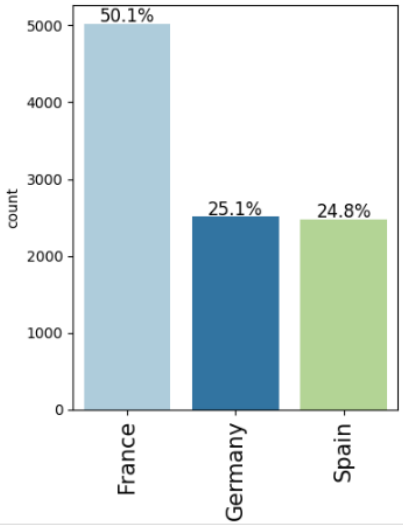
Customer Exited



Observations:

- 79.6% of the customers did not leave the bank within 6 months where 20.4% of the customer left within 6 months.

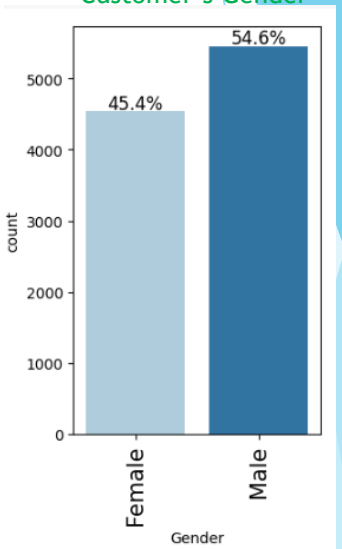
Customer's Geography



Observations:

- Half of the total number of customers are from France, and 25.1% from Germany, the rest from Spain.

Customer's Gender

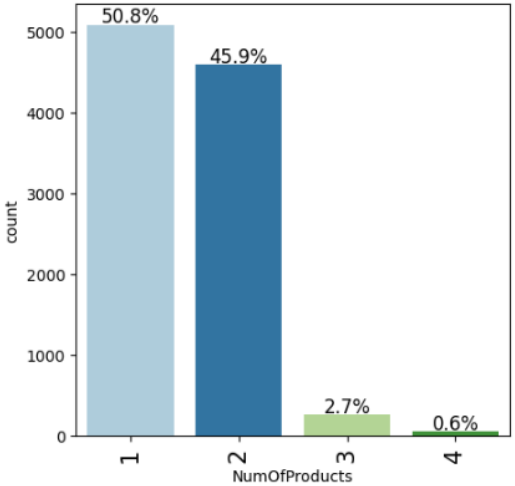


Observations:

- 54.6% of the customers are male and bypassed number of female customers

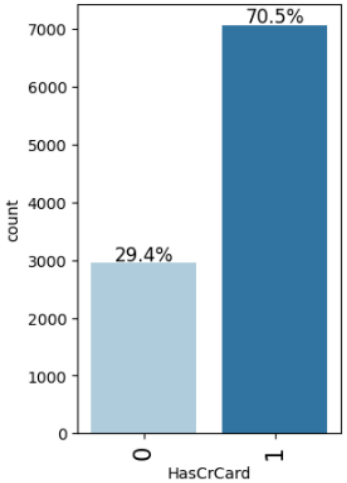
EDA Results - Univariate Data Analysis

Customer Number of Products



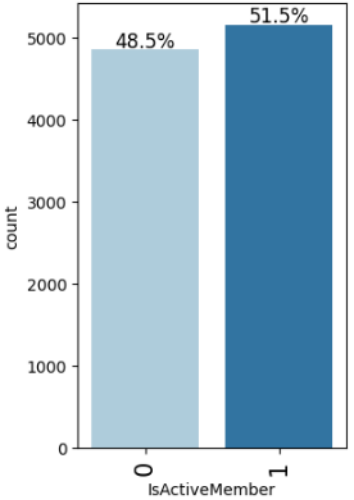
Observations:
- 50.8% of the customers have only one product from the bank, a little less than a half of the customers base have two products.

Has Credit card or Not



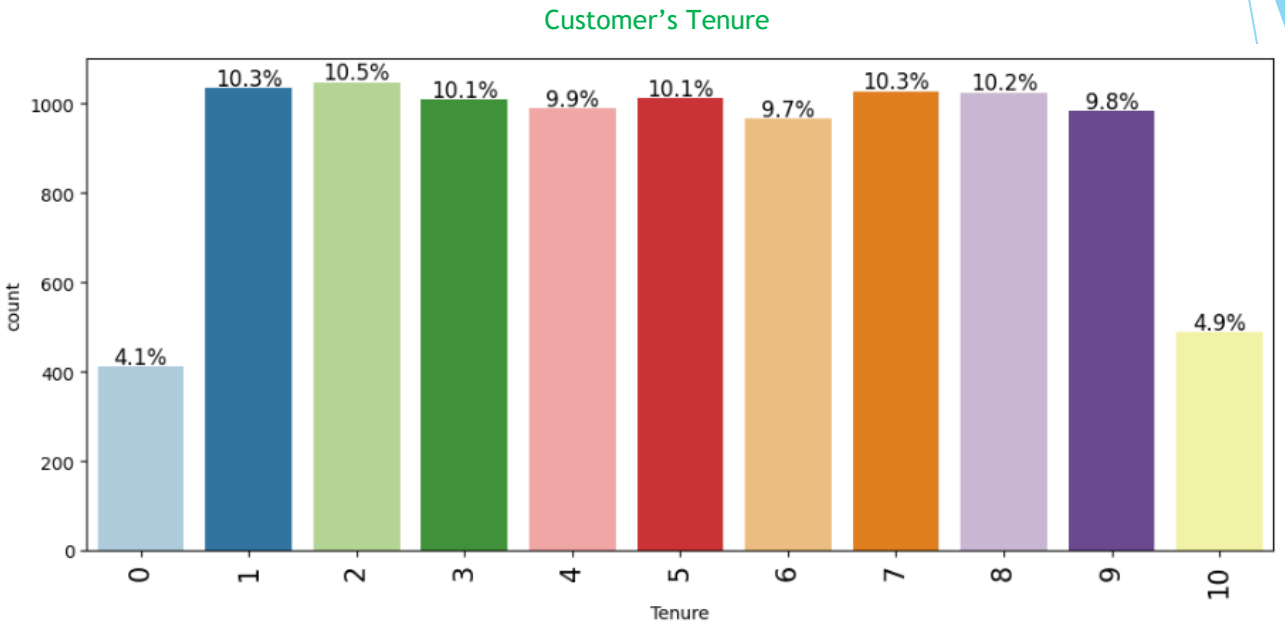
Observations:
- 70.5% of the customers have credit cards where 29.4% of the customers do not have credit cards.

Is active customer or Not



Observations:
- 51.5% are active customers and 48.5 are non-active customers.

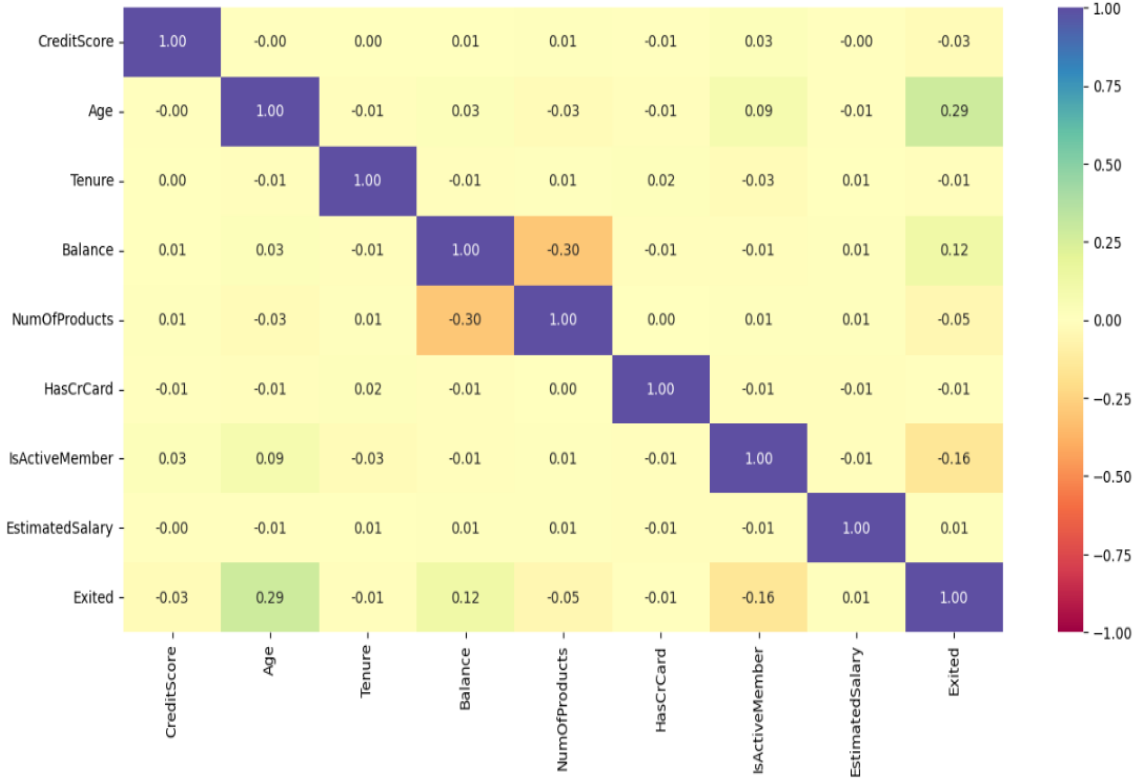
EDA Results - Univariate Data Analysis



Observations:

- Customers Tenure are distributed evenly between 1 year to 9 years with ~4% less than a year and ~5% at 10 years.

Bivariate Analysis - Correlation Matrix



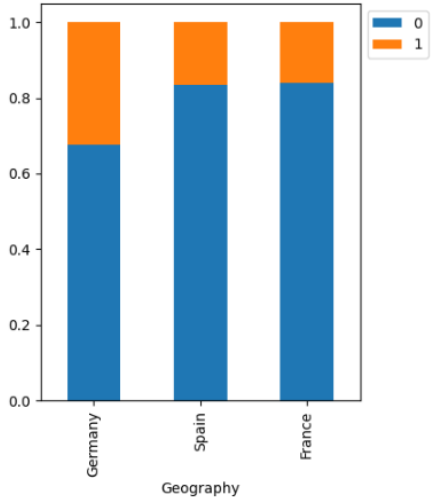
Observations:

- Customer's age has the strongest positive correlation with Customer Exited (0.29), this indicates as age goes up, the more customers exited the bank.
- Account balance also has positive correlation with Customer Exited (0.12), this indicates the higher the balance is, the more customers exited the bank.
- Balance has high negative correlation (-0.30) with number of products which indicates as balance increase the number of product decrease, the higher customer balance, the less product the customer owns.
- IsActiveCustomer flag has negative correlation with Customer Exited (-0.16), if customer is active then customer does not leave the bank.

EDA - Bivariate Analysis

Exited vs Geography

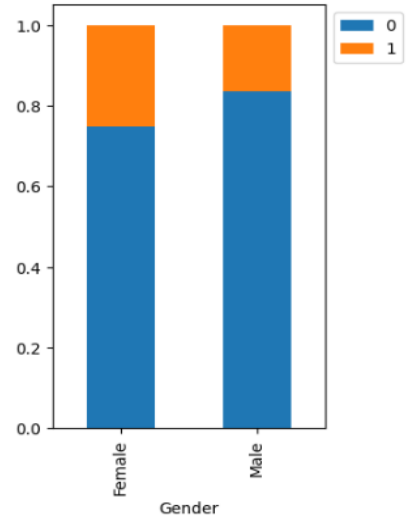
Exited	0	1	All
Geography			
All	7963	2037	10000
Germany	1695	814	2509
France	4204	810	5014
Spain	2064	413	2477



Observations: The number of customers who left the bank mostly were from Germany as compared to customers exited who were from Spain and France that are more even.

Exited vs Gender

Exited	0	1	All
Gender			
All	7963	2037	10000
Female	3404	1139	4543
Male	4559	898	5457

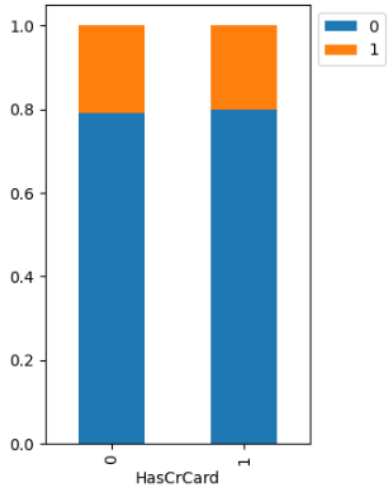


Observations: There are more female customers left the bank than male customers.

EDA - Bivariate Analysis

Exited vs Has Credit Card

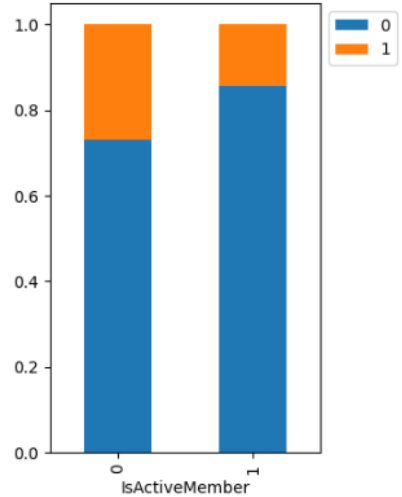
Exited	0	1	All
HasCrCard			
All	7963	2037	10000
1	5631	1424	7055
0	2332	613	2945



Observations: There are more customers who left the bank have credit cards than those who do not have credit card.

Exited vs Active Member

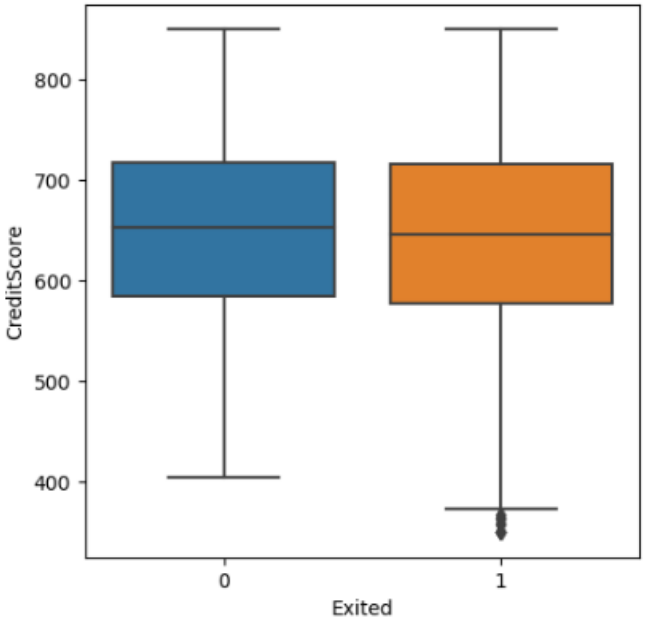
Exited	0	1	All
IsActiveMember			
All	7963	2037	10000
0	3547	1302	4849
1	4416	735	5151



Observations: There are more non-active customers left the bank than those customers that are considered active customers and left the bank within 6 months.

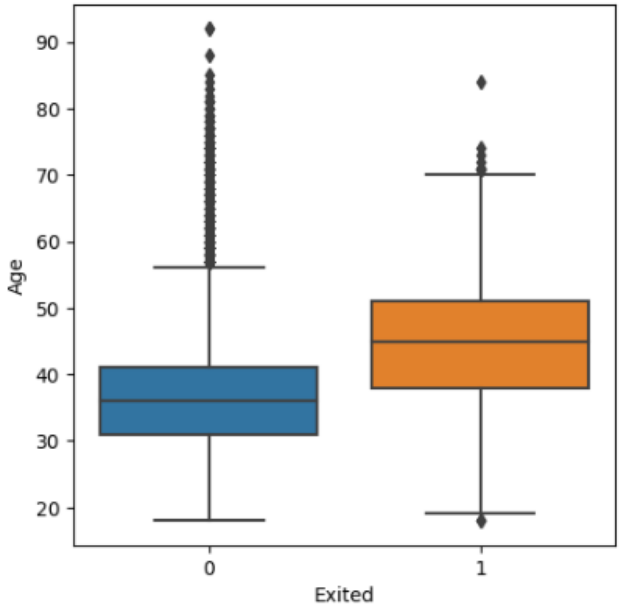
EDA - Bivariate Analysis

Exited vs Has Credit Card



Observations: the distribution and the median of the customers who have credit cards and those who do not have credit card and left the bank are almost the same, there's no significant different.

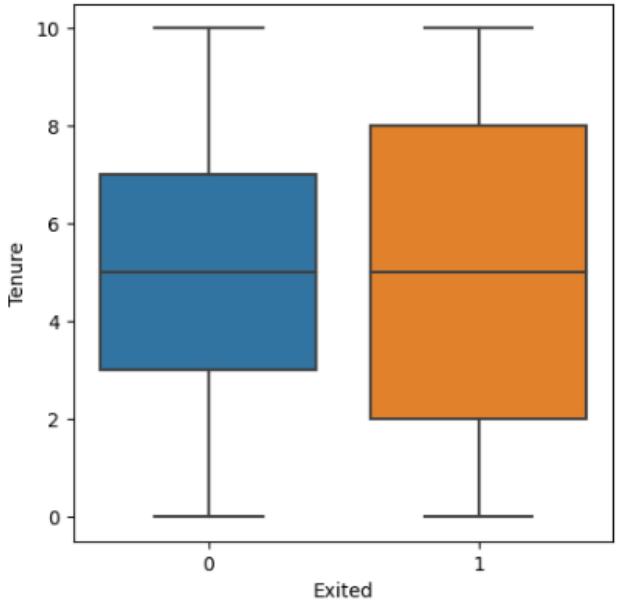
Exited vs Age



Observations: The distribution of the Age is skewed to the right mainly due to the outliers found in that dataset. When compare the distribution of the customer age vs exited. It showed the higher age (older) there are more customers in the older age group left the bank.

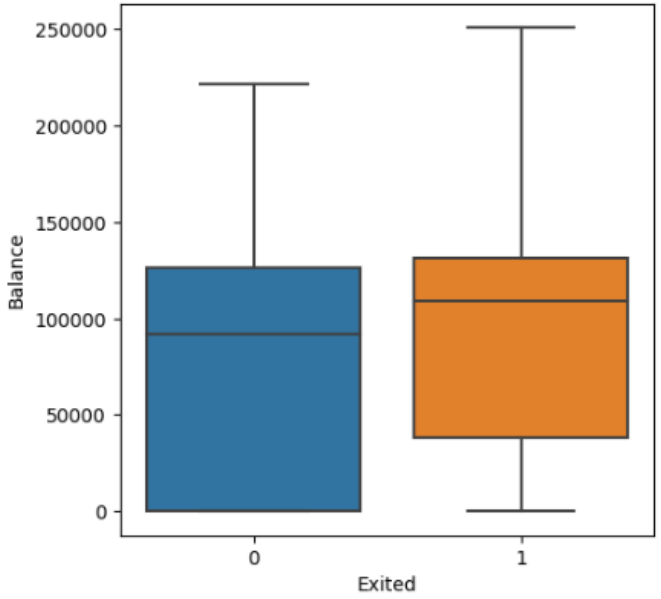
EDA - Bivariate Analysis

Exited vs Has Tenure



Observations: There are more customers who left the banks with higher number of tenure with the bank. The median of 5 years tenure are the same between the customers who left the way vs staying at the bank.

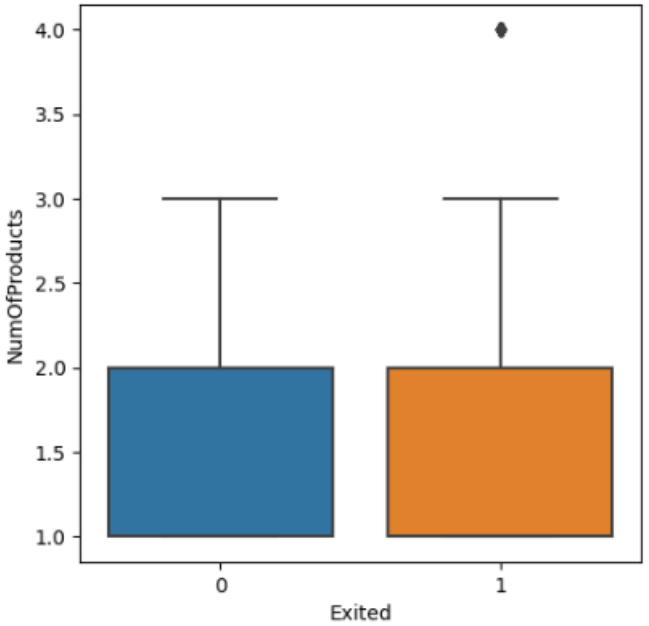
Exited vs Balance



Observations: The distribution of Balance is skewed towards the right mainly due to 36% the customer have zero balance who are staying with the bank. . The median balance of customers who left the bank show a bit higher than those who stay with the bank.

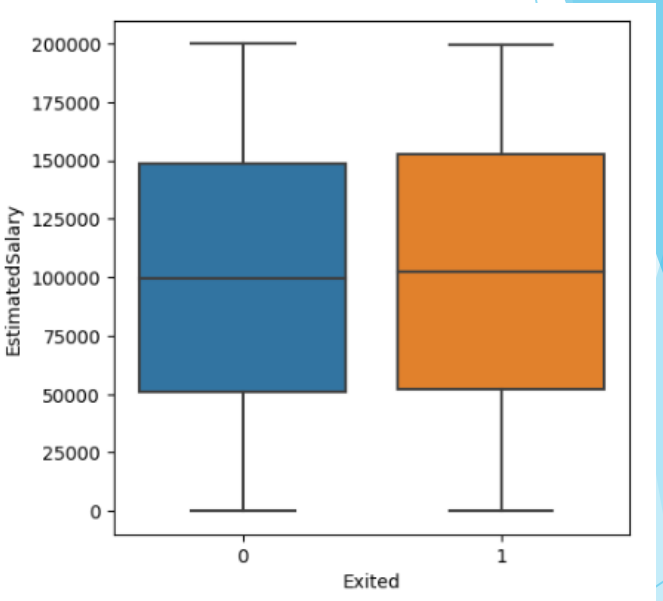
EDA - Bivariate Analysis

Exited vs Has Number of Products



Observations: The distribution of number of products is skewed towards the right mainly due to 98% of the customer have either 1 or 2 products and they are evenly distributed between the customers who left the bank vs staying with the bank, there is no significant different.

Exited vs Estimated Salary



Observations: The distribution of the estimated salary for those customers left the bank as compared to those who stay with the bank are very even, there is no significant different.

Data Preprocessing

- **Duplicate value check:** There is no duplicate row.
- **Missing value treatment:** There is no missing value in the dataset
- **Outlier check (treatment if needed):** Outliers were identified as follows, they were ignored as part of the EDA.

Columns	% of outlier
CreditScore	0.15
Age	3.59
Exited	20.37

- **Feature Engineering:** The columns CustomerIDs, Surname, and Rownumber were dropped because they are unique for all users
- **Data preparation for modeling:**

➢ The independent features and dependent features are obtained in X and y variables respectively

➢ Dummy variables were created by using the label encoding techniques for the columns Geography and Gender

➢ The columns “CreditScore”, “Age”, “Balance”, “EstimatedSalary” were scaled using standard scaler so they all will be in the same range.

Model Building

The Followings are the classification model building steps :

- EDA (Univariate and Bivariate) was performed on the data frame, to get good evaluation of the data before building the model, also data preprocessing was performed to ensure all anomalies are handled.
- Before proceeding with building the model, the data was split the data into train data (80%) and test data (20%). Then the test dataset was further split into validation data (80%) and test data (20%) to be able to evaluate the model that was built on the train data, encode categorical features and scale numerical values.
- Total rows for train data is 6,400, for validation data is 1,600, and for test data is 2,000 after the two data split executions.
- Cross validation techniques to ensure the best model is selected to achieve a generalized model performance in production.
- Neural network and hidden layer with ReLU as activation function, oversampled and under sampled data were used to produce the best model for prediction.
- Hyperparameters tuning was done to tune the models after the models were fitted with and undersampled data to get generalized performance

Model Evaluation Criterion

Model evaluation criterion

- The objective is to predict the customers who will leave the bank and reasons so that bank could improve upon those areas.

Model can make wrong predictions as:

1) Predicting a customer is exiting and the customer is not exiting

2) Predicting a customer is not exiting and customer is exiting

Which case is more important?

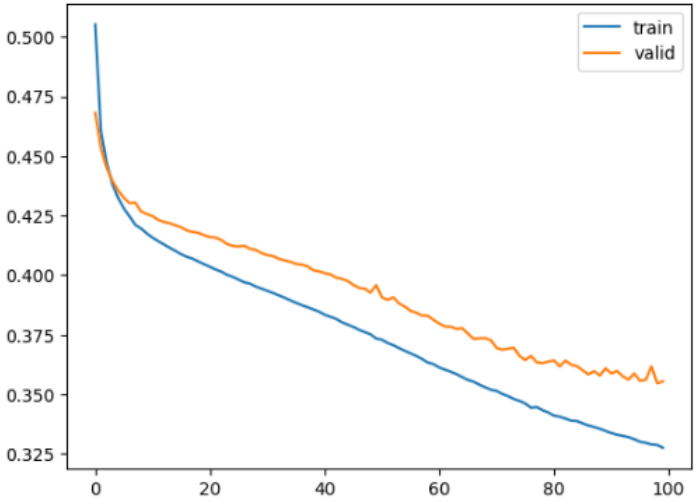
- Predicting that customer is not exiting but he/she is exiting. It might cause loss to the banks because due to wrong identification bank will not be able to take any initiative for those sensitive customers.

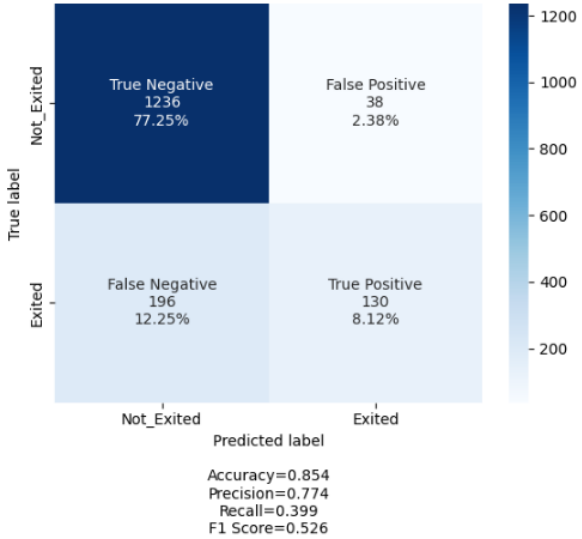
How to reduce this loss i.e need to reduce False Negatives?

- Bank would want **Recall** to be maximized, greater the Recall higher the chances of minimizing false Negative. Hence, the focus should be on increasing Recall or minimizing the false Negative or in other words identifying the True Positive(i.e. Class 1) so that the bank can retain their customers.

Model Performance Summary (Model 1)

- During initial built of the base model by defining the sequential model and add the input layer with 64 neurons with Relu as activation function with input of 11 variables. Then add the 1st hidden layer with 32 neurons, add the output layer with one node and sigmoid activation function.
- We have an output of 1 node, which is the desired dimensions of our output (stay with the bank or not) We use the sigmoid because we want probability outcomes. Predication used default 0.5 threshold. the models are overfitting.



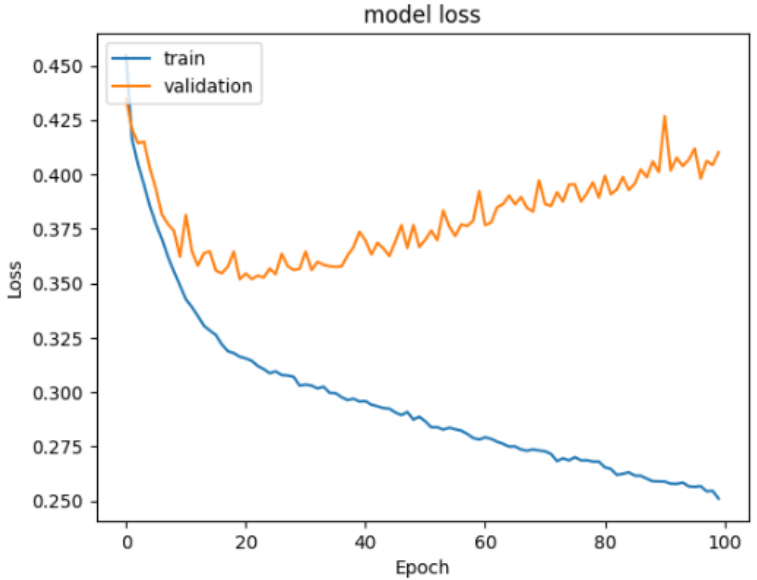


- The result has a good accuracy but a poor Recall score of **0.399** This could be due to the imbalanced dataset as we gave 0.5 as threshold to the model. We observed that the False Negative rates are also high at 12.25%, and we should minimize it.

Performance Summary use Adam Optimizier (Model 2)

- We tried tuning the model by minimizing the binary_crossentropy and used Adam optimizer
- From the below summary, we can see that this architecture trained a total of **2,881** parameters i.e. weights and biases in the network.

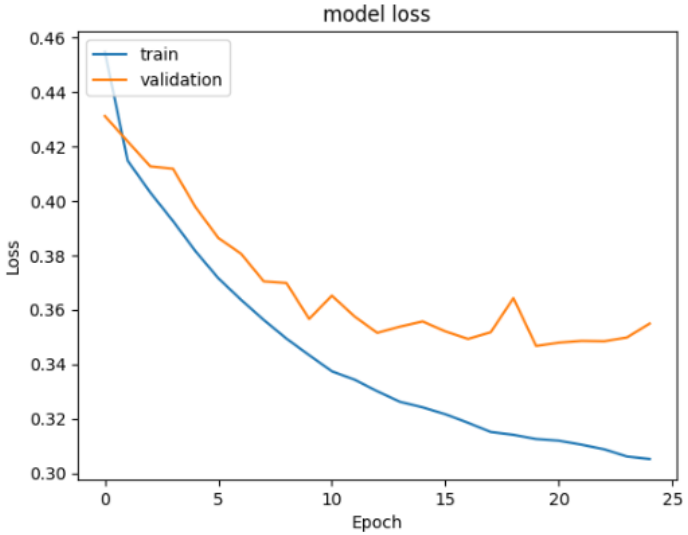
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                768

 dense_1 (Dense)             (None, 32)                2080

 dense_2 (Dense)             (None, 1)                 33

=================================================================
Total params: 2,881
Trainable params: 2,881
Non-trainable params: 0
```

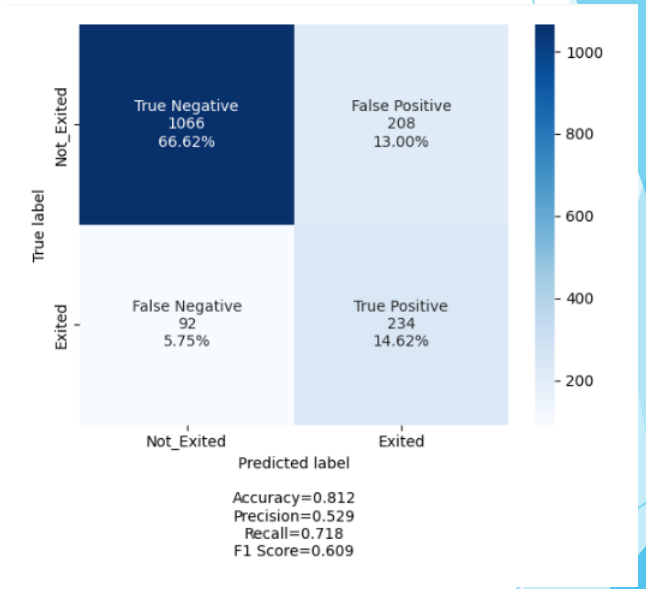


- As you can see from the above image, this model is severely overfitting. Deep learning models are very sensitive to overfitting due to a large number of parameters. We need to find the optimal point where the training should be stopped.
- The best solution for the above problem is **Early stopping** for neural network regularization

Performance Summary use Early Stopping(Model 2)

- The model was trained by using Early stopping and it's evaluated on a holdout validation dataset after each epoch. If the performance of the model on the validation dataset starts to degrade or no improvement (e.g. loss begins to increase or accuracy begins to decrease), then the training process is stopped after certain to gain good generalization performance. The model is less overfitted but still is diverged from the training curve.

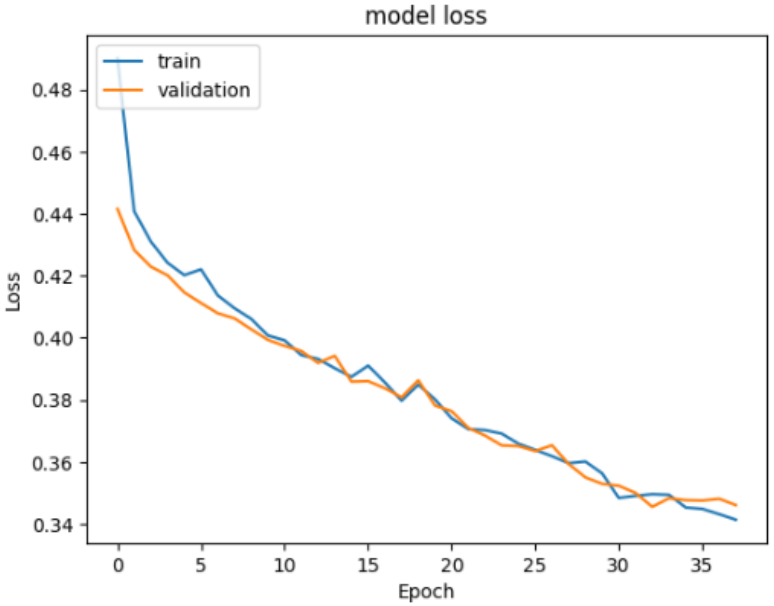


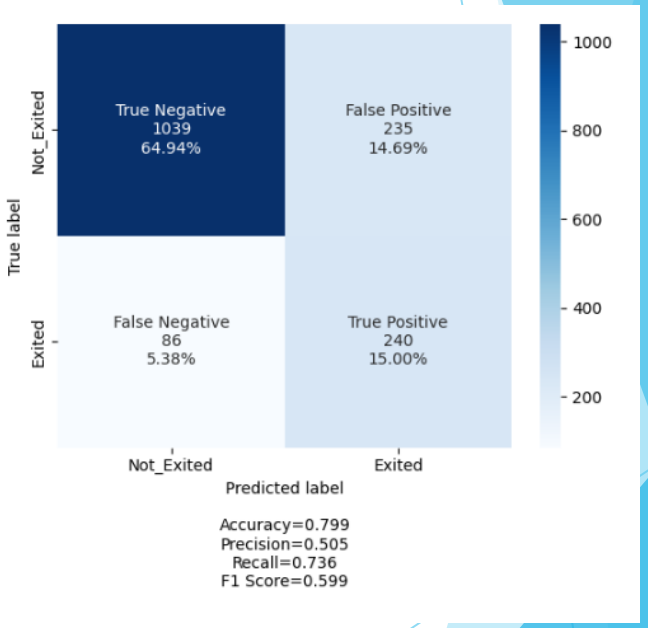


- We also tuned the model using ROC-AUC to locate the threshold for optimal balance between false positive and true positive rates. The result ran on the best threshold of 0.193 returned the predictive model that gave a better Recall score of **0.718**

Performance Summary use Dropout (Model 3)

- From the previous model loss function report, the Train and Validation curves seem to show overfitting and started to diverge, so we tried using the Dropout technique to simply the model by removing 20% neuron in every layer

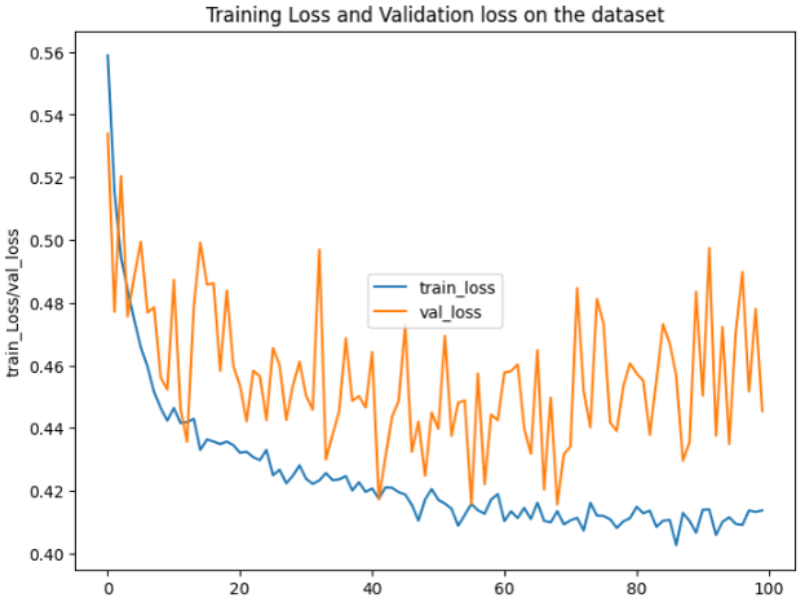


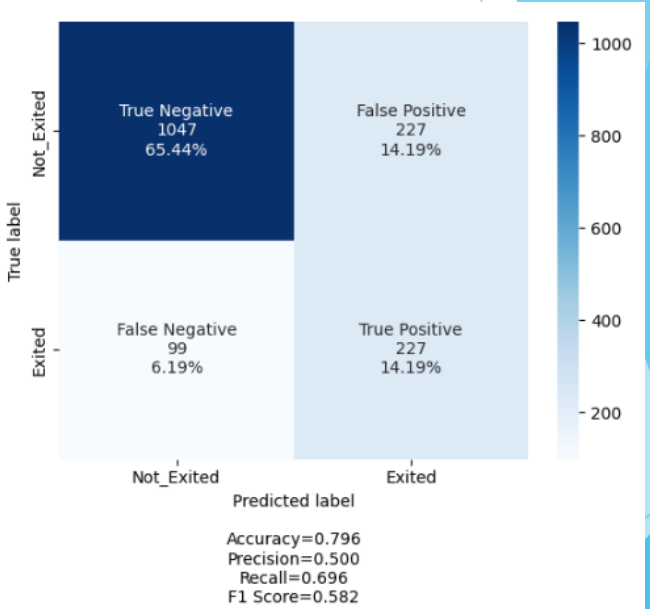


- From the above plot, we observed that the train and validation curves are smooth.
- We also tuned the model using ROC-AUC to locate the threshold for optimal balance between false positive and true positive rates. The result ran on the best threshold of 0.198 returned the predictive model that gave an improved Recall score of **0.736**

Performance Summary with Hyperparameter Grid Search tuning (Model 4)

- We further improved the model performance by hyperparameter tuning using Grid Search technique, optimized two hyperparameters called batch size, epochs.
- With batch size of [40, 64, 128] at a learning rate of [0.001, 0.0001, 0.1], after hyperparameter tuning it settled on the best batch size of 40 and learning rate

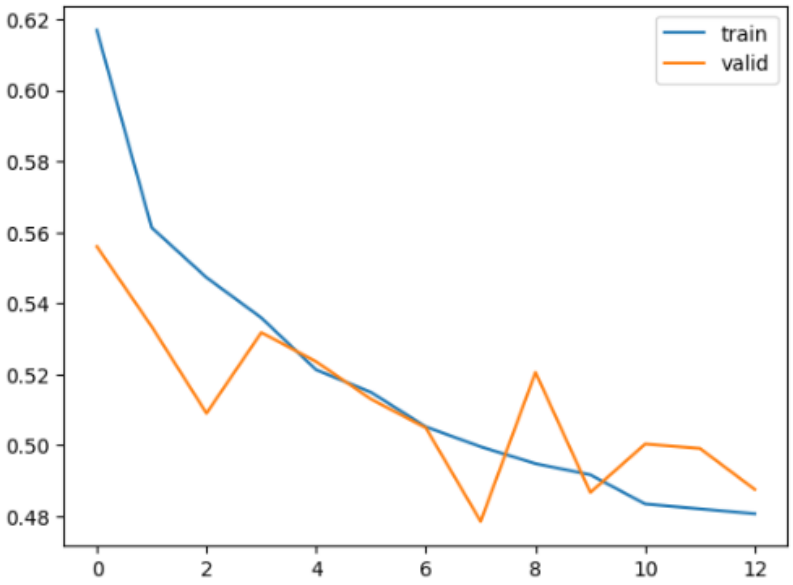


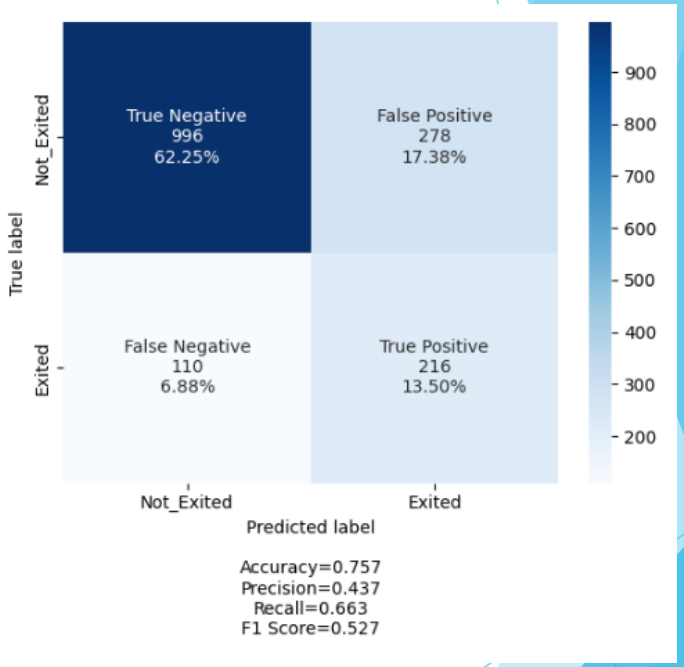


- From the above plot, we observed that there is a lot of noise in the model. We also tuned the model using ROC-AUC to locate the best threshold of 0.509,
- The result returned the predictive model that gave an dropped Recall score of **0.696**

Performance Summary with SMOTE tuning (Model 5)

- Because the previous model was severed overfitted and a lot of noise, we tried to improve the model performance by applying SMOTE to balance the dataset and then again applied hyperparameter tuning accordingly.

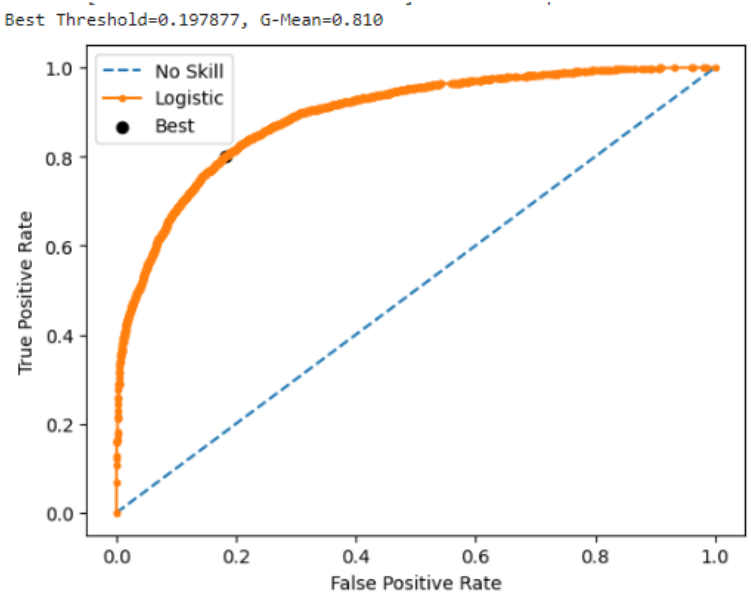


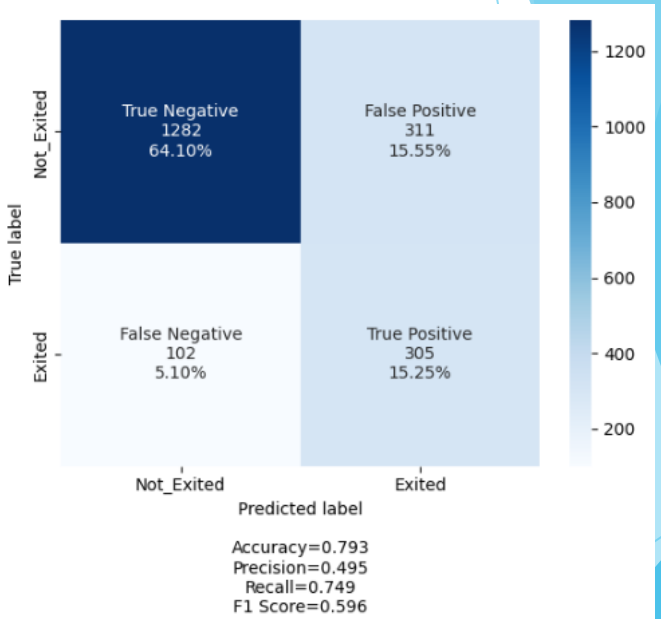


- Grid Search CV also does not seem to work that well on the SMOTE data.
- In this dataset, the SMOTE oversampling technique does not work well, as both the models we tried building have overfitted on the training dataset.
- The Recall score further dropped to **0.663**

Final Model Performance Summary

- We have built and tuned 5 models with different optimizer, decision threshold, different layer of neurons as well as other hyperparameters tuning.
- Model 3 has the best performance and therefore we selected model 3 to predict probabilities





- By using ROC-AUC to locate the best threshold of 0.1979
- The result returned the predictive model that gave a **best Recall score of 0.749**

Model Comparison and Final Model Selection

- After building 5 models, it was observed that both the Neural Network model with Dropout, exhibited strong performance on both the training and validation datasets.
- We have tuned these 4 models using different optimizer, tune the decision threshold, increase the layers and configure some other hyperparameters accordingly to improve the model's performance. The result is as followings:

Training Methods	Validation Performance (Recall Score)
Neural Network (base model)	0.399
NN model with Adam Optimizer & Early Stopping	0.718
NN model with Dropout	0.736
NN model with Grid Search Hyperparameter Tuning	0.696
NN model with SMOTE Hyperparameter Tuning	0.663

- Neural Network model trained with Dropout to generalize performance got the best recall score, therefore it's considered to be the best and final model. Model performance is then checked on unseen test data.
- Neural Network model trained with Dropout ran on unseen test data gave a **0.749** recall score which is the best Recall score.