Bank Churn Prediction Model

Project #4- Introduction to Neural Networks

Wednesday, October 16, 2024

Presented by: Sarah Lasater

Contents / Agenda

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

Executive Summary

Executive Summary

The key objective of this project is to predict customer churn accurately using neural network models, helping a bank to take proactive steps in retaining at-risk customers. Six models with various functions and parameters were evaluated based on their recall values, which indicate how well the models can identify customers likely to churn.

Key Findings:

- **Training Performance**: The Neural Network (NN) model using **SMOTE & Adam** optimization performed the best during training, achieving a recall of **0.953946**. This high recall suggests that the model is highly effective in identifying churn during the training phase, capturing a significant portion of potential churners.
- **Validation Performance**: When evaluated on the validation set, the model that included **SMOTE, Adam & Dropout** showed the best performance with a recall of **0.719902**. This result indicates that this model is better suited for generalizing to new, unseen data due to the use of dropout, which helps mitigate overfitting.

Conclusion:

Although the **NN with SMOTE & Adam** performed better on the training data, the **NN with SMOTE, Adam & Dropout** model is the most appropriate for real-world application, offering a balance between identifying churn and maintaining robust performance on new data. For practical deployment, the latter model is recommended as it provides better generalization and reduces the risk of overfitting.

This insight can guide the bank in choosing the best model for predicting customer churn, allowing targeted retention efforts to be focused on customers most likely to leave.

Business Problem Overview and Solution Approach

Business Problem Overview

Customer churn is a critical issue faced by businesses that offer services, particularly in competitive industries like banking. Customer churn refers to the phenomenon where customers stop using a company's services and switch to a competitor. For banks, losing customers can result in significant revenue loss and impact long-term profitability.

To retain customers, it is crucial to understand the factors that influence a customer's decision to leave the bank. By identifying the key drivers of churn, management can focus on targeted strategies to enhance service quality and customer satisfaction, ultimately reducing churn rates.

In this project, the bank would like a solution developed to predict customer churn. The goal is to build a neural network-based classifier that can accurately determine whether a customer is likely to leave the bank within the next six months. This predictive model will enable the bank to proactively address potential churn risks, allowing the management to implement timely retention strategies and improve overall customer loyalty.

Solution Approach

In this scenario, a false negative prediction (failing to identify a churning customer) is generally more detrimental than a true positive. Therefore, maximizing Recall will help the bank minimize false negatives, allowing for more effective retention strategies. This is the metric that will be used to evaluate and select the best prediction model for this problem.

This approach will help the bank prioritize service improvements based on data-driven insights from their customer data set, ensuring that efforts are concentrated on the aspects of service that matter most to customers at risk of leaving.

EDA Results

EDA Results- Univariate Observations & Conclusions

- **Customer Credit Score**: The majority of customers have a credit score centered around 650, with 50% falling between 584 and 718. The distribution includes several outliers with significantly lower credit scores, which may indicate financial difficulties for these customers and potential risk for churn. *(SEE U1)*
- **Customer Age**: The average customer age is 39, with 50% of customers aged between 32 and 44. Notably, there are many outliers among older customers, which may suggest that older customers either have specific needs or face different challenges that could influence their likelihood to churn.*(SEE U2)*
- **Account Balance**: A significant number of customers (3,500) have an account balance of $0, indicating a large segment of potentially inactive or low-value customers. For the remaining customers, account balances vary widely, with the upper quartile holding balances between $97,190 and $250,000, highlighting a small but significant group of high-value clients. *(SEE U3)*
- **Estimated Salary**: Salaries are evenly distributed, with an average of $100,000. This equal distribution across income ranges suggests that no particular salary bracket dominates the customer base. However, a subset of customers earns significantly less than average, which could impact their product engagement and likelihood to churn. *(SEE U4)*
- **Customer Exits**: Approximately 20.4% of customers have exited the bank. This figure highlights a notable level of churn that warrants targeted retention strategies, especially in identifying the factors contributing to the 20% churn rate. *(SEE U5)*
- **Geography**: Germany hosts the highest number of customers, followed closely by Spain and France. The geographical distribution implies that regional factors or localized customer service may impact churn, particularly in Germany where customer numbers are largest. *(SEE U6)*

EDA Results- Univariate Observations & Conclusions

- **Gender**: The customer base has more male customers (54.5%) than female customers (45.5%), though the difference is not drastic. Gender may play a role in service preferences or churn behavior, and understanding these dynamics could help tailor retention efforts. *(SEE U7)*
- **Account Tenure**: Tenure is well distributed across the customer base, with most customers having been with the bank for 1-9 years. However, there is a sharp drop-off for customers with 0 or 10 years of tenure, indicating that both very new and very long-term customers might require additional focus to prevent churn.*(SEE U8)*
- **Number of Products**: Most customers hold either one or two products, while only a small fraction hold three or four products. This suggests a significant opportunity to cross-sell additional products to the majority of customers who currently hold fewer products. *(SEE U9)*
- **Has Credit Card**: A large majority of customers (71%) have a credit card, indicating this is a common product across the base. The relatively smaller group without a credit card may represent a lower engagement segment of the customer base. *(SEE U10)*
- **Is Active Member**: The split between active (51%) and inactive (49%) members is nearly even, highlighting the need for strategies to re-engage inactive customers to prevent potential churn. This nearly balanced distribution suggests that customer engagement could be a key factor in churn prediction. *(SEE U11)*

EDA Results- Bivariate Observations & Conclusions

1. **Correlation:** There do not seem to be any categorical features with any degree of correlation above .03, which are Age and Balance. *(SEE B1)*
2. **Exited vs Geography**: Customers in Germany have a significantly higher churn rate (32.5%) compared to customers in France (16.1%) and Spain (16.7%). This indicates that geographical location plays an important role in churn, with Germany being a region that requires focused retention strategies. *(SEE B2)*
3. **Exited vs Gender**: Female customers have a higher churn rate (25%) compared to male customers (16.4%). This suggests that female customers may face unique challenges or dissatisfaction, and targeted efforts may be needed to better address their needs to reduce churn. *(SEE B3)*
4. **Exited vs Has Credit Card**: Interestingly, customers without a credit card have a much higher churn rate (29.3%) compared to those with a credit card (8.8%). This could indicate that having a credit card is a sign of higher engagement with the bank's services, and encouraging more customers to use this product might help in reducing churn. *(SEE B4)*
5. **Exited vs Is Active Member**: Inactive members show a substantially higher churn rate (44.4%) compared to active members (15%). This finding indicates that customer engagement is a critical factor in preventing churn, and strategies to increase engagement could have a significant impact on retention. *(SEE B5)*
6. **Exited vs Credit Score**: While the overall credit score ranges are similar for both exited and non-exited customers, those who churn tend to have more customers with credit scores below 585 (Q1). This suggests that customers with lower credit scores may be at a higher risk of exiting, which could be a signal for early intervention. *(SEE B6)*
7. **Exited vs Age**: Customers who have exited tend to be older, with 50% of churners aged between 39-50 years old compared to non-exited customers, who are younger (31-41 years old). This indicates that older customers may face issues that contribute to churn, and retention efforts should consider their specific needs or concerns. *(SEE B7)*

EDA Results- Bivariate Observations & Conclusions

7. **Exited vs Tenure**: Both exited and non-exited customers show similar tenure distributions, but customers with shorter tenures (less than 2 years) are more likely to churn. This suggests that early intervention in the customer journey is crucial to improving retention rates. *(SEE B8)*
8. **Exited vs Account Balance**: Customers with lower account balances are more likely to churn, with exited customers having balances between $47,000-$126,000 compared to $0-$125,000 for non-exited customers. This indicates that low account balance customers may need more attention, especially those with balances below $47,000. *(SEE B9)*
9. **Exited vs Number of Products**: Both exited and non-exited customers typically hold 1-2 products, but churners are more likely to hold fewer products. Encouraging customers to engage with more products could increase retention, especially since those with more than two products tend to churn less. *(SEE B10)*
10. **Exited vs Estimated Salary**: There is no significant difference between the salary ranges of exited and non-exited customers. Both groups have a fairly even salary distribution, which suggests that salary may not be a strong predictor of churn on its own, and other factors like engagement or product usage are more critical. *(SEE B11)*

These insights indicate that **customer engagement** (being an active member, holding a credit card) and **region-specific challenges** (customers in Germany) are key drivers of churn. Focused retention efforts should target inactive members, customers with low engagement, and those in high-risk regions. Additionally, early intervention for new customers and providing personalized services for tenured customers could further reduce churn rates.

Data Preprocessing

Data Preprocessing

Missing value treatment

- There were no missing values in the training data set therefore no need for treatment.

Train-Validation-Test Split

- The X and y were split at .20 to obtain the test set
  - (X_train.shape, X_val.shape, X_test.shape)
  - (6000, 11) (2000, 11) (2000, 11)
- The X_large and y_large were split at .25 to obtain the train and validation set
  - (y_train.shape, y_val.shape, y_test.shape)
  - (6000,) (2000,) (2000,)

Data preparation for modeling

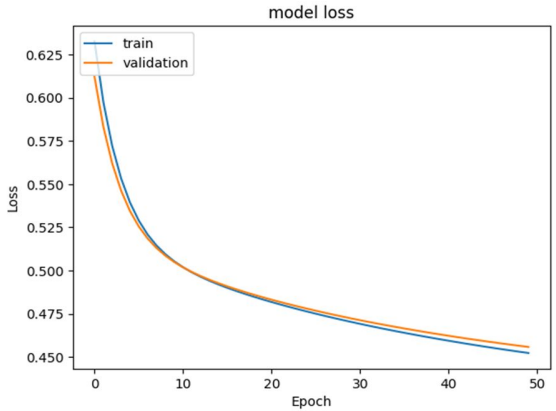
- Columns for “RowNumber”, “CustomerID”, and “Surname” are unique customer features thus were dropped
- The column “Exited” was dropped
- Dummy variables were created for all categorical columns: “Geography”, “Gender”, “Has Credit Card”, and “Is Active Member”

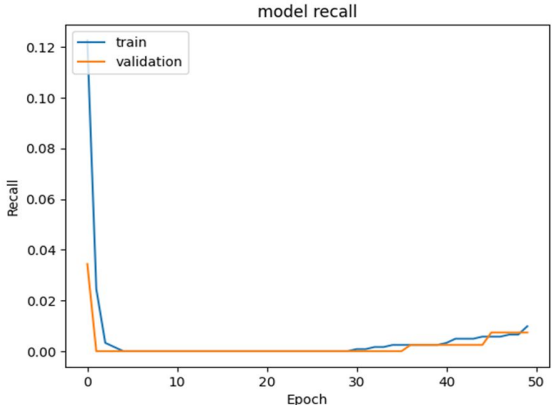
Note: You can use more than one slide if needed

Model Performance Summary

Model Performance Summary- Overview & Parameters

Model #0- Neural Network with SGD Optimizer

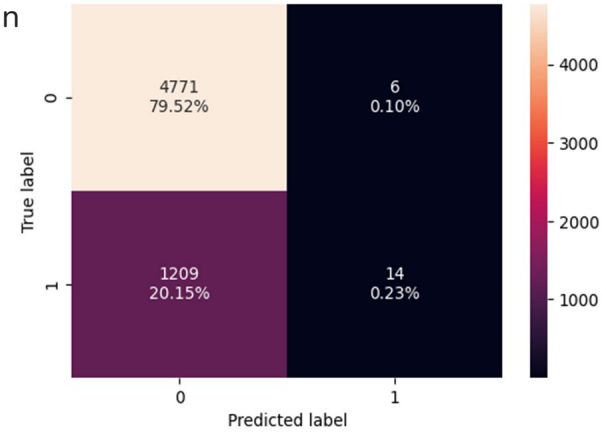




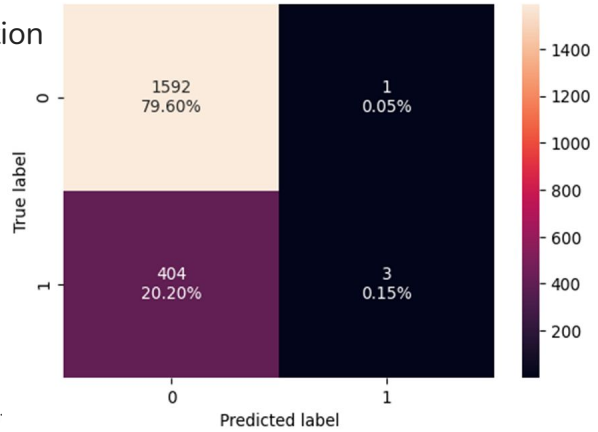
Input neurons: 64
Hidden neurons: 32
Activation: ReLU
Output neurons: 1
Batch size: 64
Epochs: 50

The model performs well in predicting non-churners, but it struggles significantly to identify actual churners. This is a critical issue because the model's primary goal is to accurately predict customer churn. Despite a good overall accuracy, the model fails to capture the majority of churners, making it unreliable for churn prediction.

Train

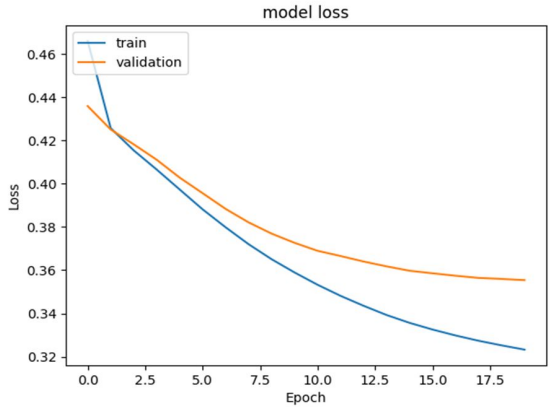


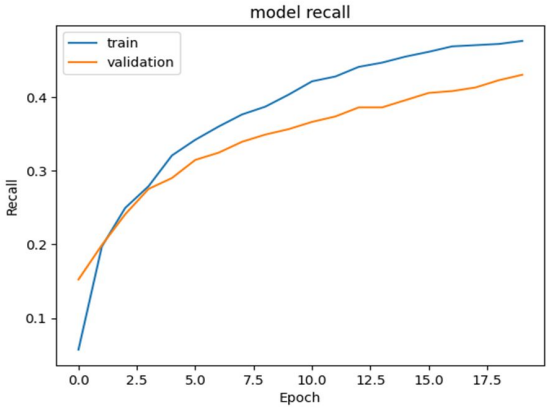
Validation



Model Performance Summary- Overview & Parameters

Model #1- Neural Network with Adam Optimizer

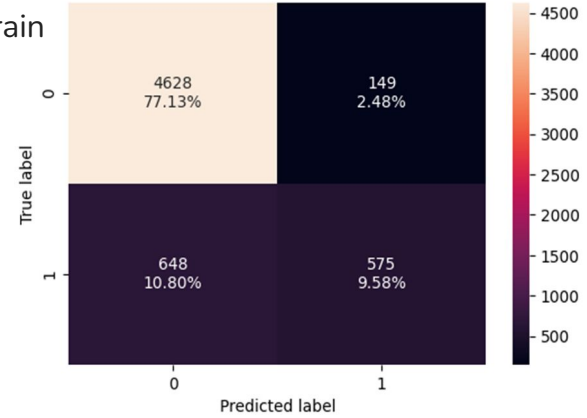




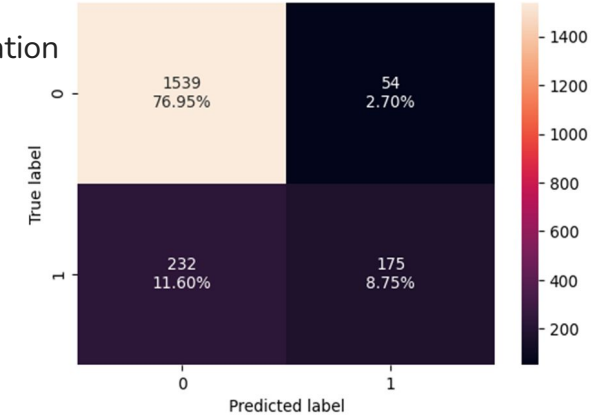
Input neurons: 32
Hidden neurons: 64
Activation: ReLU
Output neurons: 1
Batch size: 32
Epochs: 20

The model performs well at predicting non-churners however, it has difficulty identifying churners. While the overall accuracy is 86%, the model's ability to detect churners could be improved. Improving recall for churners would be essential for more balanced performance, especially in a churn prediction context.

Train

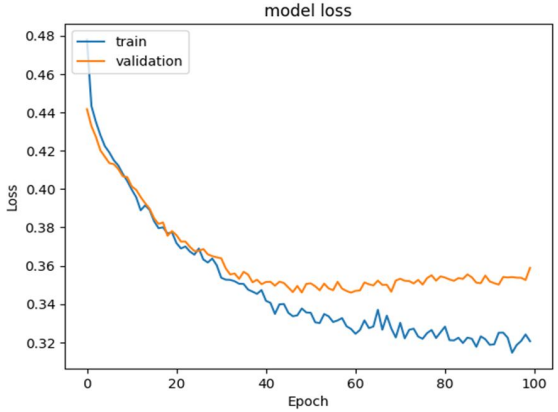


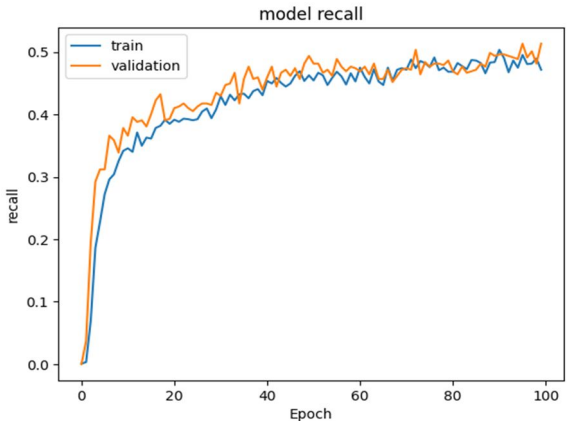
Validation



Model Performance Summary- Overview & Parameters

Model #2- Neural Network with Adam Optimizer & Dropouts

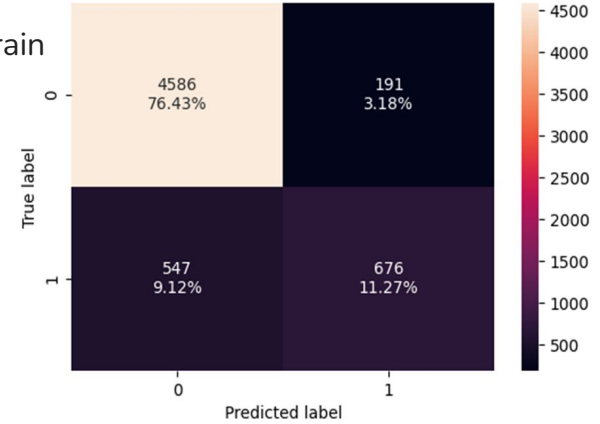




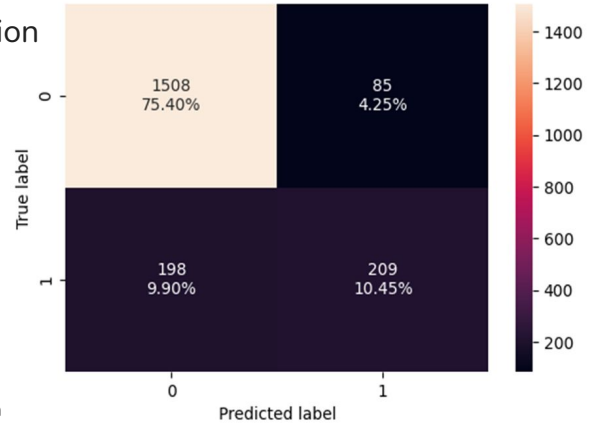
Input neurons: 32
Hidden neurons: 16
Activation: ReLU
Output neurons: 1
Dropouts: .2 / .1
Batch size: 32
Epochs: 100

The model performs well in predicting non-churners with high Recall, however, its performance in predicting churners is moderate, with a recall of 51%, meaning it misses nearly half of the churners. Improving recall for churners would enhance the model's effectiveness in identifying customers at risk of churning.

Train

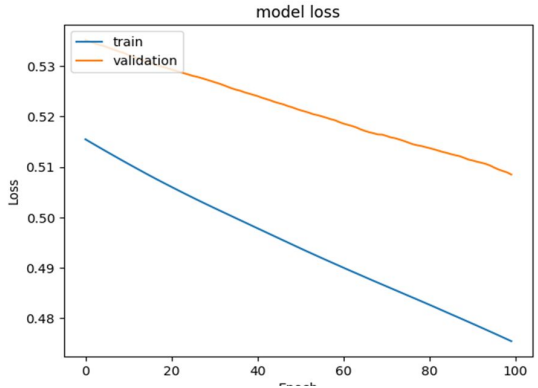


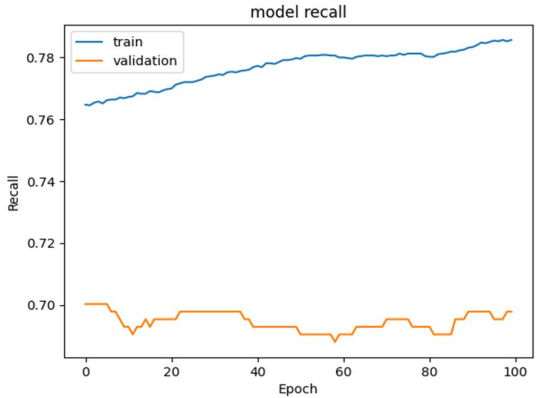
Validation



Model Performance Summary- Overview & Parameters

Model #3- Neural Network with Balanced Data (SMOTE) and SGD Optimizer

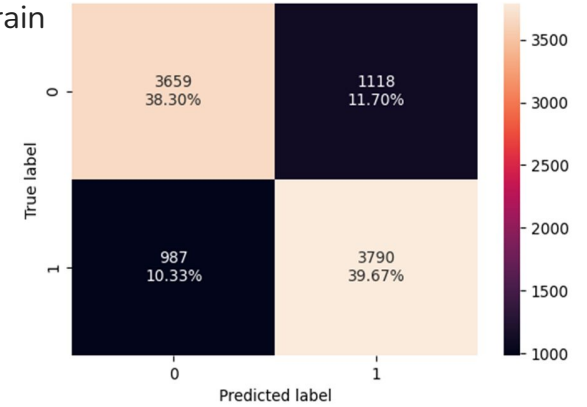




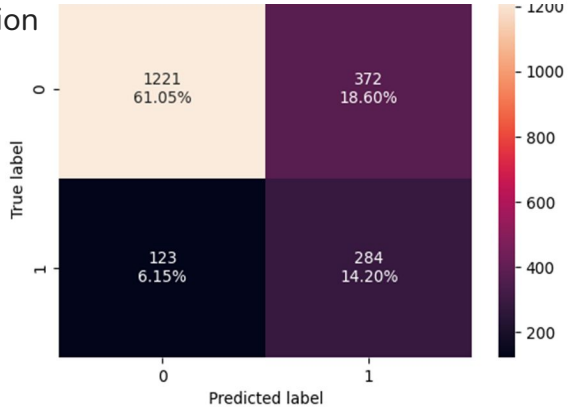
Input neurons: 32
Hidden neurons: 32
Activation: ReLU
Output neurons: 1
Batch size: 32
Epochs: 100

The model performs well in identifying non-churners, with strong recall, but struggles with churners.. While it successfully identifies 70% of actual churners, it frequently misclassifies non-churners as churners.

Train

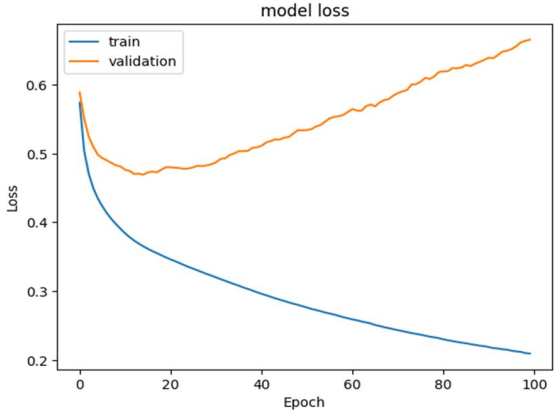


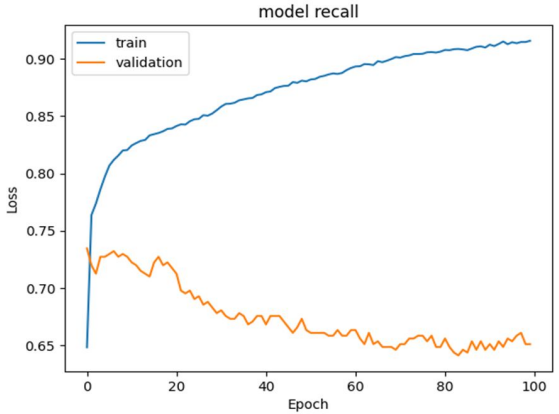
Validation



Model Performance Summary- Overview & Parameters

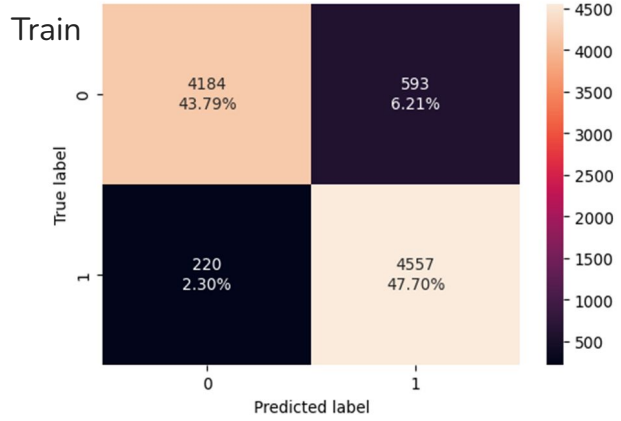
Model #4- Neural Network with Balanced Data (SMOTE) and Adam Optimizer

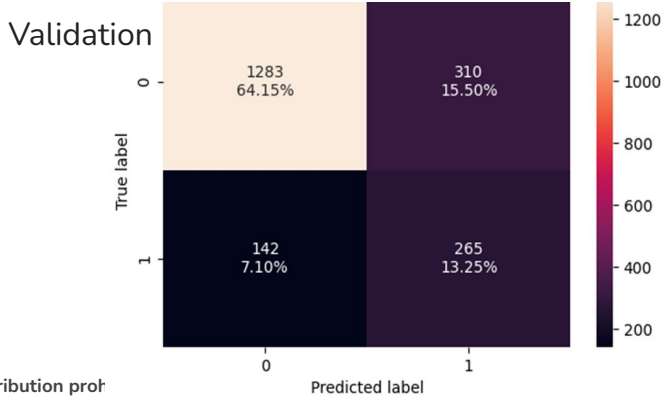




Input neurons: 32
Hidden neurons: 32
Activation: ReLU
Output neurons: 1
Batch size: 32
Epochs: 100

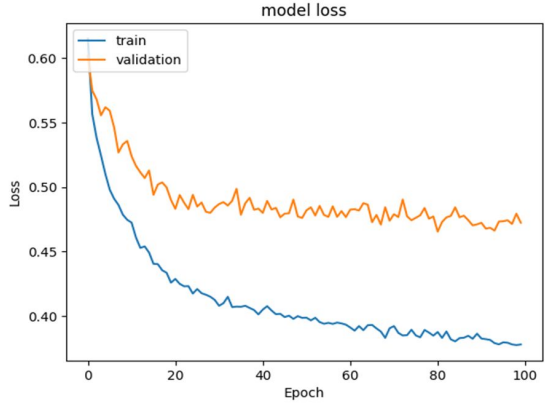
The model performs well in predicting non-churners with a good recall of 81%. However, it struggles with churners meaning there are many false positives. The recall for churners is better at 65%, but it still has difficulty in accurately predicting this class. Overall, there is room for improvement in identifying churners to reduce false positives and enhance its reliability for churn prediction.

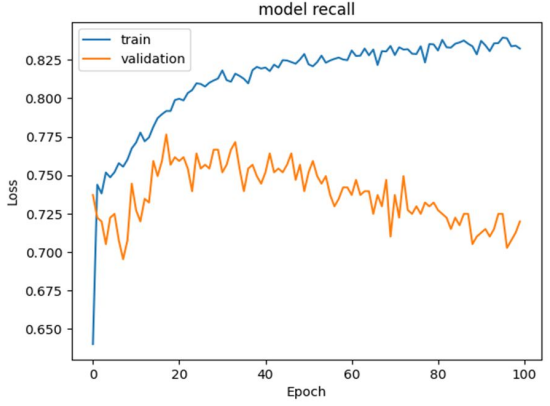




Model Performance Summary

Model #5- Neural Network with Balanced Data (SMOTE), Adam Optimizer, and Dropouts

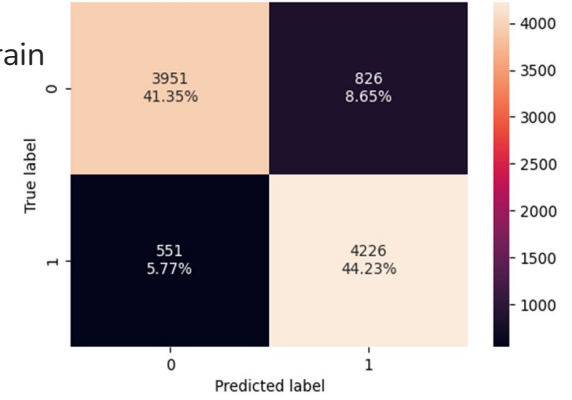




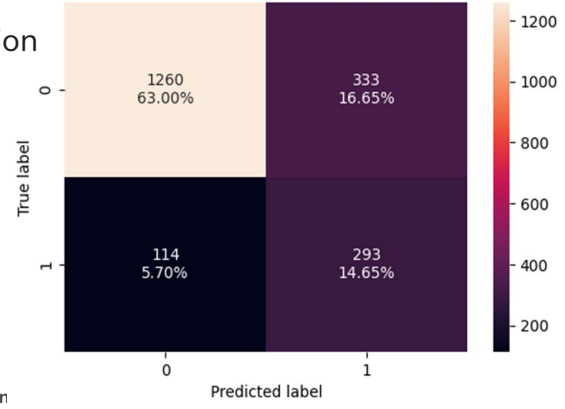
Input neurons: 32
Hidden neurons: 32
Activation: ReLU
Dropout: .20
Output neurons: 1
Batch size: 32
Epochs: 100

The model performs well in predicting non-churners with strong recall (79%). It struggles more with churners, where the precision is low (47%), indicating many false positives. Despite this, the recall for churners is relatively high (72%), meaning the model captures most churners. The overall accuracy of 78% is reasonable, but improving precision for churners would make the model more effective in predicting churn and reducing false positives.

Train



Validation



Model Summaries

Model 0

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	768
dense_1 (Dense)	(None, 32)	2,080
dense_2 (Dense)	(None, 1)	33

Total params: 2,881 (11.25 KB)
Trainable params: 2,881 (11.25 KB)
Non-trainable params: 0 (0.00 B)

Model 1

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 32)	384
dense_1 (Dense)	(None, 64)	2,112
dense_2 (Dense)	(None, 1)	65

Total params: 2,561 (10.00 KB)
Trainable params: 2,561 (10.00 KB)
Non-trainable params: 0 (0.00 B)

Model 2

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 32)	384
dropout (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 16)	528
dense_2 (Dense)	(None, 16)	272
dropout_1 (Dropout)	(None, 16)	0
dense_3 (Dense)	(None, 16)	272
dense_4 (Dense)	(None, 1)	17

Total params: 1,473 (5.75 KB)
Trainable params: 1,473 (5.75 KB)
Non-trainable params: 0 (0.00 B)

Model 3

Model: "sequential_1"

Layer (type)	Output Shape	Param #
dense_4 (Dense)	(None, 32)	384
dense_5 (Dense)	(None, 32)	1,056
dense_6 (Dense)	(None, 32)	1,056
dense_7 (Dense)	(None, 1)	33

Total params: 2,531 (9.89 KB)
Trainable params: 2,529 (9.88 KB)
Non-trainable params: 0 (0.00 B)
Optimizer params: 2 (12.00 B)

Model Summaries

Model 4

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 32)	384
dense_1 (Dense)	(None, 32)	1,056
dense_2 (Dense)	(None, 32)	1,056
dense_3 (Dense)	(None, 1)	33

Total params: 2,529 (9.88 KB)
Trainable params: 2,529 (9.88 KB)
Non-trainable params: 0 (0.00 B)

Model 5

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 32)	384
dropout (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 32)	1,056
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 8)	264
dense_3 (Dense)	(None, 1)	9

Total params: 1,713 (6.69 KB)
Trainable params: 1,713 (6.69 KB)
Non-trainable params: 0 (0.00 B)

Final Model Performance Summary

Summary of key performance metrics for training and test data

Training performance comparison

	recall
NN with SGD	0.011447
NN with Adam	0.470155
NN with Adam & Dropout	0.552739
NN with SMOTE & SGD	0.793385
NN with SMOTE & Adam	0.953946
NN with SMOTE,Adam & Dropout	0.884656

NN with SMOTE & Adam had highest Recall performance on training data set

Validation set performance comparison

	recall
NN with SGD	0.007371
NN with Adam	0.429975
NN with Adam & Dropout	0.513514
NN with SMOTE & SGD	0.697789
NN with SMOTE & Adam	0.651106
NN with SMOTE,Adam & Dropout	0.719902

NN with SMOTE, Adam, and Dropout had highest Recall performance on validation data set

```
train_metric_df - valid_metric_df
```

	recall
NN with SGD	0.004076
NN with Adam	0.040180
NN with Adam & Dropout	0.039226
NN with SMOTE & SGD	0.095596
NN with SMOTE & Adam	0.302840
NN with SMOTE,Adam & Dropout	0.164754

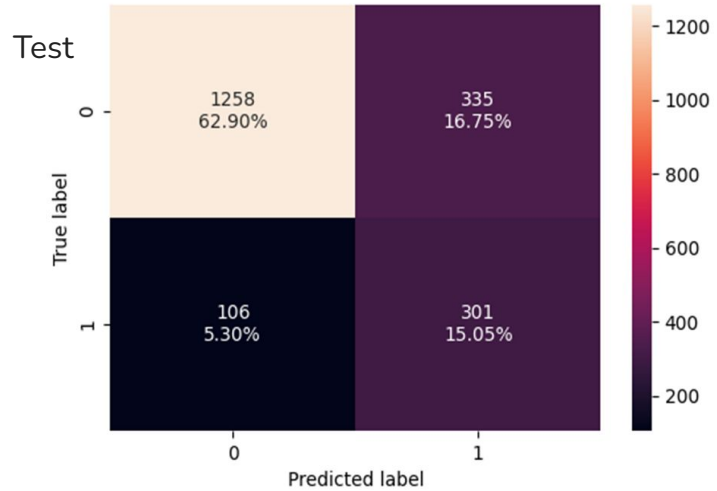
Although **NN with SMOTE, Adam, and Dropout** did NOT have smallest difference between training and validation Recall scores, it still performed well

Final Model Performance Summary

The model using SMOTE & Adam optimization achieved the highest recall during training at 0.953946, demonstrating its effectiveness in identifying potential churners. However, the model with SMOTE, Adam & Dropout performed better on the validation set with a recall of 0.719902, indicating superior generalization to new data and reduced risk of overfitting.

While the first model excels in training, the latter is more suitable for real-world deployment, providing a balance between churn prediction accuracy and robustness on unseen data. Therefore, the model with SMOTE, Adam & Dropout is recommended for practical use to help the bank focus retention efforts on customers most likely to churn.

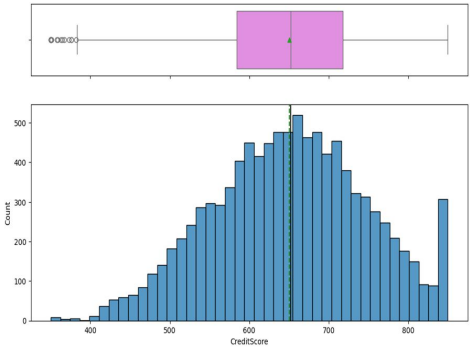
	precision	recall	f1-score	support
0	0.92	0.79	0.85	1593
1	0.47	0.74	0.58	407
accuracy			0.78	2000
macro avg	0.70	0.76	0.71	2000
weighted avg	0.83	0.78	0.80	2000



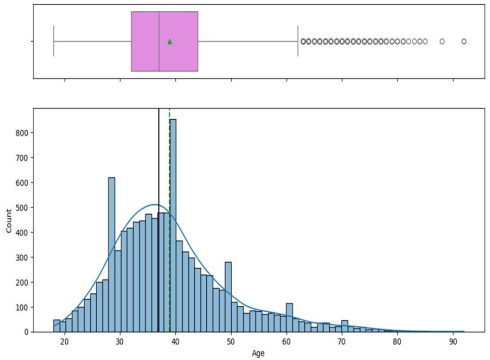
APPENDIX

Exploratory Data Analysis- Univariate Data

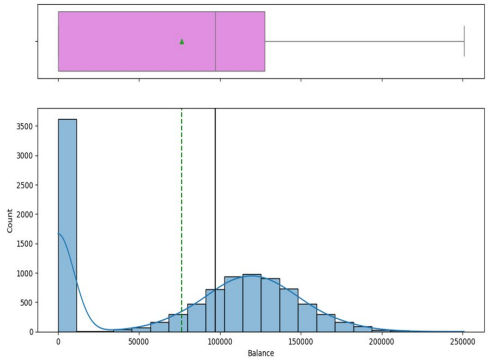
Credit Score (U1)



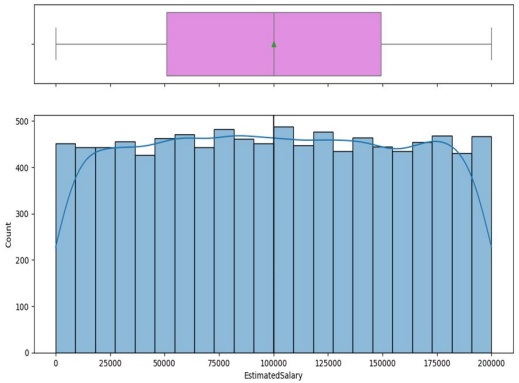
Age (U2)



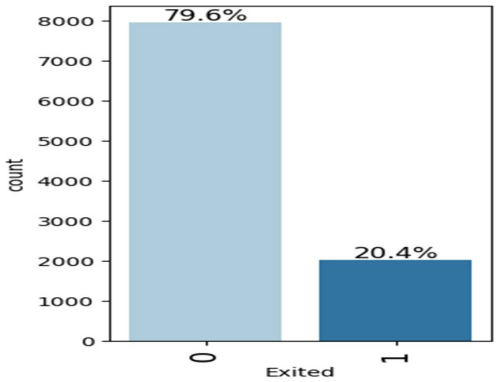
Balance (U3)



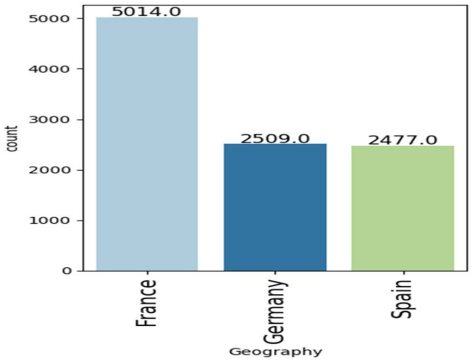
Estimated Salary (U4)



Exited (U5)

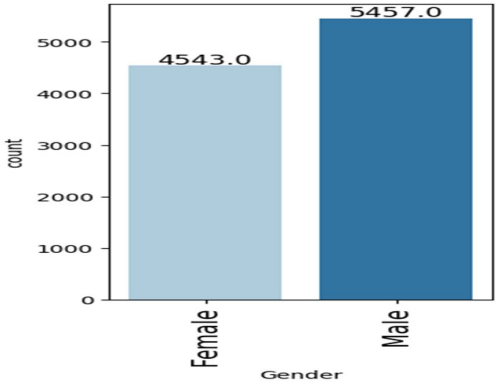


Geography (U6)

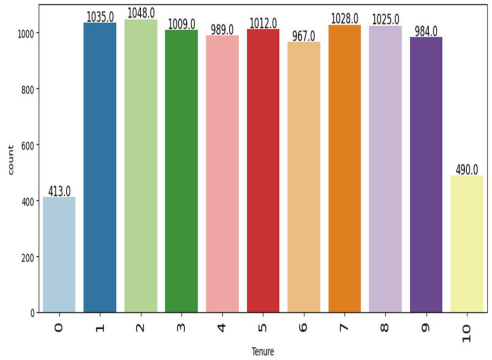


Exploratory Data Analysis- Univariate Data

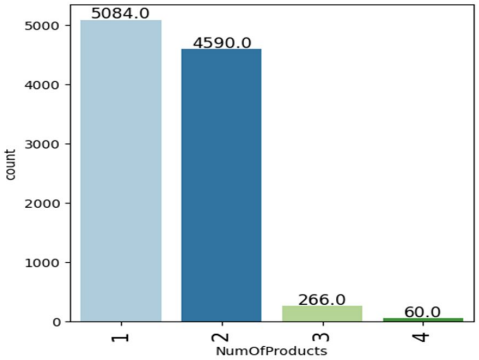
Gender (U7)



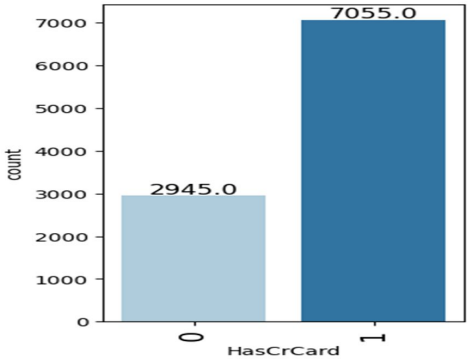
Tenure (U8)



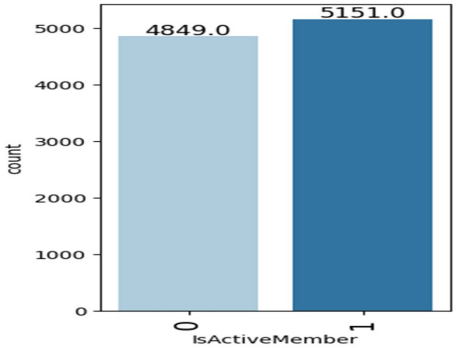
Number of Products (U9)



Has Credit Card (U10)

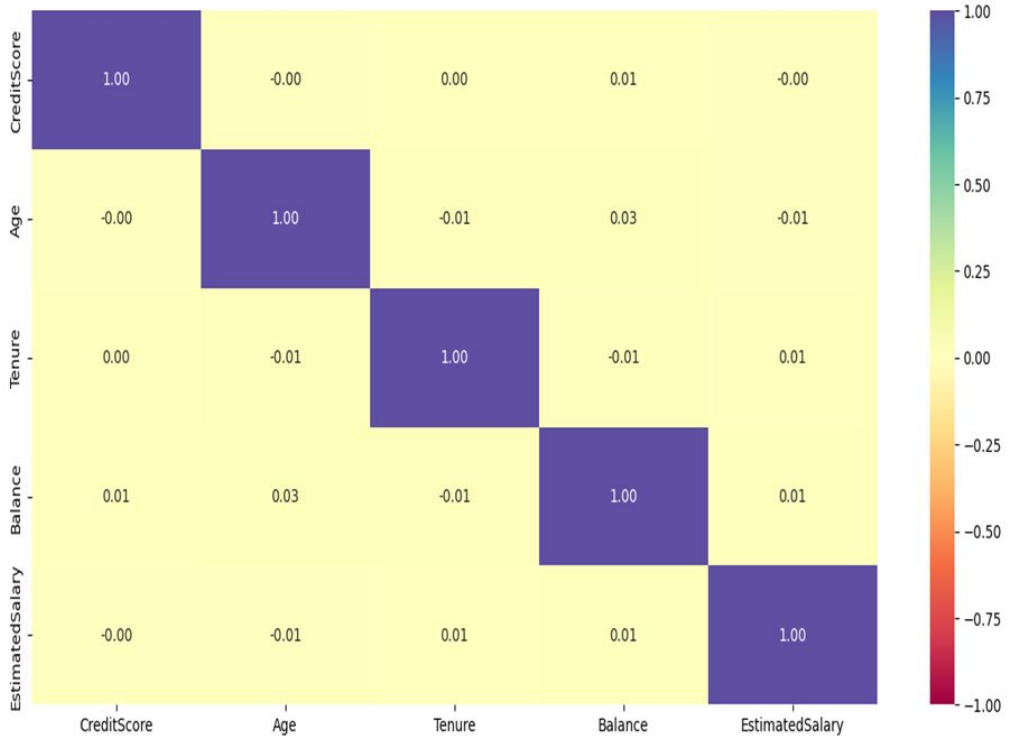


Is Active Member (U11)



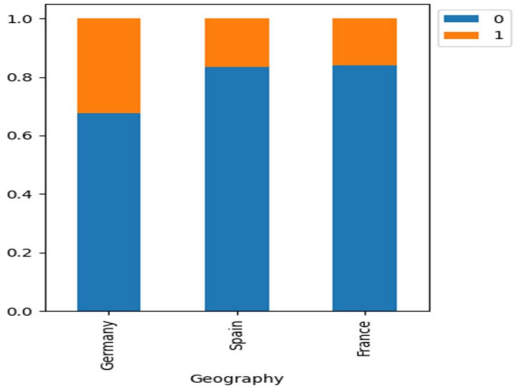
Exploratory Data Analysis- Bivariate Data

B1- Correlation Heatmap

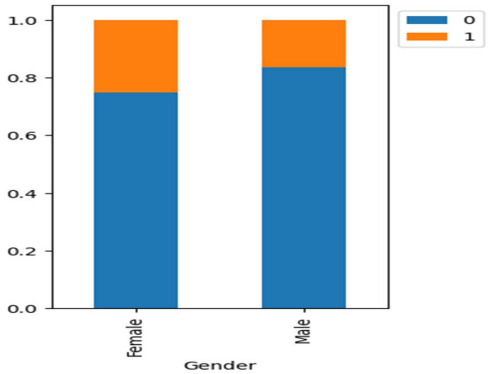


Exploratory Data Analysis- Bivariate Data

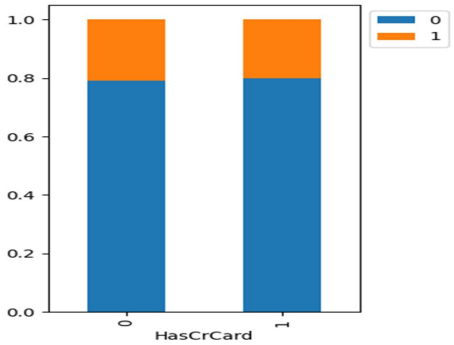
Exited vs Geography (B2)



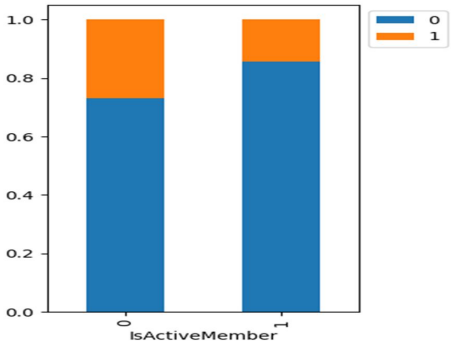
Exited vs Gender (B3)



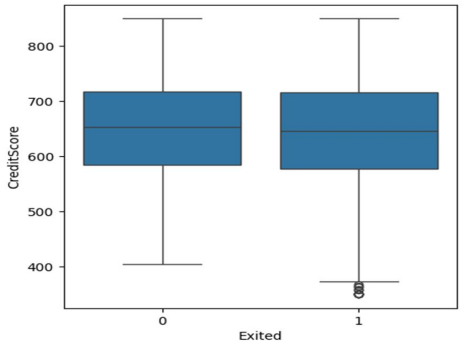
Exited vs Has Credit Card (B4)



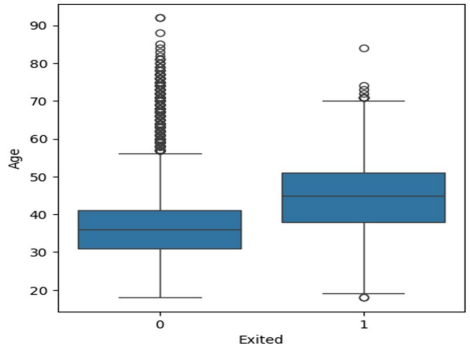
Exited vs Is Active Member (B5)



Exited vs Credit Score (B6)

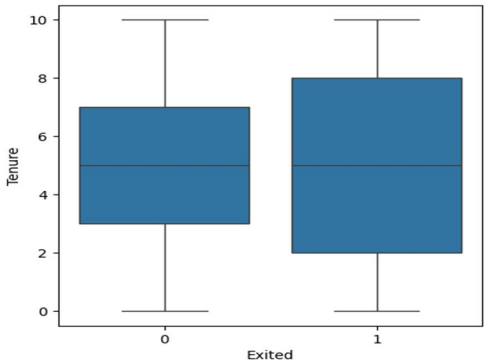


Exited vs Age (B7)

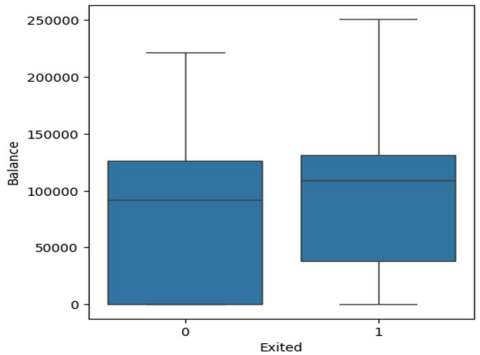


Exploratory Data Analysis- Bivariate Data

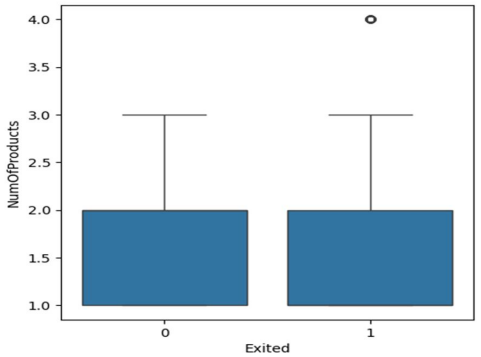
Exited vs Tenure (B8)



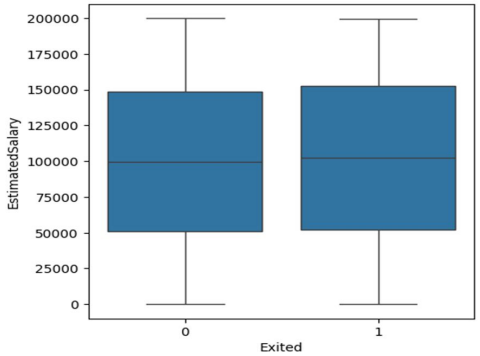
Exited vs Balance (B9)



Exited vs Number of Products (B10)



Exited vs Estimate Salary (B11)



Great Learning

Happy Learning !