Bank Churn Prediction

Project : Bank Churn Prediction
Course : Artificial Neural Networks

Document Version : 1.0
Document Owner : Rahul Kulkarni
Document ID : Project 5 - ANN – Bank Churn Prediction Project.pdf
Submission Date : 14th October 2023

Contents

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview & Analysis
- EDA - Univariate & Bivariate Analysis
- Data Preprocessing
- Model Building & Performance Summary
- Conclusion
- Key Actionable Business Insights
- Our Recommendation
- Appendix

Executive Summary

Executive Summary – Business Context

- **Business Context:** Businesses like banks that provide service have to worry about the problem of '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 then concentrate efforts on the improvement of service, keeping in mind these priorities
- **The Problem Statement :** Given a Bank customer, build a neural network-based classifier that can determine whether they will leave or not in the next 6 months.
- **Solution Approach:** In order to resolve the above problem, we will undertake the following key tasks:
  - Perform a deep-dive on the Bank's Churn dataset using libraries such as numpy and pandas for data manipulation, and seaborn and matplotlib for data visualisation
  - Perform exploratory data analysis on the dataset to deliver key findings and insights
  - Identify key customer attributes of the dataset that are most significant in driving churn
  - Build a classification model using a neural network that will be able to predict whether a customer will leave or not
  - Identify the key services that would need improvisation in order to lower customer churn
  - Recommend opportunities for improvement that will help the bank to boost customer retention and potential acquisition

Executive Summary – Model Evaluation Criteria & Approach

- **Model Evaluation Criteria :** The primary objective for building the model is to predict whether an existing customer will leave or not in the next 6 months and the key reasons for leaving the Banks' Services. Using the confusion matrix as guiding principle, it is imperative to focus on reducing the False Negatives (FN) i.e., predicting that a customer will not leave, but eventually leaves the Bank. Losing an existing customer would be a significant loss of revenue to the Bank. So, if FN is high, that means the churn will be high. This implies that **reducing False Negatives** should be of utmost importance to the business
  - Key Criteria – Recall: The bank should therefore use Recall as the key model evaluation criteria – higher the Recall, greater are the chances of minimising False Negatives
- **Model Building Approach:** We have split the data into Training, Validation and Testing datasets. We have built an initial model using SGD as an optimizer and evaluated its performance (Recall) on the training and validation dataset. To improve performance, we have built 4 additional models and identified their optimal thresholds using the ROC-AUC curves. Using these thresholds, we have evaluated the models' performance (Recall) on the training and validation datasets. Based on all the **Recall** scores, we have **finalised** the **best performing** model **(Model 2 : Neural Network model with Adam as an optimizer)**. We have then evaluated this model's performance on the Testing dataset
  - Model 1 : Neural Network model with SGD as an optimizer (Initial Model)
  - Model 2 : Neural Network model with Adam as an optimizer
  - Model 3 : Neural Network model with Dropout & Adam optimizer
  - Model 4 : Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer
  - Model 5 : Neural Network model with Balanced Data by applying SMOTE & Adam optimizer

Executive Summary – Model Feature Comparison Summary

- Following is the feature comparison summary of all models that were used on the training and validation dataset

#	Model Type			Activation Functions Used			No of Neurons per Layer			Optimizer Attributes			
		Total Layers	Total Params	Input Layer	Hidden Layer	Output Layer	Input Layer	Hidden Layer	Output Layer	Type	Loss	Metric	Optimal Threshold
1	Model 1 : Neural Network model with SGD as an optimizer	3 1 x Input 1 x Hidden 1 x Output	2881	Relu	Relu	Sigmoid	64	32	1	SGD	Binary Cross Entropy	Accuracy	0.5
2	Model 2 : Neural Network model with Adam as an optimizer	3 1 x Input 1 x Hidden 1 x Output	2881	Relu	Relu	Sigmoid	64	32	1	Adam	Binary Cross Entropy	Accuracy	0.167119
3	Model 3 : Neural Network model with Dropout & Adam optimizer	4 1 x Input 2 x Hidden 1 x Output	1057	Relu	Relu	Sigmoid	32	16 & 8	1	Adam	Binary Cross Entropy	Accuracy	0.220600
4	Model 4 : Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer	3 1 x Input 1 x Hidden 1 x Output	2881	Relu	Relu	Sigmoid	64	32	1	Adam	Binary Cross Entropy	Accuracy	0.255846
5	Model 5 : Neural Network model with Balanced Data by applying SMOTE & Adam optimizer	4 1 x Input 2 x Hidden 1 x Output	1057	Relu	Relu	Sigmoid	32	16 & 8	1	Adam	Binary Cross Entropy	Accuracy	0.486688

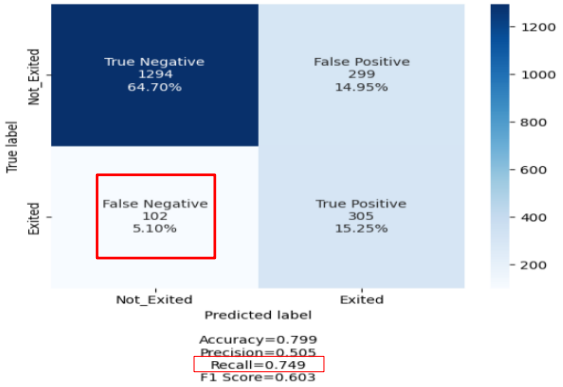
Executive Summary – Model Performance Comparison Summary

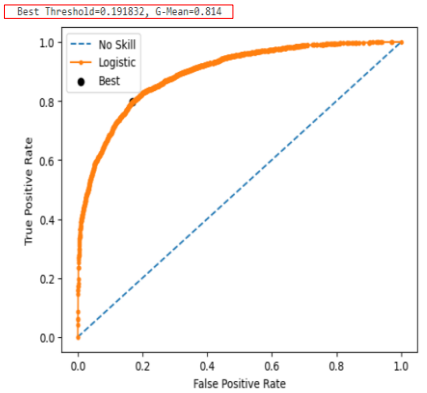
- Following is the performance metric summary of all models on the Training & Validation datasets. Model 2 : Neural Network model with Adam as an optimizer has delivered a Recall score of 76.7% and False Negative of 4.75% (76), whereas Model 3 : Neural Network model with Dropout & Adam optimizer has delivered a Recall score of 74.8% and False Negative of 5.12% (82) on the Training & Validation dataset. Based on our key criteria Recall score**, Model 2 : Neural Network model with Adam as an optimizer has the best performance followed by Model 4: Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer**

#	Model Type	Key Performance Metrics				Confusion Matrix Scores			
		Accuracy	Recall	Precision	F1	True Positive	True Negative	False Positive	False Negative
1	Model 1 : Neural Network model with SGD as an optimizer	0.849	0.371	0.771	0.501	7.56% 121	77.38% 1238	2.25% 36	12.81% 205
2	Model 2 : Neural Network model with Adam as an optimizer (1)	0.781	0.767	0.477	0.588	15.62% 250	62.50% 1000	17.12% 274	4.75% 76
3	Model 3 : Neural Network model with Dropout & Adam optimizer (2)	0.805	0.748	0.515	0.610	15.25% 244	65.25% 1044	14.37% 230	5.12% 82
4	Model 4 : Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer	0.799	0.706	0.504	0.588	14.37% 230	65.50% 1048	14.12% 226	6.00% 96
5	Model 5 : Neural Network model with Balanced Data by applying SMOTE & Adam optimizer	0.766	0.721	0.454	0.557	14.69% 235	61.94% 991	17.69% 283	5.69% 91

Executive Summary – Best & Final Model : Adam as an Optimizer

- **Best Model:** Based on the performance comparison of all the 5 models, Model 2 : Neural Network model with Adam as an optimizer has delivered a Recall score of 76.7% and False Negative of 4.75% (76) on the Training & Validation dataset. We have therefore considered this as our best candidate model and used it on the Testing dataset
- **KPI Comparison:** The model has delivered a Recall score of 0.767 with a False Negative of 4.75% (76) on the Training & Validation dataset, and a Recall score of 0.749 with a False Negative of 5.10% (102) on the Testing dataset
- **Final Model - Model 2 : Neural Network model with Adam as an optimizer :** With a Recall score of 74.9% and a False Negative of 5.10% (102) on the Testing dataset using an optimal threshold of 0.191832 and a G-Mean of 0.814, the Model 2 : Neural Network model with Adam as an optimizer has generalised its performance and can be considered as the final model





	Key Performance Metrics				Confusion Matrix Scores			
	Accuracy	Recall	Precision	F1	True Positive	True Negative	False Positive	False Negative
Training & Validation Dataset	0.781	0.767	0.477	0.588	15.62% 250	62.50% 1000	17.12% 274	4.75% 76
Testing Dataset	0.799	0.749	0.505	0.603	15.25% 305	64.70% 1294	14.95% 299	5.10% 102

Executive Summary - Conclusion

- We have built an initial model using SGD as an optimizer and evaluated its performance (Recall) on the training and validation dataset. To improve the model performance, we have built 4 additional models (below), identified their optimal thresholds using the ROC-AUC curves. Using these thresholds, we have evaluated the models' performance (Recall) on the training and validation datasets.
  - Model 1 : Neural Network model with SGD as an optimizer (Initial Model)
  - Model 2 : Neural Network model with Adam as an optimizer
  - Model 3 : Neural Network model with Dropout & Adam optimizer
  - Model 4 : Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer
  - Model 5 : Neural Network model with Balanced Data by applying SMOTE & Adam optimizer
- Based on the performance comparison (Recall & False Negative) of all the models, we have **finalised** the **best performing** model (Model 2 : Neural Network model with Adam as an optimizer). We have then evaluated this model's performance on the Testing dataset. This model has delivered a Recall score of 0.767 & 0.749, and a False Negative of 4.75% (76) and 5.10% (102) on the Training & Validation, and the Testing dataset respectively. The model's Recall score on the Testing dataset is inline to the one observed on the Training & Validation datasets. This will minimise False Negatives, which is of utmost importance to the business
- Using an optimal threshold of 0.191832 and a G-Mean of 0.814, the Model 2 : Neural Network model with Adam as an optimizer has delivered a Recall score of 74.9% and a False Negative of 5.10% (102) on the Testing dataset. This model has generalised its performance and can be considered as the final model.
- The model built can be used to predict customer churn i.e., whether a customer will leave or not. This will help the bank to target the potential customers, who have a higher probability of leaving, and proactively incentivise them in order to maximise customer retention
- The bank should focus on improving key services that are related to isActiveMember, NumOfProducts, HasCreditCard features. Improving these services will lead to customer retention (Tenure), thereby driving growth (Balance) and potential customer acquisition in new areas (Geography)

Executive Summary - Key Actionable Business Insights

- **Summarised Key Observations & Insights :**
  - There is a significant imbalance in the dataset since there are high number of existing customers (79.6%) than compared to 20.4% of attrited customers
  - The attrition levels are higher in customers who are inactive i.e., not transacting on a regular basis, than compared to customers who are actively transacting with the bank. The bank's customer service should proactively reach out to inactive members to identify and resolve the key pain-points. This might lead to a positive conversion (inactive to active) and would help customer retention
  - There is a positive correlation between Balance and Exited, which implies that as the account balance increases the customer has a high probability of the leaving the bank. This could imply that the bank probably doesn't offer a great return on investment (ROI), which does not incentivise the customer to continue the relationship. The bank's finance strategy should focus on Investments vs ROI in order to retain customers
  - There is a positive correlation between Estimated Salary and Exited, which implies that as the customer salary increases, higher is the probability of the customer leaving the bank. This could imply that bank probably doesn't offer great services to high net-worth individuals (HNI), which again does not incentivise customers to continue the relationship. The bank should provide the customers with various incentives and reward schemes, especially to high net-worth individuals, in order to enhance relationship and retain customers. The banks' customer strategy should focus on building customer relationship and providing high-end banking services in order to retain HNI customers
  - There are higher number of customers leaving the bank in Germany than compared to Spain and France. Also, there are higher number of female customers leaving the bank than compared to male customers. The bank's customer strategy should focus on a diversity and initiate a customer feedback program to identify the key challenges and take appropriate measure to resolve these challenges
  - In order to retain existing customers and acquire potential prospects, the bank should incentivise customers with cashback schemes and loyalty reward points, that can be redeemed on future purchases on using their banking products e.g. using credit cards to earn cashbacks, which might encourage customers on using their credit cards more often

Executive Summary - Our Recommendation

Based on the key observations and insights, we recommend the following areas of improvement / opportunities that will drive business growth and lead to a better customer experience

- **Implement Customer Incentivisation Scheme:** Incentivising customers by offering them cashback schemes and discounts / vouchers on purchases will encourage frequent spending and will drive customer gowth and increase revenue
- **Implement Customer Satisfaction Survey:** The bank should initiate a targeted Customer Satisfaction Survey to understand customer pain points and implement the findings to improve retention ratio of such customers
- **Implement Tier based Rewards:** The bank should introduce a Tier based Loyalty & Rewards Scheme for purchases using their banking products e.g. Credit / Debit Cards. Cumulative loyalty points above a certain threshold will promote the customer to a new tier, that will offer specific rewards such as First-Class Lounge access at Airports, Spa & Well-Being discounts etc

Business Problem Overview & Solution Approach

Business Context

- **Business Context:** Businesses like banks that provide service have to worry about the problem of '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 then concentrate efforts on the improvement of service, keeping in mind these priorities
- **The Problem Statement :** Given a Bank customer, build a neural network-based classifier that can determine whether they will leave or not in the next 6 months.
- **Solution Approach:** In order to resolve the above problem, we will undertake the following key tasks:
  - Perform a deep-dive on the Bank's Churn dataset using libraries such as numpy and pandas for data manipulation, and seaborn and matplotlib for data visualisation
  - Perform exploratory data analysis on the dataset to deliver key findings and insights
  - Identify key customer attributes of the dataset that are most significant in driving churn
  - Build a classification model using a neural network that will be able to predict whether a customer will leave or not
  - Identify the key services that would need improvisation in order to lower customer churn
  - Recommend opportunities for improvement that will help the bank to boost customer retention and potential acquisition

Data Overview & Analysis

Data Overview & Analysis

- The Bank Churners dataset has the following Data-Structure:

#	Columns	Data-type	Total Rows	Description
1	RowNumber	Integer 64	10,000	Row Number
2	CustomerId	Integer 64	10,000	Unique ID which is assigned to each customer
3	Surname	Object	10,000	Last name of the customer
4	CreditScore	Integer 64	10,000	It defines the credit history of the customer.
5	Geography	Object	10,000	A customer's location
6	Gender	Object	10,000	It defines the Gender of the customer
7	Age	Integer 64	10,000	Age of the customer
8	Tenure	Integer 64	10,000	Number of years for which the customer has been with the bank
9	Balance	Float 64	10,000	Account balance
10	NumOfProducts	Integer 64	10,000	It refers to the number of products that a customer has purchased through the bank
11	HasCrCard	Integer 64	10,000	It is a categorical variable that decides whether the customer has a credit card or not.
12	IsActiveMember	Integer 64	10,000	It is a categorical variable that decides whether the customer is an active member of the bank or not
13	EstimatedSalary	Float 64	10,000	Estimated salary
14	Exited	Integer 64	10,000	It is a categorical variable that decides whether the customer left the bank within six months or not.

Data Overview & Analysis (Cont'd)

- Shape of the Dataset: Total No. Of Columns - 14 | Total No. Of Rows - 10,000
- Column Data-types: 11 of 14 columns are of numerical data types – Float 64(2) & Integer 64(9) , and 3 are Object data types
- Missing & Duplicate Values: There are no missing values or duplicate values in the dataset
- Statistical Summary & Unique Values: Following is the statistical summary of the dataset

	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

Unique Values in Dataset

```
RowNumber          10000
CustomerId         10000
Surname             2932
CreditScore          460
Geography              3
Gender                 2
Age                   70
Tenure                11
Balance             6382
NumOfProducts          4
HasCrCard              2
IsActiveMember         2
EstimatedSalary     9999
Exited                 2
dtype: int64
```

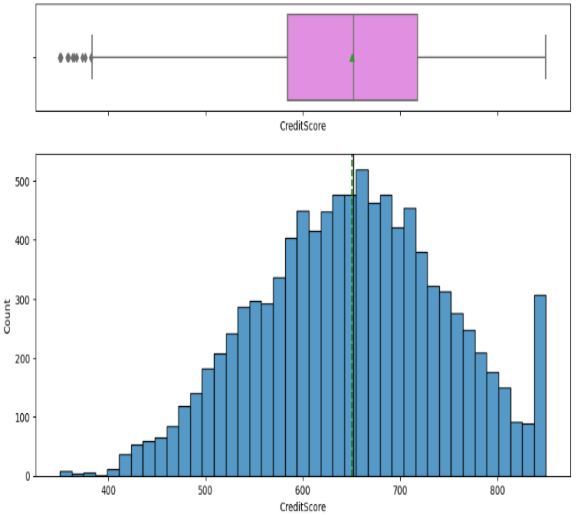
Data Overview & Analysis – Key Observations & Insights

- **Key Observations & Insights:**
  - The minimum and maximum Customer Age is 18 years and 92 years respectively, whereas the mean Age is 39 years
  - The minimum and maximum number of Tenure is 0 and 10 respectively, whereas the mean and median Tenure is approximately 5 years
  - The minimum and maximum customer balance is $0 and $250K respectively, whereas the mean balance is $76K and the median balance is $97K
  - The minimum and maximum number of products is 1 and 4 respectively, whereas the mean number of products is 1.5 and the median number of products is 1
  - Circa. 7,055 customers have a credit than compared to 2,945 that do not have a credit card
  - Circa. 5,151 customers are active members of the bank i.e. transacting with the bank regularly than compared to 4,849 that inactive member
  - The minimum and maximum estimated salary is $11.58 and $200K respectively, whereas the mean and median estimated salary is approximately the same - $100K
  - Circa. 2,037 customers have exited the bank whereas 7,963 are still with the bank
  - RowNumber, CustomerId, and Surname columns do not add any value and hence should be dropped from the dataset

EDA - Univariate Analysis

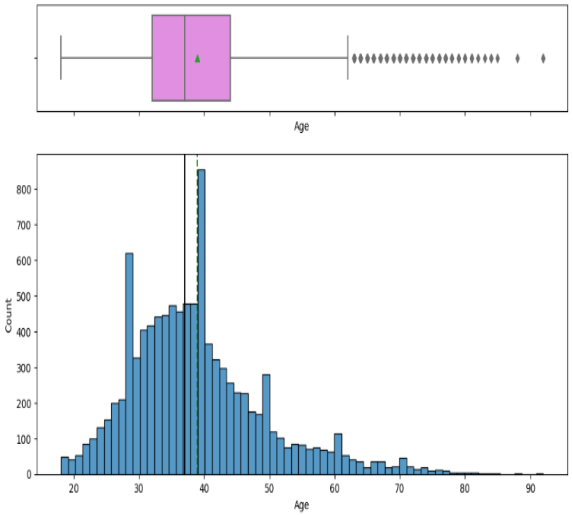
EDA - Univariate Analysis – Credit Score

- **Credit Score:**
  - Minimum: The minimum credit score is 350
  - Q1: 25% of the customers have a credit score of less than 584
  - Q3: 75% of the customers have a credit score of less than 718
  - Maximum: The maximum credit score is 850
  - Median: The median score is 652
  - Mean: The mean credit score is 650.52
  - Outliers: There are a few outliers having a credit score between 350-382
  - Skewness: From the plot, it can be observed that it has a near to normal distribution



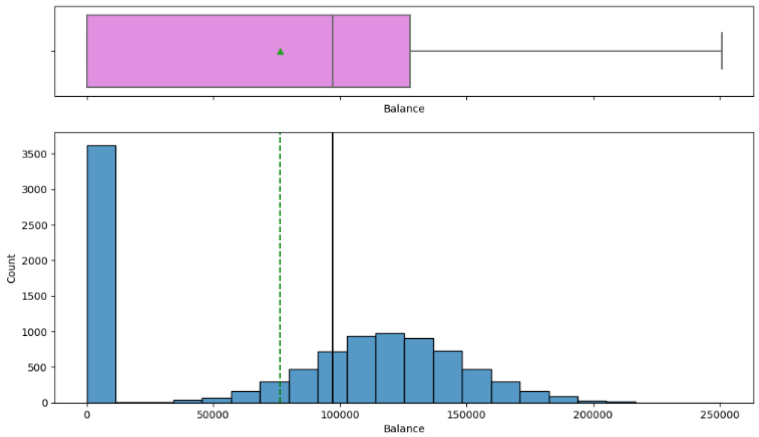
EDA - Univariate Analysis – Age

- **Age:**
  - Highest Age Group: There are circa 477 customers 38 years of age
  - Minimum: The minimum age is 18 years
  - Q1: 25% of the population are less 32 years of age
  - Q3: 75% of the population are less 44 years of age
  - Maximum: The maximum age is 92 years
  - Median & Mean: The median age is 37 years, and the mean age is approx. 39 years
  - Outliers: There are several outliers ranging from 63 to 92 years of age
  - Skewness: From the plot, it can be observed that the graph has near to normal distribution



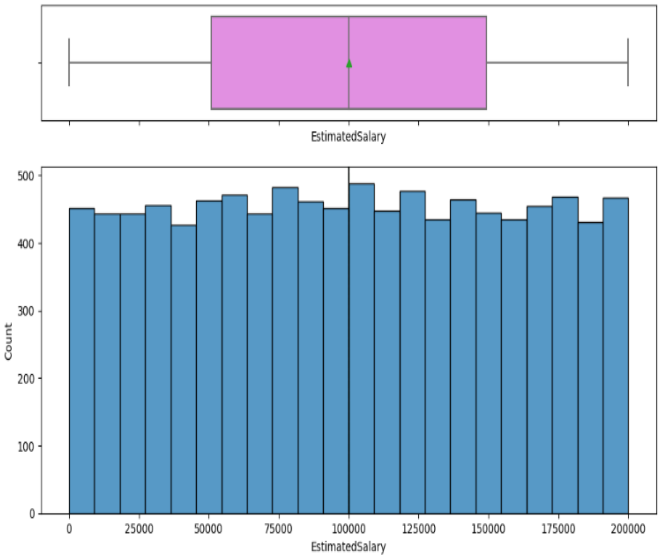
EDA - Univariate Analysis – Balance

- **Balance:**
  - Minimum: The minimum account balance is $0
  - Q1: 25% of the customers have a balance of less than $0
  - Q3: 75% of the customers have a balance of less than $127K
  - Maximum: The maximum balance is circa $250K
  - Median: The median balance is circa $97K
  - Mean: The mean balance is circa $76K
  - Outliers: There are no outliers



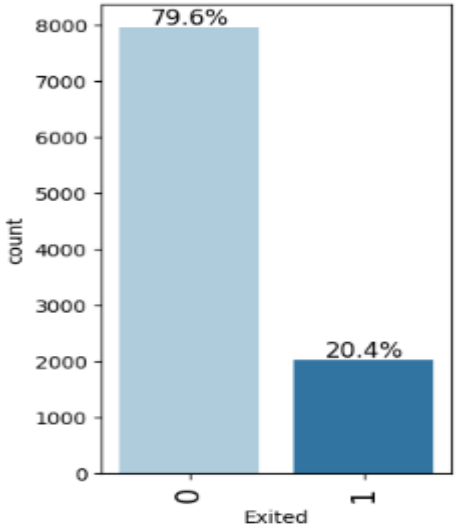
EDA - Univariate Analysis – Estimated Salary

- **Estimated Salary :**
  - Minimum: The minimum estimated salary is $11.58
  - Q1: 25% of the customers have an estimated salary of circa less than $51K
  - Q3: 75% of the customers have an estimated salary of circa less than $149K
  - Maximum: The maximum estimated salary is circa $200K
  - Mean & Median: The median and mean estimated salary is circa $100K
  - Outliers: There are no outliers
  - Skewness: From the plot, it can be observed that the graph has a normal distribution



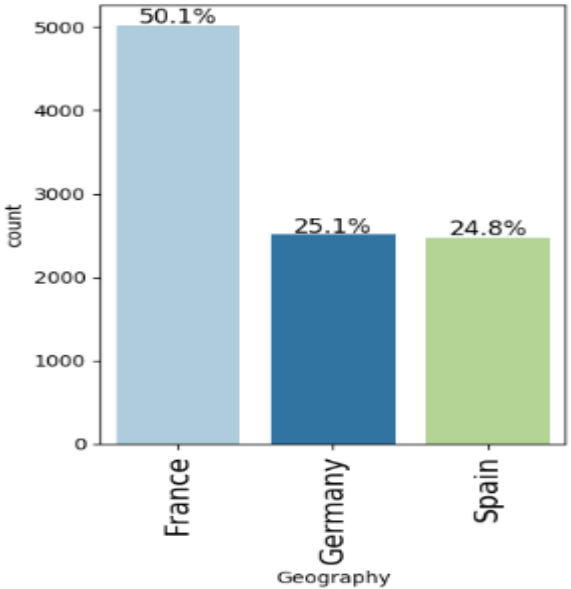
EDA - Univariate Analysis – Exited

- **Exited:**
  - There is a significant imbalance in data since there are high number of existing customers than compared to those that have attrited
  - Approx. 79.6% (7,963) are existing that compared to 20.4% (2,037) customers that have left the bank



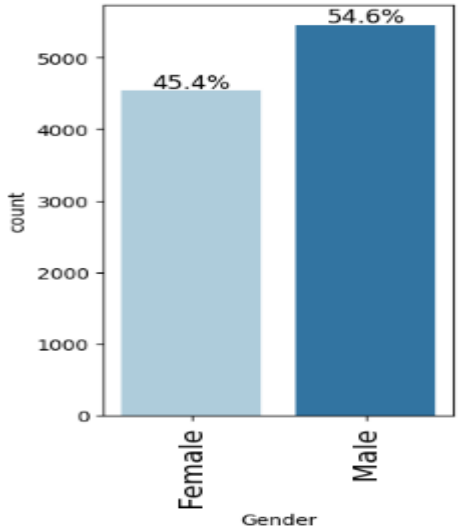
EDA - Univariate Analysis – Geography

- **Geography :**
  - There are higher number of customers in France than compared to Germany and Spain
  - There are approx. 50.01% (5,014) customers in France, 25.1%(2,509) in Germany and 24.8%(2,477) in Spain
  -



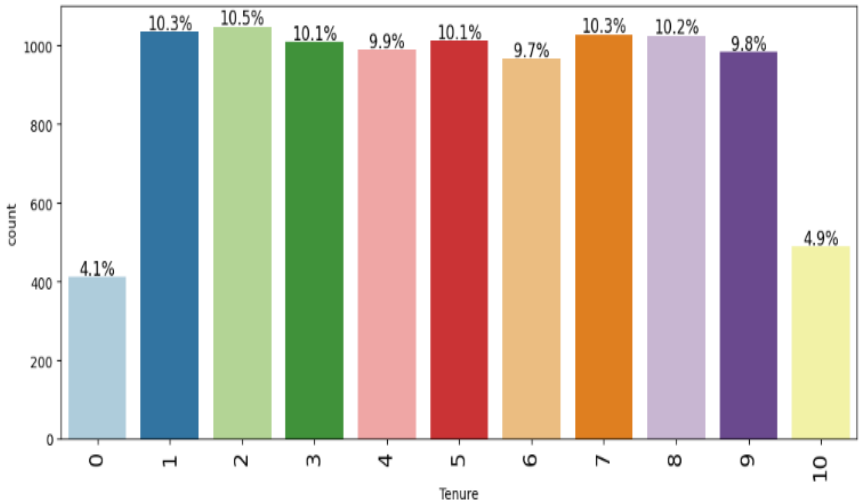
EDA - Univariate Analysis – Gender

- **Gender :**
  - There are higher number of male customers than female customers
  - There are approx. 54.6% (5,457) male customers than compared to 45.4%(4,543) female customers
  -



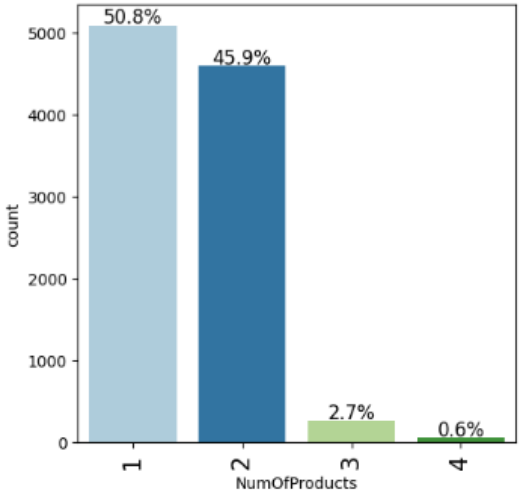
EDA - Univariate Analysis – Tenure

- **Tenure :**
  - Only 4.9% (490) customers have been with the bank for 10 years
  - 9.8% (984) customers have been with the bank for 9 years
  - 10.2% (1,025) customers have been with the bank for 8 years
  - 10.3% (1,028) customers have been with the bank for 7 years
  - 9.7% (967) customers have been with the bank for 6 years
  - 10.1% (1,012) customers have been with the bank for 5 years
  - 9.9% (989) customers have been with the bank for 4 years
  - 10.1% (1,009) customers have been with the bank for 3 years
  - 10.5% (1,048) customers have been with the bank for 2 years
  - 10.3% (1,035) customers have been with the bank for 1 years
  - 4.1% (413) customers have been with the bank for less than 1 year



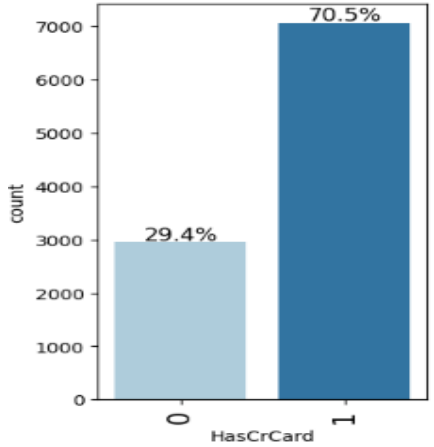
EDA - Univariate Analysis – Number Of Products

- **Number Of Products :**
  - Circa 95% of customers have less than 3 products
  - Circa 50.8% (5,084) customers have 1 product, 45.9% (4,590) have 2 products, 2.7% (266) customers have 3 products, and 0.6% (60) customers have 4 products



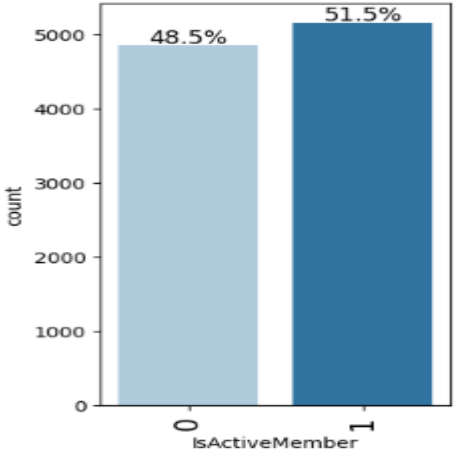
EDA - Univariate Analysis – Has Credit Card

- **Has Credit Card:**
  - Majority of customers have a credit card
  - Circa. 70.5% (7,055) of customers have a credit card than compared to 29.4% (2,945) customers that do not have a credit card



EDA - Univariate Analysis – Is Active Member

- **Is Active Member :**
  - Approximately half the customers are actively transacting with the bank, whereas the other half are inactive
  - Circa 51.5% (5,151) of customers have a credit card than compared to 48.5% (4,849) customers that do not have a credit card



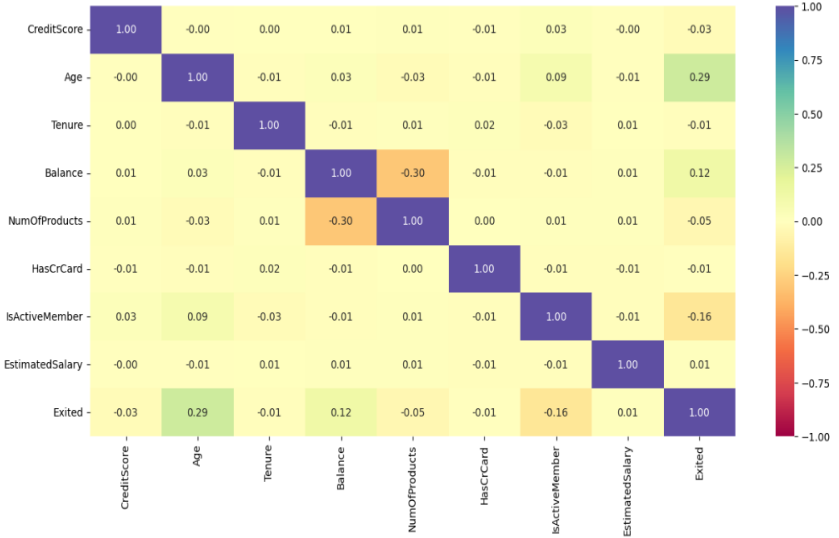
EDA - Univariate Analysis – Key Observations & Insights

- **Key Observations & Insights:**
  - The minimum and maximum credit score is 350 and 850 respectively, whereas the mean and median credit score is approximately the same
  - The minimum and maximum customer age is 18 and 92 respectively, whereas the mean and median age is approximately similar
  - The minimum and maximum customer balance is $0 and $250K respectively, whereas the mean balance is $76K and the median balance is $97K
  - The minimum and maximum estimated salary is $11.58 and $200K respectively, whereas the mean and median estimated salary is approximately the same
  - There is a significant imbalance in data since there are high number of existing customers than compared to those that have attrited
  - There are higher number of customers in France than compared to Germany and Spain
  - There are higher number of male customers than female customers
  - Only 4.9% (490) customers have been with the bank for 10 years, whereas only 4.1% (413) customers have been with the bank for less than 1 year
  - Circa 95% of customers have less than 3 products
  - Majority of customers have a credit card
  - Approximately half of the customers are actively transacting with the bank, whereas the other half of the customers are inactive

EDA - Bivariate Analysis

EDA - Bivariate Analysis – Correlation Check

- **Correlation Amongst Variables :**
  - There is a weak positive correlation of 0.29 between Age and Exited, which implies that as age increases the likelihood of customers exiting the services increases
  - There is a weak negative correlation of -0.30 between Balance and NumOfProducts, which implies that as the number of products increases the balance decreases. It is possible that the customer has bought various other products from the bank such as fixed deposits, mutual funds, credit cards etc or invested in other banking financial assets sold by the bank. This could have potentially led to the depletion the account balance. On the contrary, lesser the number of products purchased from the bank, higher is the customer balance



- There is an insignificant negative correlation of -0.16 between isActiveMember and Exited, which implies that customers who aren't actively transacting with the bank are more likely to exit the services compared to those who are are actively transacting
- There is an insignificant positive correlation of 0.12 between Balance and Exited, which implies that as the account balance increases the customer has high probability of the leaving the banking services. This could imply that the Bank probably doesn't offer a significant return on investment (ROI) or provides poor services to high net-worth individuals

EDA - Bivariate Analysis – Correlation Check (Cont'd)

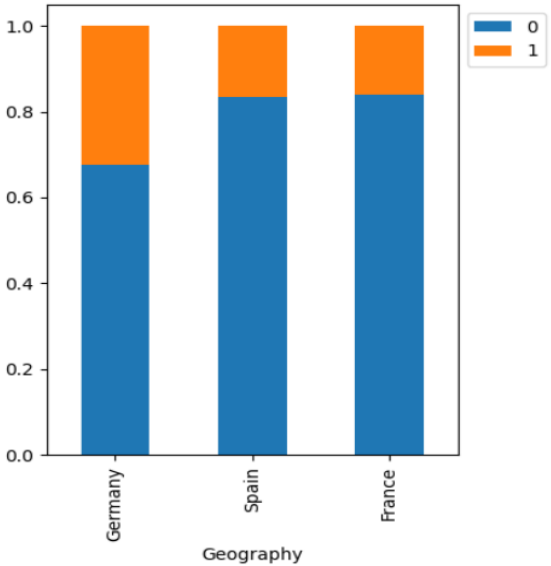
- **Correlation Amongst Variables (Cont'd) :**
  - There is an insignificant negative correlation of -0.03 between CreditScore and Exited, which implies that as the credit score of the customer increases, the probability of the customer leaving the bank decreases. There is also an insignificant positive correlation of 0.01 between Balance, Number of Products and CreditScore. Considering all the factors above, implies that certain customer have a good ROI, which increases their account balance, and hence customers would be investing into other banking products to diversify their portfolio. This would lead to higher credit worthiness
  - There is an insignificant negative correlation of -0.05 between NumOfProducts and Exited, which implies that bigger the banking portfolio lesser are the chances of the customer's leaving the bank and vice-versa. There is also an insignificant positive correlation of 0.01 between isActiveMember, Estimated Salary and NumOfProducts. Considering all the factors above, implies that high salaried individuals tend to transact regularly with the bank to diversify their portfolio by purchasing new banking products
  - There is an insignificant negative correlation of -0.01 between Tenure and Exited, which implies higher the tenure, lesser are the chances of leaving the bank and vice-versa
  - There is an insignificant negative correlation of -0.01 between Balance and Exited. This could imply that a select few customer are having a good return on investment and hence choose to stay with the bank as the account balance increases
  - There is an insignificant negative correlation of -0.01 between HasCard and Exited, which implies that customers with credit cards are less likely to leave the bank and vice versa
  - There is an insignificant positive correlation of 0.01 between EstimatedSalary and Exited, which implies that higher the salary, greater are the chances of customers leaving the bank and vice-versa. This could mean that the bank would not be offering a high return on investment

EDA - Bivariate Analysis – Correlation Check (Cont'd)

- **Correlation Amongst Variables (Cont'd) :**
  - There is an insignificant positive correlation of 0.09 between isActiveMember and Age, which implies that higher the age, more often does the customer transact with the bank and vice-versa. There is also an insignificant positive correlation of 0.03 between Balance and Age, which implies that customers tend to save more / have a higher account balance as age increases
  - There is an insignificant negative correlation of -0.01 between HasCard, Tenure and Age, which implies that higher the age, lesser are the chances of the customers having a credit card and lesser is the tenure with the bank i.e. higher probability of exiting the bank
  - There is an insignificant positive correlation of 0.03 between isActiveMember and Creditscore, which implies that higher the banking activity i.e. transactions with the bank, greater is the credit score and vice-versa

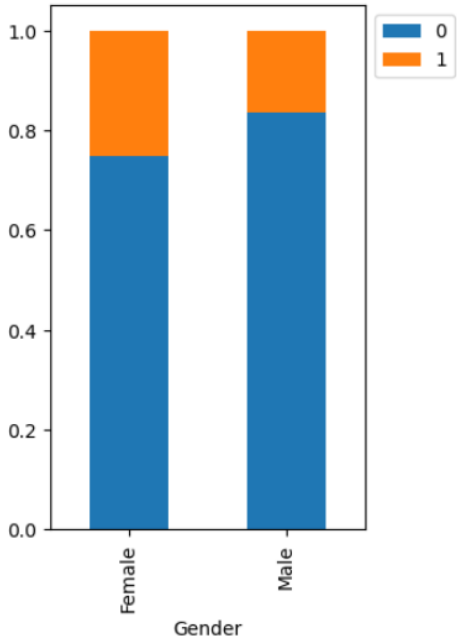
EDA - Bivariate Analysis – Exited Vs Geography

- **Exited vs Geography:**
  - The attrition levels are higher in Germany than compared to Spain & France
  - 32.44% (814 of 2,509) of customers in Germany have attrited the bank than compared to 16.67% (413 of 2,477) of customers in Spain and 16.15% (810 of 5,014) of customers in France
  - The attrition levels in Spain and France are similar



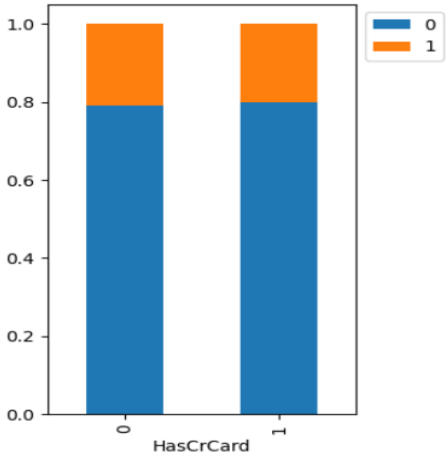
EDA - Bivariate Analysis – Exited Vs Gender

- **Exited vs Gender:**
  - The attrition levels are higher in females than compared to males
  - 25% (1,139 of 4,543) of female customers have attrited the bank than compared to 16.45% (898 of 5,457) of male customers who have left the bank



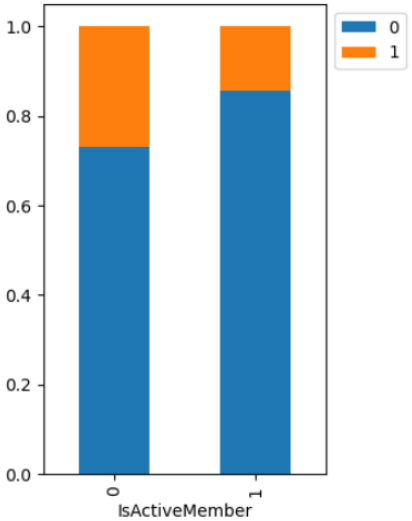
EDA - Bivariate Analysis – Exited Vs Has Credit Card

- **Exited vs Has Credit Card:**
  - The attrition levels are similar with customers having a credit as opposed to customers that do not have a credit
  - 20.81% (613 of 2,945) of customers who do not have a credit card have attrited the bank than compared to 20.18% (1,424 of 7,055) of customers that have a credit card who have left the bank



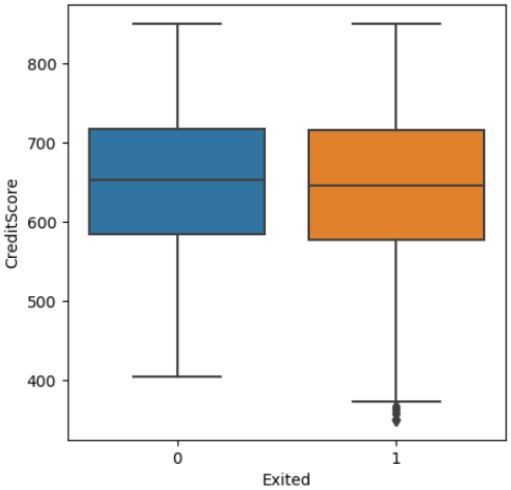
EDA - Bivariate Analysis – Exited Vs Is Active Member

- **Exited vs Is Active Member :**
  - The attrition levels are higher in customers who are inactive than compared to customers who are actively transacting with the bank
  - 26.85% (1302 of 4,849) of customers who do not have actively transact with the bank have attrited than compared to 14.26% (735 of 5,151) of customers who actively transact with the bank



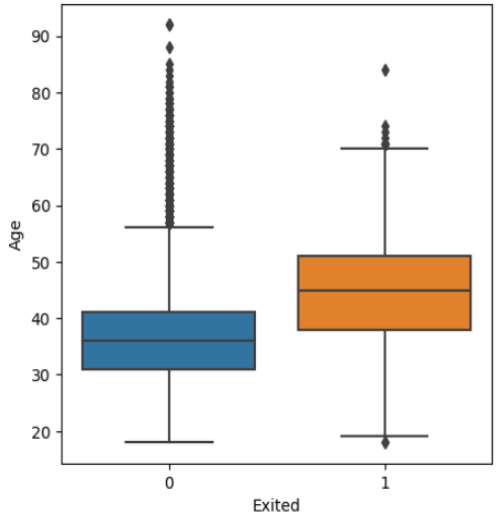
EDA - Bivariate Analysis – Exited Vs Credit Score

- **Exited vs Is Credit Score :**
  - The credit score of customers that left the bank are similar as compared to those who did not leave the bank
  - Customers that exited had a median credit score of 646 and a mean credit score of 645 as opposed to those who did not exit, who had a median credit score of 653 and a mean credit score of 651
  - Q1: 25% of customers who left the bank had a credit score of less than 578, than compared to customers who did not leave the bank and a credit score of less than 585
  - Q3: 75% of customers who left the bank had a credit score of less than 716, than compared to customers who did not leave the bank and a credit score of less than 718
  - Min: Customers who left the bank had a minimum credit score of 350 than compared to customers who did not leave the bank and minimum credit score of 405
  - Max: The maximum credit score for both - customers who left the bank and those who did not - had a maximum credit score of 850



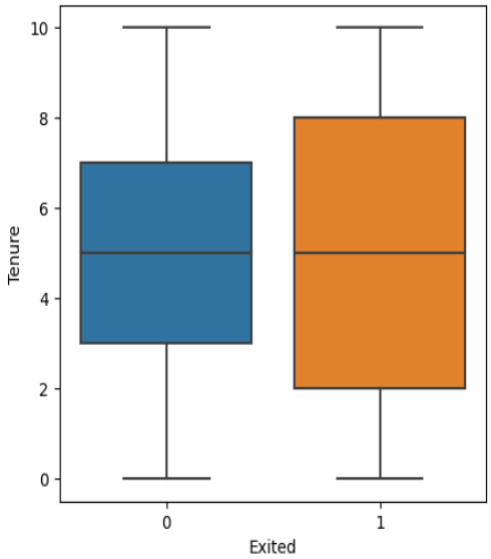
EDA - Bivariate Analysis – Exited Vs Age

- **Exited vs Is Age :**
  - Customers that exited had a median age of 45 and a mean age of 44 as opposed to those who did not exit, who had a median age of 36 and a mean age of 37
  - Q1: 25% of customers who left the bank were less than 38 years old, than compared to customers who did not leave the bank, and were less than 31 years old
  - Q3: 75% of customers who left the were less than 51 years old, than compared to customers who did not leave the bank, and were less than 41 years old
  - Min: The minimum age for both (customers who left the bank and those who did not) was 18 years of age
  - Max: The maximum age within customers who left the bank was 84, whereas the maximum age within customers who did not leave the bank was 92



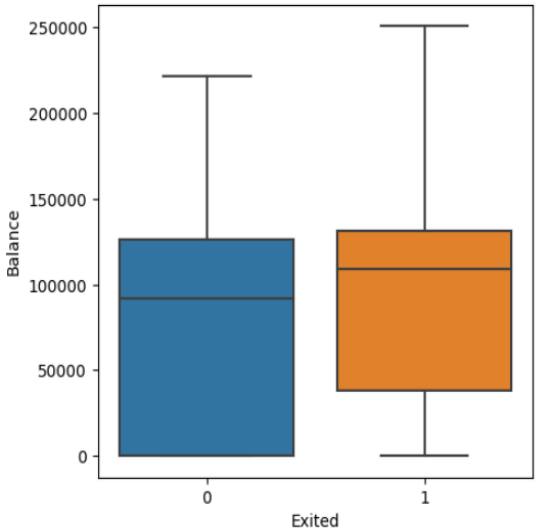
EDA - Bivariate Analysis – Exited Vs Tenure

- **Exited vs Tenure :**
  - Customers that exited the bank had an approximately a similar tenure to those who did not leave the bank
  - Customers that exited the bank had a median tenure of 5 years and a mean tenure of 4.93 years as opposed to those who did not exit, who had a median tenure of 5 years and a mean tenure of 5.03 years
  - Q1: 25% of customers who left the bank had a tenure of less than 2 years, than compared to customers who did not leave the bank, and had a tenure of less than 3 years
  - Q3: 75% of customers who left the bank had a tenure of less than 8 years, than compared to customers who did not leave the bank, and had a tenure of less than 7 years
  - Min: The minimum tenure for customers who left the bank and those who did not was the same – 0 years
  - Max: The maximum tenure for customers who left the bank and those who did not was the same – 10 years



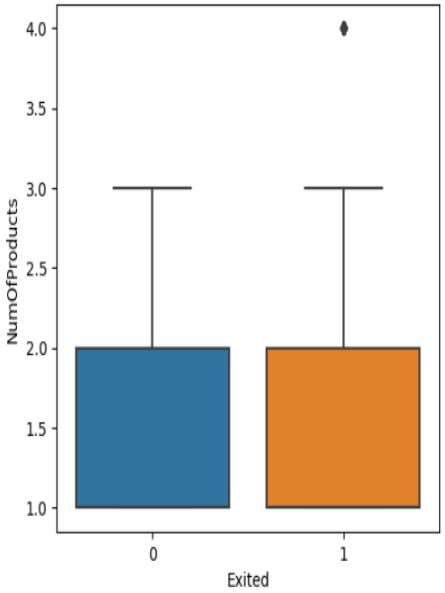
EDA - Bivariate Analysis – Exited Vs Balance

- **Exited vs Balance :**
  - Customers that had a higher balance have exited the bank than compared to those who did not leave the bank
  - Customers that exited the bank had a median balance of $109K and a mean balance of $91K as opposed to those who did not exit, who had a median balance of $92K and a mean balance of $73K
  - Q1: 25% of customers who left the bank had a balance of less than $38K, than compared to customers who did not leave the bank, and had a $0 balance
  - Q3: 75% of customers who left the bank had a balance of less than $131K, than compared to customers who did not leave the bank, and had a balance of less than $126K
  - Min: The minimum balance for customers who left the bank and those who did not was $0
  - Max: The maximum balance for customers who left the bank was circa $250K, than compared to customers who did not leave the bank was circa $221.5K



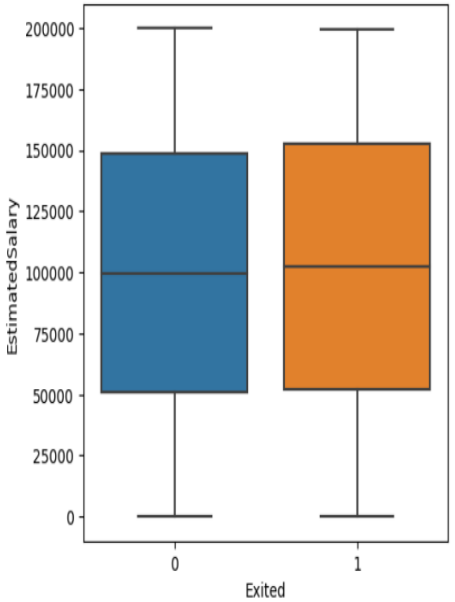
EDA - Bivariate Analysis – Exited Vs Number Of Products

- **Exited vs Number of Products :**
  - Customers who exited the bank had similar number of products compared to customers who did not leave the bank
  - Customers that exited the bank had a median of 1 product and a mean of 1.5 products as opposed to those who did not exit, who had a median of 2 products and a mean of 1.5 products
  - Q1: 25% of customers who left the bank and those who did not leave the bank had less than 1 product
  - Q3: 75% of customers who left the bank and those who did not leave the bank had less than 2 products
  - Min: Customers who left the bank and those who did not leave the bank had minimum of 1 product
  - Max: Customers who left the bank had a maximum of 4 products than compared to customers who did not leave the bank and had a maximum of 3 products



EDA - Bivariate Analysis – Exited Vs Estimated Salary

- **Exited vs Estimated Salary:**
  - Customers who exited the bank and those who did not exit the bank had a similar estimated salary
  - Customers that exited the bank had a median estimated salary of $102K and a mean estimated salary of $101K as opposed to those who did not exit, who had a median estimated salary of $99.6K and a mean estimated salary of $99.7K
  - Q1: 25% of customers who left the bank had an estimated salary of less than $52K than compared to customers who did not leave the bank and had an estimated salary of less than $51K
  - Q3: 75% of customers who left the bank had an estimated salary of less than $152K than compared to customers who did not leave the bank and had an estimated salary of less than $147K
  - Min: Customers who left the bank had an estimated salary of circa $12 than compared to customers who did not leave the bank and had an estimated salary of circa $90
  - Max: Customers who left the bank and those who did not leave the bank had a similar estimated salary of circa $200K



EDA - Bivariate Analysis – Key Observations & Insights

- **Key Observations & Insights :**
  - There are higher number of female customers leaving the bank than compared to male customers
  - The attrition levels are higher in customers who are inactive than compared to customers who are actively transacting with the bank
  - The attrition levels are similar with customers having a credit as opposed to customers that do not have a credit
  - Customers that had a higher balance have exited the bank than compared to those who did not leave the bank
  - The credit score of customers that left the bank are similar as compared to those who did not leave the bank
  - Customers that exited the bank had an approximately a similar tenure to those who did not leave the bank
  - Customers who exited the bank had similar number of products compared to customers who did not leave the bank
  - Customers who exited the bank and those who did not exit the bank had a similar estimated salary
  - There is an insignificant positive correlation of 0.01 between EstimatedSalary and Exited, which implies that higher the salary, greater are the chances of customers leaving the bank and vice-versa. This could mean that the bank would not be offering a high return on investment
  - There is an insignificant positive correlation of 0.12 between Balance and Exited, which implies that as the account balance increases the customer has high probability of the leaving the banking services. This could imply that the Bank probably doesn't offer a significant return on investment (ROI) or provides poor services to high net-worth individuals
  - There is a weak positive correlation of 0.29 between Age and Exited, which implies that as age increases the likelihood of customers exiting the services increases

EDA - Bivariate Analysis – Key Observations & Insights (Cont'd)

- **Key Observations & Insights (Cont'd) :**
  - There is a weak negative correlation of -0.30 between Balance and NumOfProducts, which implies that as number of products increases the balance decrases.It is possible that the customer has bought various othert products from the bank such as fixed deposits, mutual funds, credit cards etc or invested in other banking financial assets sold by the bank. This could have potentially led to the depletion the account balance. On the contrary, lesses the number of products purchased from the bank, higher is the customer balance
  - There is an insignificant negative correlation of -0.01 between Tenure and Exited, which implies higher the tenure, lesser are the chances of leaving the bank and vice-versa
  - There is an insignificant negative correlation of -0.16 between isActiveMember and Exited, which implies that customers who aren't active i.e., not using bank products regularly, making transactions, etc are more likely to exit the services vis-a-vis customers who are actively transacting with the Bank
  - There is an insignificant negative correlation of -0.03 between CreditScore and Exited, which implies that as the credit score of the customer increases, the probability of the customer leaving the bank decreases. There is also an insignificant positive correlation of 0.01 between Balance, Number of Products and CreditScore. Considering all the factors above, implies that certain customer have a good ROI, which increases their account balance, and hence customers would be investing into other banking products to diversify their portfolio. This would lead to higher credit worthiness
  - There is an insignificant negative correlation of -0.05 between NumOfProducts and Exited, which implies that bigger the banking portfolio lesser are the chances of the customer's leaving the bank and vice-versa. There is also an insignificant positive correlation of 0.01 between isActiveMember, Estimated Salary and NumOfProducts. Considering all the factors above, implies that high salaried individuals tend to transact regularly with the bank to diversify their portfolio by purchasing new banking products
  - There is an insignificant negative correlation of -0.01 between Balance and Exited. This could imply that a select few customer are having a good return on investment and hence choose to stay with the bank as the account balance increases

EDA - Bivariate Analysis – Key Observations & Insights (Cont'd)

- **Key Observations & Insights (Cont'd) :**
  - There is an insignificant negative correlation of -0.01 between HasCard and Exited, which implies that customers with credit cards are less likely to leave the bank and vice versa
  - There is an insignificant positive correlation of 0.09 between isActiveMember and Age, which implies that higher the age, more often does the customer transact with the bank and vice-versa. There is also an insignificant positive correlation of 0.03 between Balance and Age, which implies that customers tend to save more / have a higher account balance as age increases.
  - There is an insignificant negative correlation of -0.01 between HasCard, Tenure and Age, which implies that higher the age, lesser are the chances of the customers having a credit card and lesser is the tenure with the bank i.e. higher probability of exiting the bank
  - There is an insignificant positive correlation of 0.03 between isActiveMember and Creditscore, which implies that higher the banking activity i.e. transactions with the bank, greater is the credit score and vice-versa

Data Preprocessing

Data Preprocessing – Key Observations & Insights

- **Feature Engineering:**
  - RowNumber, CustomerId and Surname columns have been dropped since they do not add value to the dataset
- **Dataset Segregation:**
  - The dependent (Exited) and independent valriables have been split
  - The data has been split into Training, Validation & Testing datasets.
- **Missing Value Treatment:**
  - There are no missing values
- **Encoding for String Variables:**
  - Geography & Gender variables have been encoded within the Training, Validation & Testing
- **Data Normalisation for Numeric Variables:**
  - CreditScore, Age, Tenure, Balance and EstimatedSalary variables have been normalised using the StandardScalar function for the Training, Validation & Testing dataset

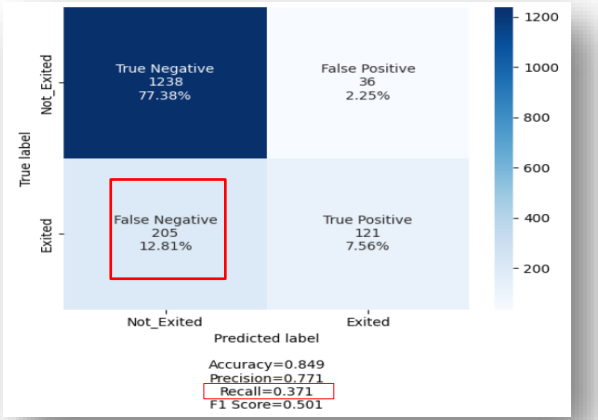
Model Building

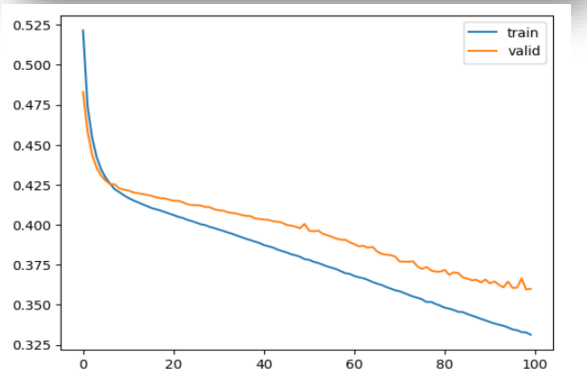
Model Evaluation Criteria & Approach

- **Model Evaluation Criteria :** The primary objective for building the model is to predict whether an existing customer will leave or not in the next 6 months and the key reasons for leaving the Banks' Services. Using the confusion matrix as guiding principle, it is imperative to focus on reducing the False Negatives (FN) i.e., predicting that a customer will not leave, but eventually leaves the Bank. Losing an existing customer would be a significant loss of revenue to the Bank. So, if FN is high, that means the churn will be high. This implies that **reducing False Negatives** should be of utmost importance to the business
    - Key Criteria – Recall: The bank should therefore use Recall as the key model evaluation criteria – higher the Recall, greater are the chances of minimising False Negatives
- **Model Building Approach:** We have split the data into Training, Validation and Testing datasets. We have built an initial model using SGD as an optimizer and evaluated its performance (Recall) on the training and validation dataset. To improve performance, we have built 4 additional models and identified their optimal thresholds using the ROC-AUC curves. Using these thresholds, we have evaluated the models' performance (Recall) on the training and validation datasets. Based on all the **Recall** scores, we have **finalised** the **best performing** model (Model 2 : Neural Network model with Adam as an optimizer). We have then evaluated this model's performance on the Testing dataset
    - Model 1 : Neural Network model with SGD as an optimizer (Initial Model)
    - Model 2 : Neural Network model with Adam as an optimizer
    - Model 3 : Neural Network model with Dropout & Adam optimizer
    - Model 4 : Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer
    - Model 5 : Neural Network model with Balanced Data by applying SMOTE & Adam optimizer

Model Building - Model 1 : SGD as an Optimizer

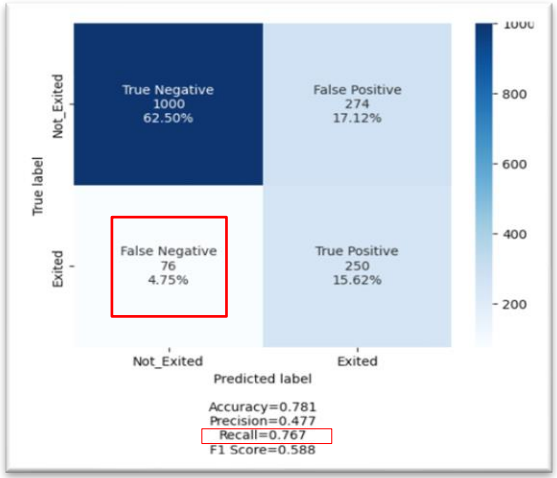
- **Model 1 - Neural Network model with SGD as an optimizer :** Following are the key model attributes, that delivered a significantly low Recall score of 37.1%
  - Optimizer Type: Stochastic Gradient Descent
  - Loss Function: Binary Cross Entropy
  - Total Inputs: 11
  - Total Layers: 3 : Input Layer – 1, Hidden Layer – 1 and Output Layer-1
  - Total Parameters: 2,881
  - Activation Functions: Input Layer – Relu, Hidden Layer – Relu and Output Layer – Sigmoid
  - No. of Neurons: Input Layer – 64, Hidden Layer – 32 and Output Layer – 1
  - Metric: Accuracy
  - Optimal Threshold: 0.5
- Based on the above inputs and using the optimal threshold of 0.5, the model delivered a Recall of 0.371 and a False Negative of 12.81% (205) on the Training & Validation dataset



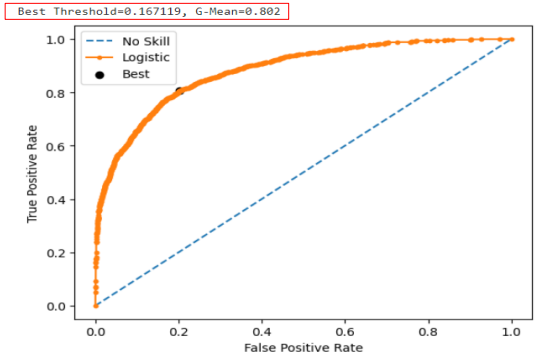


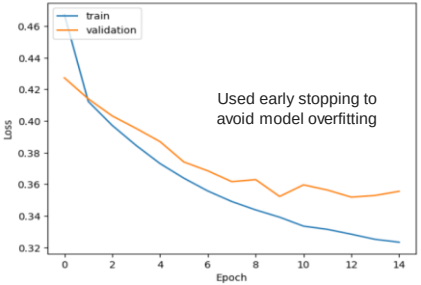
Model Building - Model 2 : Adam as an Optimizer

- **Model 2 - Neural Network model with Adam as an optimizer :** Using the following key attributes, the model delivered a Recall score of 76.6%
  - Optimizer Type: Adam
  - Loss Function: Binary Cross Entropy
  - Total Layers: 3 : Input Layer – 1, Hidden Layer – 1 and Output Layer-1
  - Total Parameters: 2,881
  - Activation Functions: Input Layer – Relu, Hidden Layer – Relu and Output Layer – Sigmoid
  - No. of Neurons: Input Layer – 64, Hidden Layer – 32 and Output Layer – 1
  - Metric: Accuracy
  - ROC AUC Threshold : 0.167119 | G-Mean: 0.802
- Early Stopping – We have used early-stopping / callbacks to prevent model overfitting as seen in the figure on the bottom right side
- Based on the above inputs and using the optimal threshold of 0.167119, the model delivered a Recall of 0.7671 and a False Negative of 4.75% (76) on the Training & Validation dataset



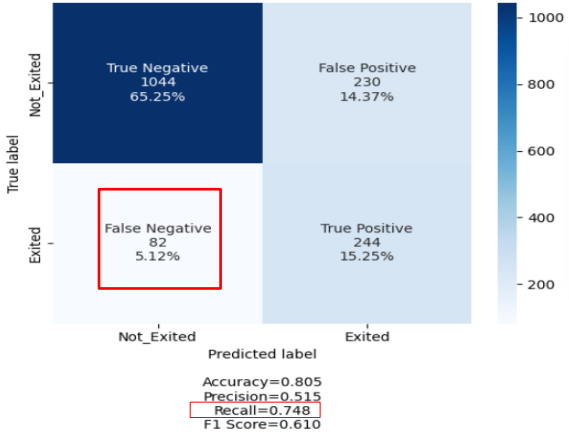
	precision	recall	f1-score	support
0	0.93	0.78	0.85	1274
1	0.48	0.77	0.59	326
accuracy			0.78	1600
macro avg	0.70	0.78	0.72	1600
eighted avg	0.84	0.78	0.80	1600



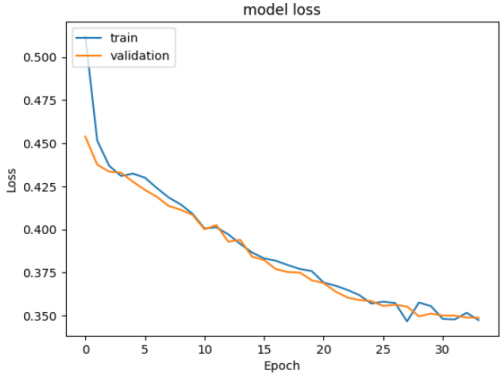


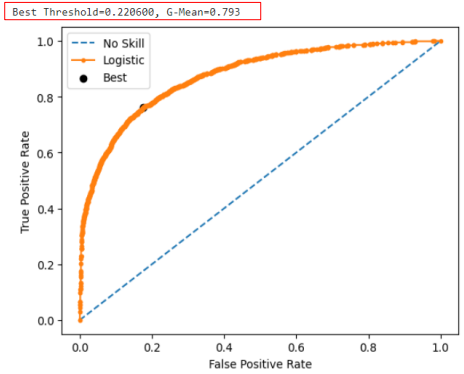
Model Building - Model 3 : Dropout & Adam as an Optimizer

- **Model 3 - Neural Network model with Dropout & Adam as an optimizer :** Using the following key attributes, the model delivered a Recall score of 74.8%
  - Optimizer Type: Adam
  - Loss Function: Binary Cross Entropy
  - Total Layers: 4 : Input Layer – 1, Hidden Layer – 2 and Output Layer-1
  - Total Parameters: 1,057
  - Activation Functions: Input Layer – Relu, Hidden Layer – Relu and Output Layer – Sigmoid
  - No. of Neurons: Input Layer – 32, Hidden Layer1 – 16, Hidden Layer2 – 8 and Output Layer – 1
  - Metric: Accuracy
  - Optimal Threshold: 0.167119
  - Dropout: Dropout Rate at Input Layer is 0.2 and Hidden Layer1 is 0.1
  - ROC AUC Threshold: 0.220600 | G-Mean: 0.793
- Based on the above inputs and using the optimal threshold of 0.220600 , the model delivered a Recall of 0.748 and a False Negative of 5.12% (82) on the Training & Validation dataset



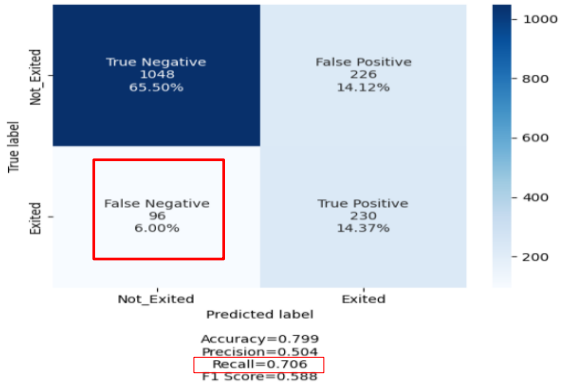
	precision	recall	f1-score	support
0	0.93	0.82	0.87	1274
1	0.51	0.75	0.61	326
accuracy			0.81	1600
macro avg	0.72	0.78	0.74	1600
weighted avg	0.84	0.81	0.82	1600



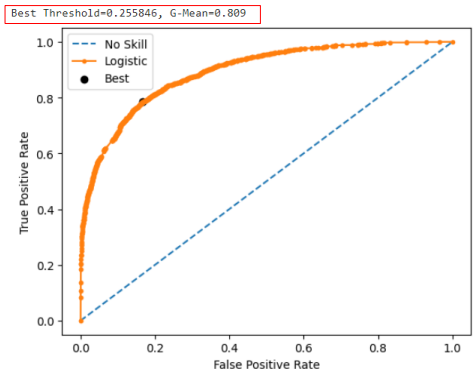


Model Building - Model 4 : Hyperparameter & Adam Optimizer

- **Model 4 - Neural Network model Hyperparameter tuning using Grid search & Adam optimizer :** Using the following key attributes, the model delivered a Recall score of 70.6%
  - Optimizer Type: Adam
  - Loss Function: Binary Cross Entropy
  - Total Layers: 3 : Input Layer – 1, Hidden Layer – 1 and Output Layer-1
  - Total Parameters: 2,881
  - Activation Functions: Input Layer – Relu, Hidden Layer – Relu and Output Layer – Sigmoid
  - No. of Neurons: Input Layer – 64, Hidden Layer – 32, and Output Layer – 1
  - Metric: Accuracy
  - Optimal Threshold: 0.255846
  - Dropout: Dropout Rate at Input Layer is 0.5
  - Optimal Learning Rate = 0.01 | Optimal Batch Size: 64
  - ROC AUC Threshold: 0.255846 | G-Mean: 0.793
- Based on the above inputs and using the optimal threshold of 0.255846 , the model delivered a Recall of 0.706 and a False Negative of 6% (96) on the Training & Validation dataset

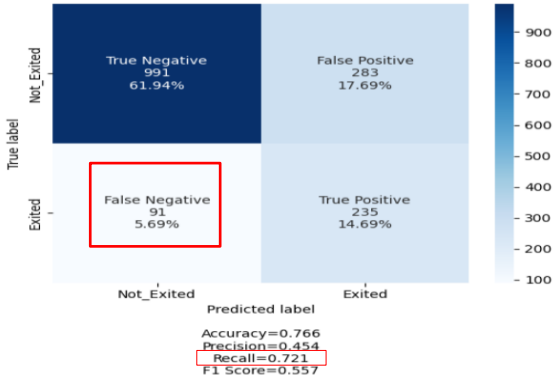


	precision	recall	f1-score	support
0	0.92	0.82	0.87	1274
1	0.50	0.71	0.59	326
accuracy			0.80	1600
macro avg	0.71	0.76	0.73	1600
weighted avg	0.83	0.80	0.81	1600

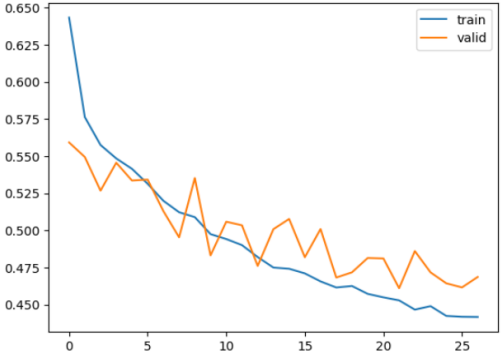


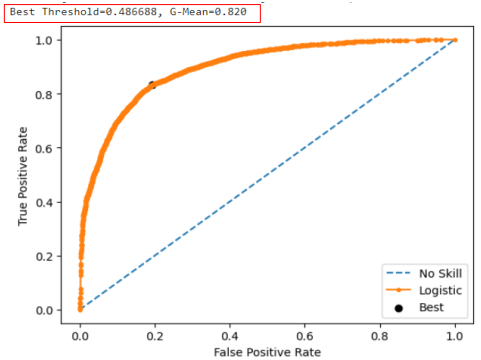
Model Building - Model 5 : Balanced Data & Adam Optimizer

- **Model 5 - Neural Network model with Balanced Data by applying SMOTE & Adam optimizer :** Using the following key attributes, the model delivered a Recall score of 72.1%
  - Optimizer Type: Adam
  - Loss Function: Binary Cross Entropy
  - Total Layers: 4 : Input Layer – 1, Hidden Layer – 2 and Output Layer-1
  - Total Parameters: 1,057
  - Activation Functions: Input Layer – Relu, Hidden Layer – Relu and Output Layer – Sigmoid
  - No. of Neurons: Input Layer – 32, Hidden Layer1 – 16, Hidden Layer2 – 8 and Output Layer – 1
  - Metric: Accuracy
  - Optimal Threshold: 0.486688
  - Dropout: Dropout Rate at Input Layer is 0.2 and Hidden Layer1 is 0.1
  - ROC AUC Threshold: 0.486688 | G-Mean: 0.820
- Based on the above inputs and using the optimal threshold of 0.486688 , the model delivered a Recall of 0.721 and a False Negative of 5.69% (91) on the Training & Validation dataset



	precision	recall	f1-score	support
0	0.92	0.78	0.84	1274
1	0.45	0.72	0.56	326
accuracy			0.77	1600
macro avg	0.68	0.75	0.70	1600
weighted avg	0.82	0.77	0.78	1600





Model Feature Comparison Summary - All Models

- Following is the feature comparison summary of all models that were used on the training and validation dataset

#	Model Type			Activation Functions Used			No of Neurons per Layer			Optimizer Attributes			
		Total Layers	Total Params	Input Layer	Hidden Layer	Output Layer	Input Layer	Hidden Layer	Output Layer	Type	Loss	Metric	Optimal Threshold
1	Model 1 : Neural Network model with SGD as an optimizer	3 1 x Input 1 x Hidden 1 x Output	2881	Relu	Relu	Sigmoid	64	32	1	SGD	Binary Cross Entropy	Accuracy	0.5
2	Model 2 : Neural Network model with Adam as an optimizer	3 1 x Input 1 x Hidden 1 x Output	2881	Relu	Relu	Sigmoid	64	32	1	Adam	Binary Cross Entropy	Accuracy	0.167119
3	Model 3 : Neural Network model with Dropout & Adam optimizer	4 1 x Input 2 x Hidden 1 x Output	1057	Relu	Relu	Sigmoid	32	16 & 8	1	Adam	Binary Cross Entropy	Accuracy	0.220600
4	Model 4 : Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer	3 1 x Input 1 x Hidden 1 x Output	2881	Relu	Relu	Sigmoid	64	32	1	Adam	Binary Cross Entropy	Accuracy	0.255846
5	Model 5 : Neural Network model with Balanced Data by applying SMOTE & Adam optimizer	4 1 x Input 2 x Hidden 1 x Output	1057	Relu	Relu	Sigmoid	32	16 & 8	1	Adam	Binary Cross Entropy	Accuracy	0.486688

Model Performance Comparison Summary - All Models

- Following is the performance metric summary of all models on the Training & Validation datasets. Model 2 : Neural Network model with Adam as an optimizer has delivered a Recall score of 76.7% and False Negative of 4.75% (76), whereas Model 3 : Neural Network model with Dropout & Adam optimizer has delivered a Recall score of 74.8% and False Negative of 5.12% (82) on the Training & Validation dataset. Based on our key criteria Recall score, **Model 2 : Neural Network model with Adam as an optimizer has the best performance followed by Model 4: Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer**

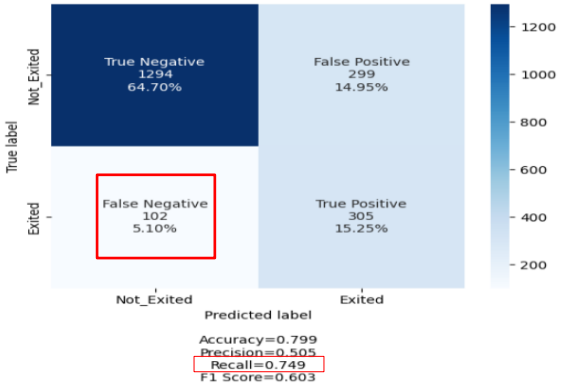
#	Model Type	Key Performance Metrics				Confusion Matrix Scores			
		Accuracy	Recall	Precision	F1	True Positive	True Negative	False Positive	False Negative
1	Model 1 : Neural Network model with SGD as an optimizer	0.849	0.371	0.771	0.501	7.56% 121	77.38% 1238	2.25% 36	12.81% 205
2	Model 2 : Neural Network model with Adam as an optimizer (1)	0.781	0.767	0.477	0.588	15.62% 250	62.50% 1000	17.12% 274	4.75% 76
3	Model 3 : Neural Network model with Dropout & Adam optimizer (2)	0.805	0.748	0.515	0.610	15.25% 244	65.25% 1044	14.37% 230	5.12% 82
4	Model 4 : Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer	0.799	0.706	0.504	0.588	14.37% 230	65.50% 1048	14.12% 226	6.00% 96
5	Model 5 : Neural Network model with Balanced Data by applying SMOTE & Adam optimizer	0.766	0.721	0.454	0.557	14.69% 235	61.94% 991	17.69% 283	5.69% 91

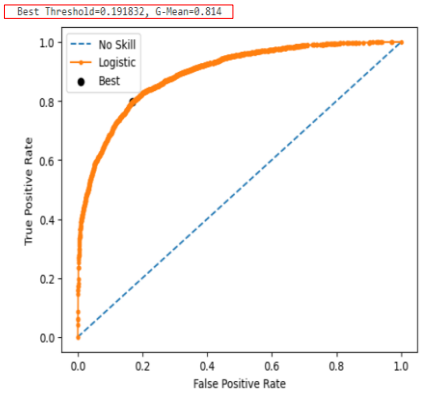
Model Building – Best & Final Model : Adam as an Optimizer

- **Best Model:** Based on the performance comparison of all the 5 models, Model 2 : Neural Network model with Adam as an optimizer has delivered a Recall score of 76.7% and False Negative of 4.75% (76) on the Training & Validation dataset. We have therefore considered this as our best candidate model and used it on the Testing dataset

- **KPI Comparison:** The model has delivered a Recall score of 0.767 with a False Negative of 4.75% (76) on the Training & Validation dataset, and a Recall score of 0.749 with a False Negative of 5.10% (102) on the Testing dataset

- **Final Model - Model 2 : Neural Network model with Adam as an optimizer :** With a Recall score of 74.9% and a False Negative of 5.10% (102) on the Testing dataset using an optimal threshold of 0.191832 and a G-Mean of 0.814, the Model 2 : Neural Network model with Adam as an optimizer has generalised its performance and can be considered as the final model





	Key Performance Metrics				Confusion Matrix Scores			
	Accuracy	Recall	Precision	F1	True Positive	True Negative	False Positive	False Negative
Training & Validation Dataset	0.781	0.767	0.477	0.588	15.62% 250	62.50% 1000	17.12% 274	4.75% 76
Testing Dataset	0.799	0.749	0.505	0.603	15.25% 305	64.70% 1294	14.95% 299	5.10% 102

Conclusion

- We have built an initial model using SGD as an optimizer and evaluated its performance (Recall) on the training and validation dataset. To improve the model performance, we have built 4 additional models (below), identified their optimal thresholds using the ROC-AUC curves. Using these thresholds, we have evaluated the models' performance (Recall) on the training and validation datasets.
  - Model 1 : Neural Network model with SGD as an optimizer (Initial Model)
  - Model 2 : Neural Network model with Adam as an optimizer
  - Model 3 : Neural Network model with Dropout & Adam optimizer
  - Model 4 : Neural Network model with Hyperparameter tuning using Grid search & Adam optimizer
  - Model 5 : Neural Network model with Balanced Data by applying SMOTE & Adam optimizer
- Based on the performance comparison (Recall & False Negative) of all the models, we have **finalised** the **best performing** model (Model 2 : Neural Network model with Adam as an optimizer). We have then evaluated this model's performance on the Testing dataset. This model has delivered a Recall score of 0.767 & 0.749, and a False Negative of 4.75% (76) and 5.10% (102) on the Training & Validation, and the Testing dataset respectively. The model's Recall score on the Testing dataset is inline to the one observed on the Training & Validation datasets. This will minimise False Negatives, which is of utmost importance to the business
- Using an optimal threshold of 0.191832 and a G-Mean of 0.814, the Model 2 : Neural Network model with Adam as an optimizer has delivered a Recall score of 74.9% and a False Negative of 5.10% (102) on the Testing dataset. This model has generalised its performance and can be considered as the final model.
- The model built can be used to predict customer churn i.e., whether a customer will leave or not. This will help the bank to target the potential customers, who have a higher probability of leaving, and proactively incentivise them in order to maximise customer retention
- The bank should focus on improving key services that are related to isActiveMember, NumOfProducts, HasCreditCard features. Improving these services will lead to customer retention (Tenure), thereby driving growth (Balance) and potential customer acquisition in new areas (Geography)

Key Actionable Business Insights

- **Summarised Key Observations & Insights :**
  - There is a significant imbalance in the dataset since there are high number of existing customers (79.6%) than compared to 20.4% of attrited customers
  - The attrition levels are higher in customers who are inactive i.e., not transacting on a regular basis, than compared to customers who are actively transacting with the bank. The bank's customer service should proactively reach out to inactive members to identify and resolve the key pain-points. This might lead to a positive conversion (inactive to active) and would help customer retention
  - There is a positive correlation between Balance and Exited, which implies that as the account balance increases the customer has a high probability of the leaving the bank. This could imply that the bank probably doesn't offer a great return on investment (ROI), which does not incentivise the customer to continue the relationship. The bank's finance strategy should focus on Investments vs ROI in order to retain customers
  - There is a positive correlation between Estimated Salary and Exited, which implies that as the customer salary increases, higher is the probability of the customer leaving the bank. This could imply that bank probably doesn't offer great services to high net-worth individuals (HNI), which again does not incentivise customers to continue the relationship. The bank should provide the customers with various incentives and reward schemes, especially to high net-worth individuals, in order to enhance relationship and retain customers. The banks' customer strategy should focus on building customer relationship and providing high-end banking services in order to retain HNI customers
  - There are higher number of customers leaving the bank in Germany than compared to Spain and France. Also, there are higher number of female customers leaving the bank than compared to male customers. The bank's customer strategy should focus on a diversity and initiate a customer feedback program to identify the key challenges and take appropriate measure to resolve these challenges
  - In order to retain existing customers and acquire potential prospects, the bank should incentivise customers with cashback schemes and loyalty reward points, that can be redeemed on future purchases on using their banking products e.g. using credit cards to earn cashbacks, which might encourage customers on using their credit cards more often

Our Recommendation

Based on the key observations and insights, we recommend the following areas of improvement / opportunities that will drive business growth and lead to a better customer experience

- **Implement Customer Incentivisation Scheme:** Incentivising customers by offering them cashback schemes and discounts / vouchers on purchases will encourage frequent spending and will drive customer gowth and increase revenue
- **Implement Customer Satisfaction Survey:** The bank should initiate a targeted Customer Satisfaction Survey to understand customer pain points and implement the findings to improve retention ratio of such customers
- **Implement Tier based Rewards:** The bank should introduce a Tier based Loyalty & Rewards Scheme for purchases using their banking products e.g. Credit / Debit Cards. Cumulative loyalty points above a certain threshold will promote the customer to a new tier, that will offer specific rewards such as First-Class Lounge access at Airports, Spa & Well-Being discounts etc

APPENDIX

Appendix - Notes

- Further analysis would be required on a comprehensive dataset to provide customer segmentation strategies

Great Learning

Happy Learning !