



INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT045-3-M-ABAV

ADVANCED BUSINESS ANALYTICS AND VISUALIZATION

APDMF2112DSBA(DE)(PR)

AUGUST 2022

**TITLE: IMPROVING BANKING CUSTOMER RETENTION
BY CUSTOMER CHURN PREDICTION**

LEE KEAN LIM

TP065778

LECTURER: DR. PREETHI SUBRAMANIAN

TABLE OF CONTENTS

TABLE OF CONTENTS.....	i
LIST OF TABLES.....	iii
LIST OF FIGURES	iv
LIST OF ABBREVIATIONS.....	vi
SECTION 1: INTRODUCTION	1
1.1 BACKGROUND.....	1
1.2 PROBLEM STATEMENT	1
1.3 AIM & OBJECTIVES	2
1.3.1 Aim	2
1.3.2 Objectives	2
1.4 SCOPE OF STUDY	2
SECTION 2: METHODOLOGY	3
SECTION 3: EXPLORATORY DATA ANALYSIS	5
3.1 INTRODUCTION.....	5
3.2 DATASET & IMPORT	5
3.2.1 Dataset.....	5
3.2.2 Dataset Import.....	6
3.3 EXPLORATORY DATA ANALYSIS.....	8
3.3.1 Univariate Analysis of Categorical Variables.....	8
3.3.2 Univariate Analysis of Numerical Variables	15
3.3.3 Bivariate Analysis of Variables	21
SECTION 4: IMPLEMENTATION.....	30
4.1 INTRODUCTION.....	30
4.2 DATA PRE-PROCESSING.....	30
4.2.1 Feature Selection.....	30

4.2.2	Feature Encoding	32
4.2.3	Feature Transformation.....	33
4.2.4	Data Balancing.....	34
4.2.5	Data Partitioning	35
4.3	MODEL DEVELOPMENT	36
4.3.1	Logistic Regression Model	36
4.3.2	Neural Network Model	42
4.4	DISCUSSION	52
SECTION 5: CONCLUSION.....		55
REFERENCES		56
APPENDIX A: MODEL STATISTICS & OUTPUTS		57
A.1	Baseline Regression Model Outputs	57
A.2	Poly-2-Regression Model Outputs.....	59
A.3	Poly-3-Regression Model Outputs.....	61
A.4	1 Hidden Layer Neural Network Model Outputs.....	63
A.5	2 Hidden Layer Neural Network Model Outputs.....	64
A.6	3 Hidden Layer Neural Network Model Outputs.....	65
A.7	4 Hidden Layer Neural Network Model Outputs.....	66
A.8	Surrogate Tree Complete List of Node Rules	67

LIST OF TABLES

Table 3.1: Dataset metadata	5
Table 3.2: Snapshot of dataset	7
Table 3.3: Summary statistics of variable Exited and Geography	21
Table 3.4: Summary statistics of variable Exited and Gender	23
Table 3.5: Summary statistics of variable Exited and IsActiveMember	24
Table 3.6: Summary statistics of variable Exited and NumOfProd	26
Table 3.7: Summary statistics of variable Exited and Age	28
Table 4.1: Variable worth	31
Table 4.2: Evaluation metrics for regression baseline model	38
Table 4.3: Evaluation metrics for regression polynomial degree 2 model	39
Table 4.4: Evaluation metrics for regression polynomial degree 3 model	40
Table 4.5: ROC index for regression models	41
Table 4.6: Validation evaluation metrics for regression models	41
Table 4.7: Misclassification of 1 hidden layer neural network models	43
Table 4.8: Evaluation metrics for optimal 1 hidden layer neural network model	44
Table 4.9: Misclassification of 2 hidden layer neural network models	44
Table 4.10: Evaluation metrics for optimal 2 hidden layer neural network model	45
Table 4.11: Misclassification of 3 hidden layer neural network models	46
Table 4.12: Evaluation metrics for optimal 3 hidden layer neural network model	47
Table 4.13: Misclassification of 4 hidden layer neural network models	48
Table 4.14: Evaluation metrics for optimal 4 hidden layer neural network model	49
Table 4.15: ROC index for optimal neural network models	50
Table 4.16: Validation evaluation metrics for optimal neural network models	50

LIST OF FIGURES

Figure 3.1: Dataset import and variable role settings	7
Figure 3.2: Table summary statistics of categorical variables	8
Figure 3.3: Univariate analysis of the variable Gender	9
Figure 3.4: Univariate analysis of the variable Geography	10
Figure 3.5: Univariate analysis of the variable HasCrCard	11
Figure 3.6: Univariate analysis of the variable IsActiveMember	12
Figure 3.7: Univariate analysis of the variable Exited.....	13
Figure 3.8: Univariate analysis of the variable Exited.....	14
Figure 3.9: Table summary statistics of numerical variables	15
Figure 3.10: Univariate analysis of the variable Age	16
Figure 3.11: Box plot for variable Age	16
Figure 3.12: Univariate analysis of the variable Balance	17
Figure 3.13: Box plot for variable Balance.....	17
Figure 3.14: Univariate analysis of the variable CreditScore	18
Figure 3.15: Box plot for variable CreditScore	18
Figure 3.16: Univariate analysis of the variable EstimatedSalary	19
Figure 3.17: Box plot for variable EstimatedSalary	19
Figure 3.18: Univariate analysis of the variable Tenure.....	20
Figure 3.19: Box plot for variable Tenure	20
Figure 3.20: Bivariate analysis of the variables Exited and Geography.....	22
Figure 3.21: Bivariate analysis of the variables Exited and Gender.....	23
Figure 3.22: Bivariate analysis of the variables Exited and IsActiveMember	25
Figure 3.23: Bivariate analysis of the variables Exited and NumOfProducts	27
Figure 3.24: Bivariate analysis of the variables Exited and Age.....	28
Figure 4.1: Overall process flow diagram.....	30
Figure 4.2: Variable worth.....	31
Figure 4.3: Variables dropping node settings	32
Figure 4.4: Feature encoding node settings	33
Figure 4.5: Feature transformation node settings	34
Figure 4.6: Interval variables after log transformation	34
Figure 4.7: Sample node settings	35

Figure 4.8: Data sample after sampling	35
Figure 4.9: Data partition node settings	36
Figure 4.10: Partial process flow diagram for logistic regression model	37
Figure 4.11: Classification result regression baseline model.....	37
Figure 4.12: Classification result regression polynomial degree 2 model.....	38
Figure 4.13: Classification result regression polynomial degree 3 model.....	39
Figure 4.14: ROC chart for regression models	40
Figure 4.15: Maximum likelihood estimates for baseline regression model.....	41
Figure 4.16: Partial process flow diagram for neural network model	42
Figure 4.17: Classification result of optimal 1 hidden layer neural network model.....	43
Figure 4.18: Classification result of optimal 2 hidden layer neural network model.....	45
Figure 4.19: Classification result of optimal 3 hidden layer neural network model.....	47
Figure 4.20: Classification result of optimal 4 hidden layer neural network model.....	48
Figure 4.21: ROC chart for neural network models	49
Figure 4.22: Misclassification rate after pruning for surrogate model	51
Figure 4.23: Node rule number 15 from surrogate model	52
Figure A.1: Analysis of effects for baseline regression model.....	57
Figure A.2: Maximum likelihood estimates for baseline regression model	57
Figure A.3: Fit statistics for baseline regression model.....	58
Figure A.4: Analysis of effects for Poly-2-Regression model.....	59
Figure A.5: Maximum likelihood estimates for Poly-2-Regression model.....	59
Figure A.6: Fit statistics for Poly-2-Regression model	60
Figure A.7: Analysis of effects for Poly-3-Regression model.....	61
Figure A.8: Maximum likelihood estimates for Poly-3-Regression model.....	61
Figure A.9: Fit statistics for Poly-3-Regression model	62
Figure A.10: Fit statistics for 1 hidden layer neural network model	63
Figure A.11: Misclassification rate for 1 hidden layer neural network model	63
Figure A.12: Fit statistics for 2 hidden layer neural network model	64
Figure A.13: Misclassification rate for 2 hidden layer neural network model	64
Figure A.14: Fit statistics for 3 hidden layer neural network model	65
Figure A.15: Misclassification rate for 3 hidden layer neural network model	65
Figure A.16: Fit statistics for 4 hidden layer neural network model	66
Figure A.17: Misclassification rate for 4 hidden layer neural network model	66
Figure A.18: Formula for log transformed variables.....	71

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
DT	Decision Tree
LR	Logistic Regression
ROC	Receiver Operating Characteristic

SECTION 1

INTRODUCTION

1.1 BACKGROUND

The rise of fintech and adoption of newer technologies has introduced greater competition to the traditional banks. According to de Lima Lemos et al. (2022), two thirds of existing banking customers are intending to or has already shifted to the use of fintech as compared to using traditional banks. In addition, bank managers and executives foresee the rise of fintech as a threat to the traditional banking. Therefore, a new challenge exists for the banks to retain the existing customers.

Improving the customer relationship management thus becomes a critical task for the banks to reduce customer churn rate. Which according to Tang et al. (2020), retaining existing customers cost much less than trying to attract new customers in using the banking services. Various studies adopted the machine learning approach in predicting the customer churn rate to provide targeted campaigns or services in order to retain the customers (Broby, 2022; Wang et al., 2020).

1.2 PROBLEM STATEMENT

Predicting customer churn rate in the banking sector carries a few challenges. Generally, banks are working with millions of customers, which utilizing methods of human interventions in predicting the churn rate is not feasible. In addition, customers may have a change in requests at any time which is to be adapted by the banking executives in a very quick manner. However, each banking executive may be handling a large customer pool at any given time which may introduce error in the process of amending the requests by customers. Therefore, there is a need to automate the process of customer churn prediction which allows the detection of early signs of potential customer churn. The use of machine learning would facilitate the customer churn prediction, providing a high predictive accuracy and quick results. Which, targeted strategies can be adopted to such customers to improve the retention rate.

1.3 AIM & OBJECTIVES

1.3.1 Aim

The aim of this study is to predict bank customer churn by utilizing different machine learning algorithms to improve the customer retention rate of a bank.

1.3.2 Objectives

The objectives of this study are as followed:

1. To develop predictive models using different machine learning algorithms for predicting bank customer churn.
2. To evaluate performance of different predictive models in predicting bank customer churn.
3. To identify causes leading to the likelihood of customer churn.
4. To propose suggestions for the bank to implement to reduce customer churn rate.

1.4 SCOPE OF STUDY

1. The study is conducted based on a fictitious banking dataset. The dataset provided comprises of customer data and a binary target variable indicating whether customer is churned or retained. There are 10,000 observations and 14 features available from the dataset. The dataset is retrieved from Kaggle and can be accessed from the following link: <https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers>
2. The data mining process will be performed using SAS Enterprise Miner application. The application provides a drag and drop graphical user interface in the form of a process flow diagram that can be easily navigated when performing the data mining process. The visual view of the data mining process facilitates high interpretability and faster model development time. In addition, various modeling techniques are supported by the application which allow variation and comparison of performance from using different models.
3. The causes of customer churn will be determined based on the information derived only from the descriptive and predictive analysis of the dataset.
4. The selection of predictive algorithms will be limited to what is provided by the SAS Enterprise Miner application.

SECTION 2

METHODOLOGY

This study adopts the SEMMA methodology as a guide to the process of data mining. The methodology comprises of five stages namely **S**ample, **E**xplore, **M**odify, **M**odel, and **A**ssess. The following shows the description of tasks involved in each stage of the SEMMA methodology applicable in this study.

Stage 1: Sample

This stage involves splitting the dataset into two groups, namely the training data and the testing data. The dataset would be partitioned, 80% as the training data and 20% as the testing data. This is performed as an evaluation step for the fitted model. The target variable from the testing dataset will be dropped to allow the fitted model to predict the outcome and evaluate the performance of the prediction.

Stage 2: Explore

Exploration of the data can be performed through exploratory data analysis with the aid of visualization and statistics. Through data exploratory analysis, trends and anomalies in the dataset can be observed which provides a better understanding of the dataset. Visualization methods such as bar chart, histogram, box and whisker plot, etc., can be utilized to understand the relationship between variables and identifying anomalies. The use of visualization provides a graphical means of interpreting the dataset. In addition, complementing the visualization, performing descriptive statistics such as measuring central tendency and identifying frequency distribution would provide further information regarding the dataset.

Stage 3: Modify

The dataset contains a mix of categorical and numerical variables. Encoding of the categorical variables are expected due to many algorithms are not able to work directly with character labels (Tang et al., 2020). Example, label encoding can be performed for the “Gender” feature to convert the character labels into numeric format. Feature scaling is to be performed for numerical variables due to the different ranges of values of each feature. This is to ensure all features are weighed equally by the model thus minimizing the bias.

Several features can be removed from the dataset as it does not provide any impact to the predictability of the model. Such features identified from the dataset are “RowNumber”, “CustomerId”, and “Surname”.

Missing values are not identified from the dataset. However, outliers are present and should be identified and processed.

Stage 4: Model

Two prediction models will be developed using different machine learning techniques, namely Logistic Regression (LR) and Artificial Neural Networks (ANN). The LR models is chosen due to its quick computation and provides high explainability. This allows the identifying of features with high impact and business rules for application. In contrast, ANN does not provide high explainability. However, expectation that the ANN model would yield higher prediction accuracy due to their better fitting ability. Hyperparameter tuning will be performed on the models to attain better prediction results.

Stage 5: Assess

The models will be evaluated by the testing dataset to determine the fitness of model. An iterative process of evaluation and hyperparameter tuning will be performed to achieve a good fitted model for each algorithm. The accuracy, precision, recall, and confusion matrix will be used as the evaluation metrics for the prediction models. Multiple evaluation metrics are used to provide a deeper evaluation of the models, due to the imbalanced data classes typically observed in churn dataset (Tang et al., 2020).

SECTION 3

EXPLORATORY DATA ANALYSIS

3.1 INTRODUCTION

This section outlines the dataset utilized in this study and performs exploratory data analysis to identify potential trends and data quality issues. An in-depth discussion on the identified data trend will be provided. In addition, approaches for data quality issues will be suggested.

3.2 DATASET & IMPORT

3.2.1 Dataset

The dataset utilized in this study represents the demographics and banking information for bank customers in Europe. The dataset consists of 10,000 observations with 14 features. Table 3.1 shows the metadata which provides description for each feature in the dataset. The objective of utilizing this dataset is to be able to predict the potential of customer churn. Therefore, of the 14 features, the “**Exited**” feature is the **dependent variable** representing the churn status of customers, while the other features are the independent variables.

Table 3.1: Dataset metadata


No.	Variable	Description	Measurement Level	Sample Data
1	RowNumber	Row index number	Nominal	1, 2, 3, ..., 10000
2	CustomerID	Index number of customers	Nominal	15634602, 15737173, ...
3	Surname	Surname of customers	Nominal	McDonald, Hsiao, Chiazagomekpere, ...
4	CreditScore	A value representing creditworthiness of customers	Interval	350, 484, 510, ..., 850
5	Geography	Location of customers	Nominal	Spain, Germany, France
6	Gender	Gender of customers	Nominal	Male, Female
7	Age	Age of customers	Interval	18, 39, 50, ..., 92
8	Tenure	Duration of customer as a client	Interval	0, 1, 2, 3, 4, ..., 10


		with the bank in years		
9	Balance	Amount of balance in the bank accounts of customers	Interval	0, 15966.8, 115046.74, ..., 250898.09
10	NumOfProducts	Number of products that a customer has bought from the bank	Nominal	0, 1, 2, 3, 4
11	HasCrCard	Indicates whether a customer is with (1) or without (0) a credit card	Binary	0, 1
12	IsActiveMember	Member activity status either active (1) or inactive (0)	Binary	0, 1
13	EstimatedSalary	Salary of customers	Interval	79084.1, 119346.88, ..., 199992.48
14	Exited	Indicate customer churned (1) or retained (0)	Binary	0, 1

3.2.2 Dataset Import

The dataset will be imported to SAS Enterprise Miner using the File Import tool. The tool converts the dataset file which is a comma separated values file into a format which SAS Enterprise Miner recognizes. Figure 3.1 shows the dataset imported and setting the role of the variables. Several variables are identified to have no impact on the analysis and output, namely “CustomerID”, “Surname”, and “RowNumber”. These variables will be dropped from the dataset. Since the “Exited” variable is the dependent variable, the role of the variable will be set as the target. In addition, the level of each variable will be set according to the metadata provided.

Shown in Table 3.2, the snapshot of the imported dataset with the previously mentioned variables excluded. Thus, there are 11 variables remaining in the dataset.

 **File Import**

 **Variables - FIMPORT**

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining

Name	Role	Level	Report	Order	Drop
Age	Input	Interval	No		No
Balance	Input	Interval	No		No
CreditScore	Input	Interval	No		No
CustomerId	Rejected	Interval	No		Yes
EstimatedSalary	Input	Interval	No		No
Exited	Target	Binary	No		No
Gender	Input	Nominal	No		No
Geography	Input	Nominal	No		No
HasCrCard	Input	Binary	No		No
IsActiveMember	Input	Binary	No		No
NumOfProducts	Input	Nominal	No		No
RowNumber	Rejected	Interval	No		Yes
Surname	Rejected	Nominal	No		Yes
Tenure	Input	Interval	No		No

Figure 3.1: Dataset import and variable role settings

Table 3.2: Snapshot of dataset

CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
619	France	Female	42	2	0	1	1	1	101348.88	1
608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
502	France	Female	42	8	159660.8	3	1	0	113931.57	1
699	France	Female	39	1	0	2	0	0	93826.63	0
850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
822	France	Male	50	7	0	2	1	1	10062.8	0
376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
501	France	Male	44	4	142051.07	2	0	1	74940.5	0
684	France	Male	27	2	134603.88	1	1	1	71725.73	0
528	France	Male	31	6	102016.72	2	0	0	80181.12	0
497	Spain	Male	24	3	0	2	1	0	76390.01	0
476	France	Female	34	10	0	2	1	0	26260.98	0
549	France	Female	25	5	0	2	0	0	190857.79	0
635	Spain	Female	35	7	0	2	1	1	65951.65	0
616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0
653	Germany	Male	58	1	132602.88	1	1	0	5097.67	1
549	Spain	Female	24	9	0	2	1	1	14406.41	0
587	Spain	Male	45	6	0	1	0	0	158684.81	0
726	France	Female	24	6	0	2	1	1	54724.03	0
732	France	Male	41	8	0	2	1	1	170886.17	0
636	Spain	Female	32	8	0	2	1	0	138555.46	0
510	Spain	Female	38	4	0	1	1	0	118913.53	1

3.3 EXPLORATORY DATA ANALYSIS

Univariate analysis will be performed for every categorical and numerical variable to identify trends and data anomalies. Following that, bivariate analysis for selected combination of variables will be performed to further identify trends and hidden patterns in the dataset.

3.3.1 Univariate Analysis of Categorical Variables

This section analyzes the seven categorical variables found in the dataset. The analysis method used would include univariate analysis for each categorical variable. Interpretation of the results to identify trends and anomalies will be performed.

Figure 3.2 shows the table summary statistics for categorical variables. The results are generated using the StatExplore tool. While the following bar charts are generated using the Graph Explore tool.

Class Variable Summary Statistics (maximum 500 observations printed)								
Data Role=TRAIN								
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Gender	INPUT	2	0	Male	54.57	Female	45.43
TRAIN	Geography	INPUT	3	0	France	50.14	Germany	25.09
TRAIN	HasCrCard	INPUT	2	0	1	70.55	0	29.45
TRAIN	IsActiveMember	INPUT	2	0	1	51.51	0	48.49
TRAIN	NumOfProducts	INPUT	4	0	1	50.84	2	45.90
TRAIN	Exited	TARGET	2	0	0	79.63	1	20.37

Figure 3.2: Table summary statistics of categorical variables

3.3.1.1 Univariate Analysis for Categorical Variable – Gender

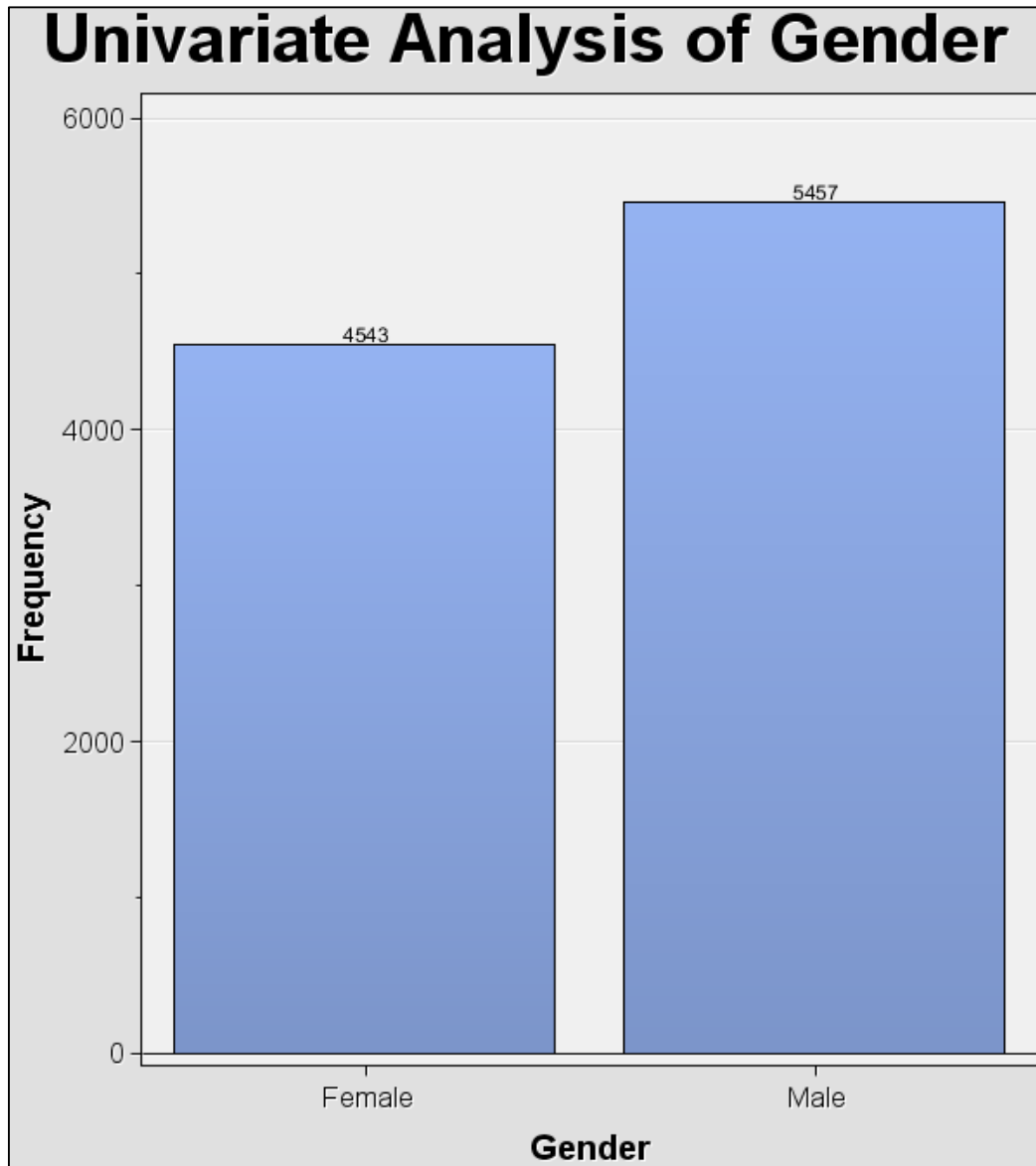


Figure 3.3: Univariate analysis of the variable Gender

Figure 3.3 shows the graphical analysis output from univariate analysis for the variable Gender. Based on Figure 3.2, there is no missing value identified in this variable. Of the 10,000 customers, there are 5457 male customers (54.57%) and 4543 female customers (45.43%). In overall, this indicates that there are more male customers as compared to female customers.

3.3.1.2 Univariate Analysis for Categorical Variable – Geography

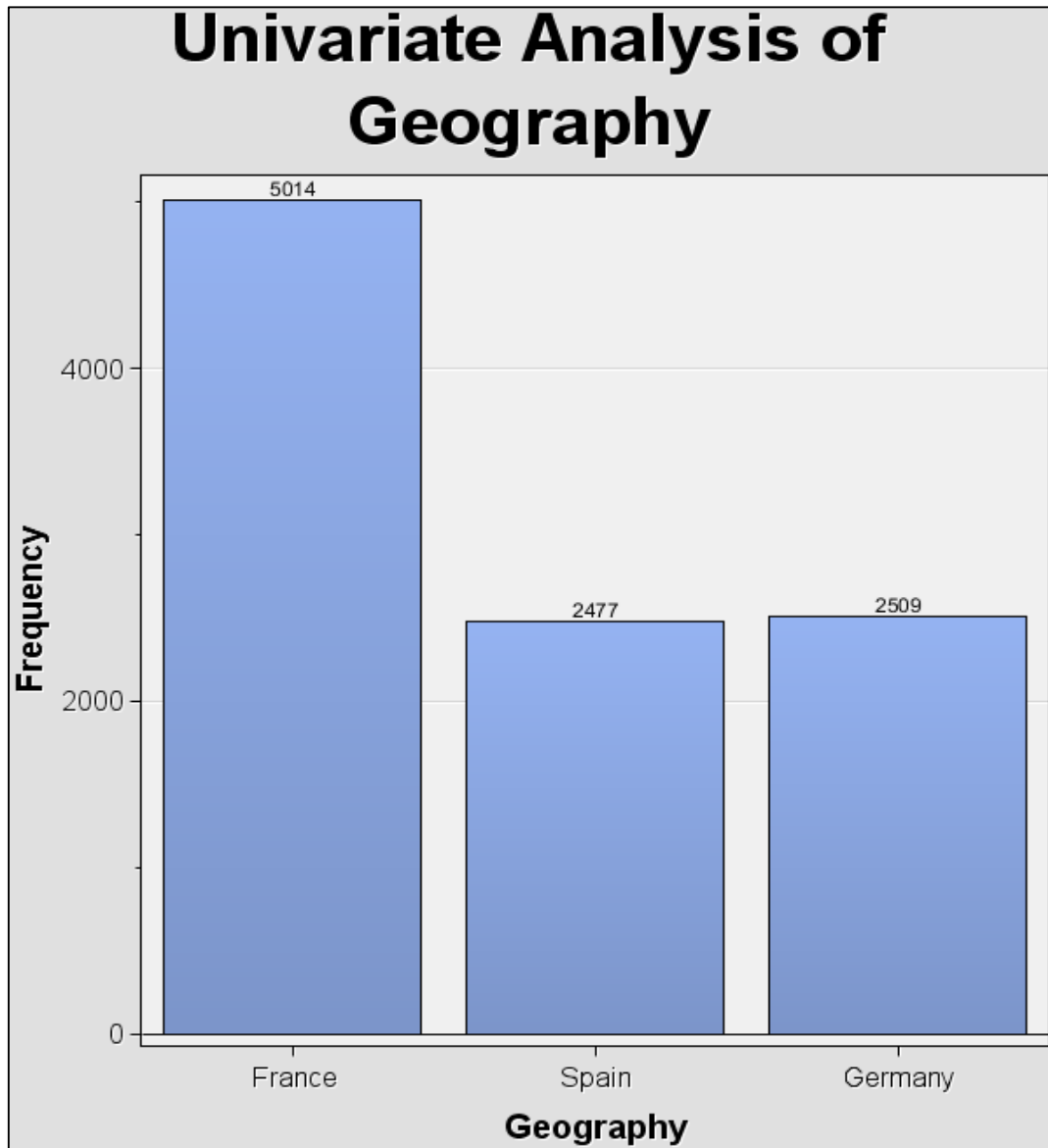


Figure 3.4: Univariate analysis of the variable Geography

Figure 3.4 shows the graphical analysis output from univariate analysis for the variable Geography. Based on Figure 3.2, there is no missing value identified in this variable. Of the 10,000 customers, there are 5014 customers from France (50.14%), 2477 customers from Spain (24.77%), and 2509 customers from Germany (25.09%). In overall, this indicates that France has the highest proportion of customers as compared to Spain and Germany. While Spain and Germany have a roughly equivalent number of customers.

3.3.1.3 Univariate Analysis for Categorical Variable – HasCrCard

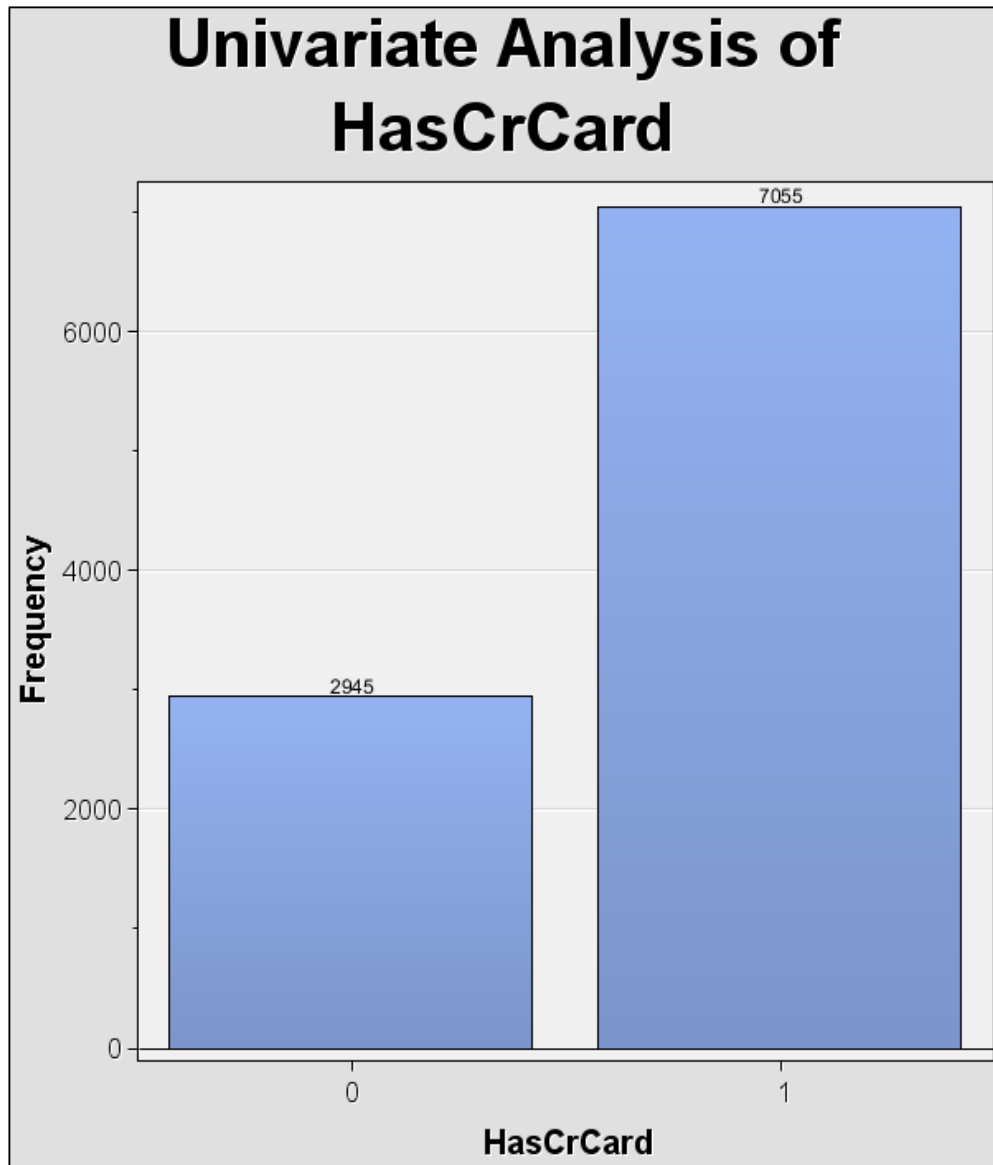


Figure 3.5: Univariate analysis of the variable HasCrCard

Figure 3.5 shows the graphical analysis output from univariate analysis for the variable HasCrCard. Based on Figure 3.2, there is no missing value identified in this variable. Of the 10,000 customers, there are 2945 customers without a credit card (29.45%) and 7055 customers with at least one credit card (70.55%). In overall, this indicates that more customers have credit card as compared to customers without a credit card.

3.3.1.4 Univariate Analysis for Categorical Variable – IsActiveMember

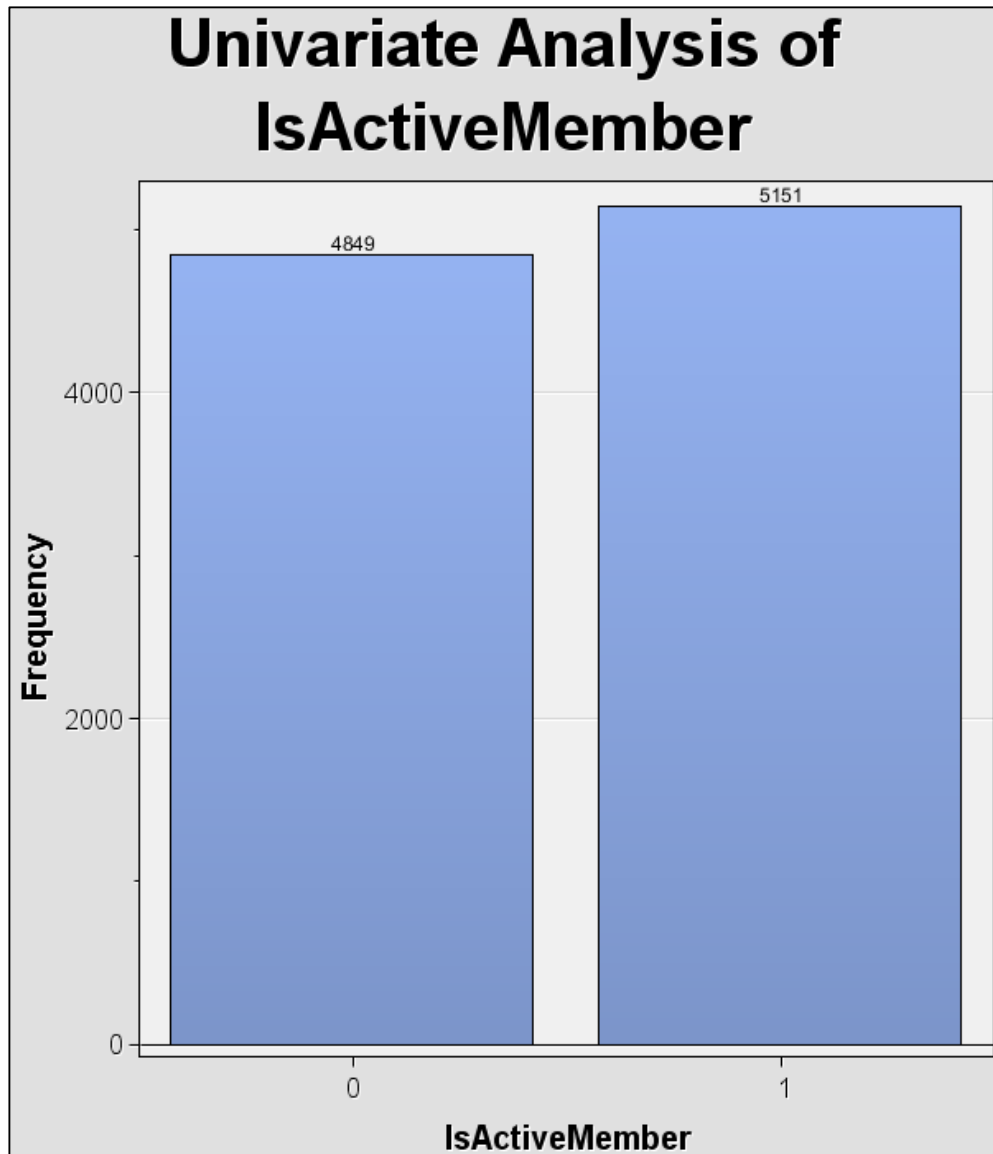


Figure 3.6: Univariate analysis of the variable IsActiveMember

Figure 3.6 shows the graphical analysis output from univariate analysis for the variable IsActiveMember. Based on Figure 3.2, there is no missing value identified in this variable. Of the 10,000 customers, there are 4849 customers who are non-active (48.49%) and 5151 customers who are active (51.51%). In overall, this indicates that there is slightly more active customers as compared to non-active customers.

3.3.1.5 Univariate Analysis for Categorical Variable – NumOfProducts

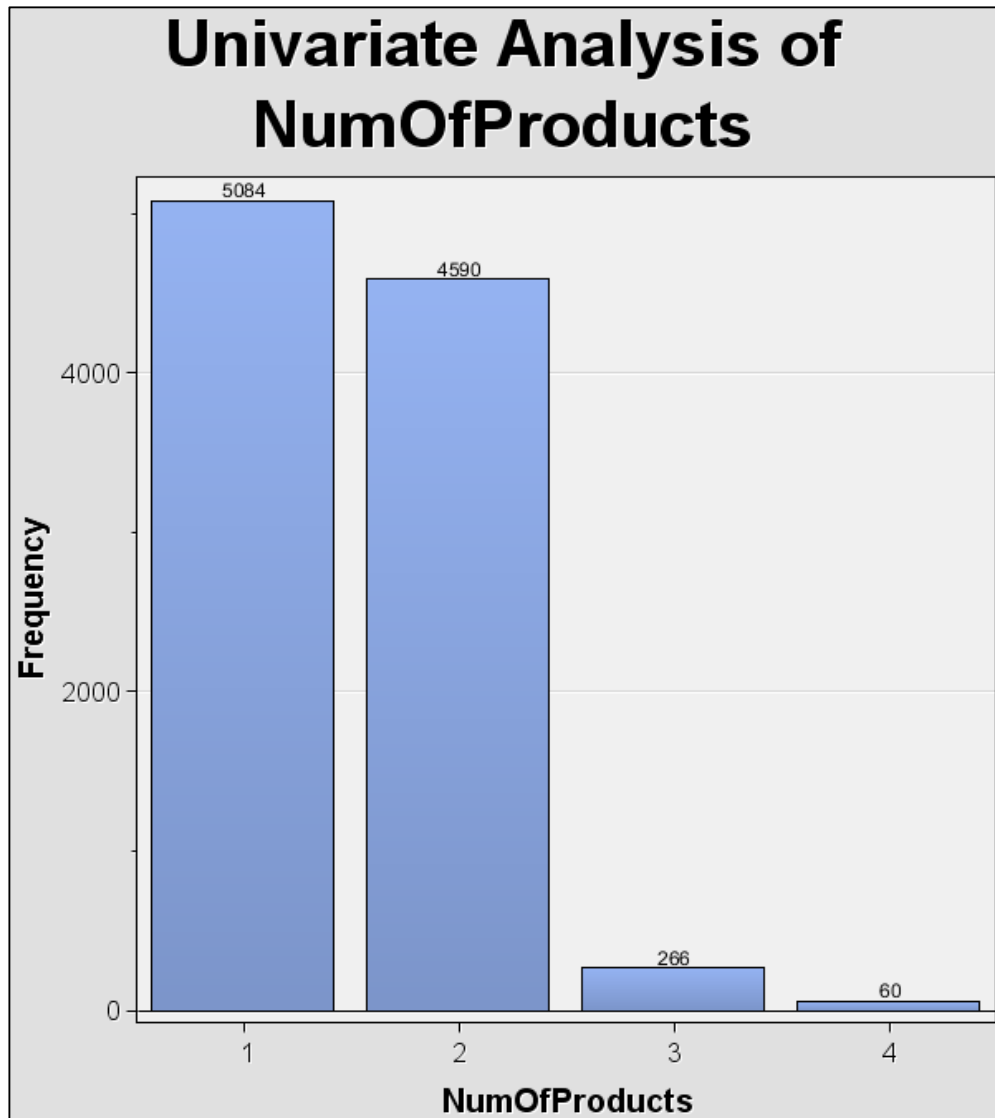


Figure 3.7: Univariate analysis of the variable Exited

Figure 3.7 shows the graphical analysis output from univariate analysis for the variable NumOfProducts. Based on Figure 3.2, there is no missing value identified in this variable. Of the 10,000 customers, there are 5084 customers who purchased one product (50.84%), 4590 customers purchased two products (45.90%), 266 customers purchased three products (2.66%), and 60 customers purchased four products (0.60%). In overall, this indicates that majority of the customers are purchasing one or two products from the bank while very minority of customers purchased more than two products from the bank.

3.3.1.6 Univariate Analysis for Categorical Variable – Exited

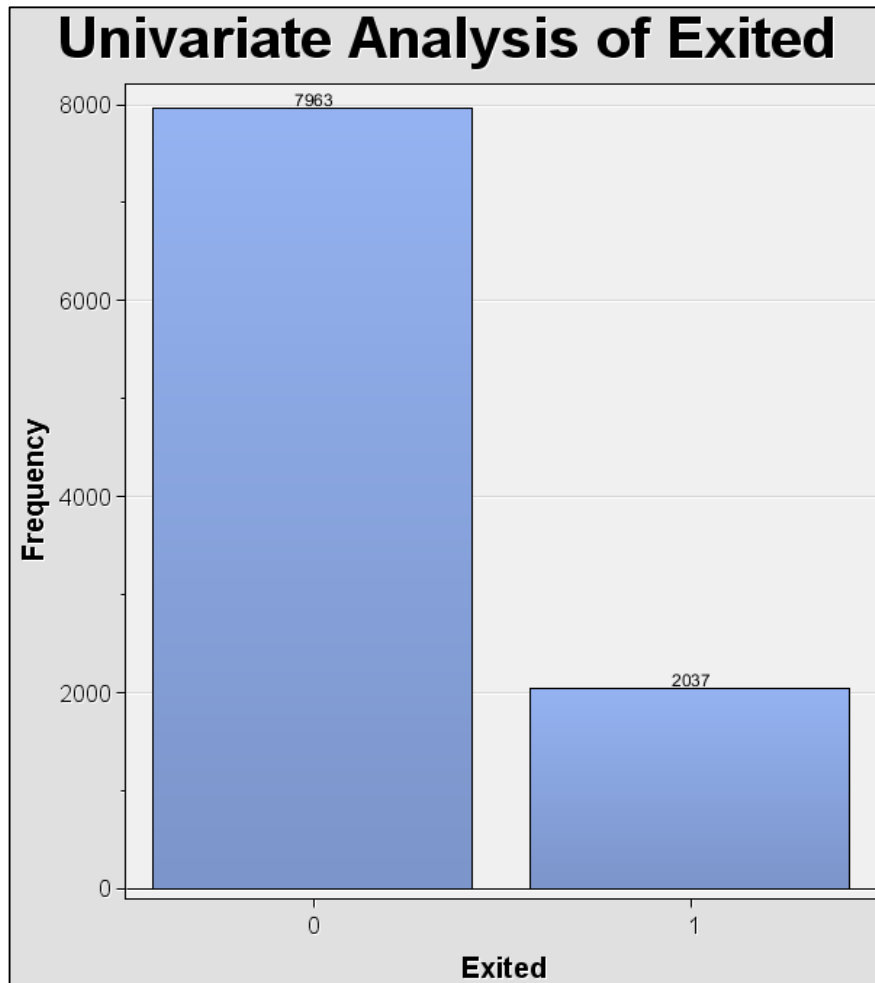


Figure 3.8: Univariate analysis of the variable Exited

Figure 3.8 shows the graphical analysis output from univariate analysis for the variable Exited. Based on Figure 3.2, there is no missing value identified in this variable. Of the 10,000 customers, there are 7963 customers retained (79.63%) and 2037 customers have churned (20.37%). In overall, this indicates that there are more customers retained as compared to customers who have churned.

This variable is the target variable of this study. A data imbalance issue between the two classes of this variable is observed. Bias can be introduced in the prediction models due to the highly disproportionate classes. Data balancing is highly encouraged for this dataset to ensure bias in the prediction models are minimized.

3.3.2 Univariate Analysis of Numerical Variables

This section analyzes the five numerical variables found in the dataset. The analysis method used would include univariate analysis for each numerical variable. Interpretation of the results to identify trends and anomalies will be performed.

Figure 3.7 shows the table summary statistics for every numerical variable in the dataset. The results are generated using the StatExplore tool. While the following histograms and box plots for each variable are generated using the Graph Explore tool.

Interval Variable Summary Statistics (maximum 500 observations printed)										
Data Role=TRAIN										
Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	38.9218	10.48781	10000	0	18	37	92	1.01132	1.395347
Balance	INPUT	76485.89	62397.41	10000	0	0	97188.62	250898.1	-0.14111	-1.48941
CreditScore	INPUT	650.5288	96.6533	10000	0	350	652	850	-0.07161	-0.42573
EstimatedSalary	INPUT	100090.2	57510.49	10000	0	11.58	100187.4	199992.5	0.002085	-1.18152
Tenure	INPUT	5.0128	2.892174	10000	0	0	5	10	0.010991	-1.16523

Figure 3.9: Table summary statistics of numerical variables

3.3.2.1 Univariate Analysis for Numerical Variable – Age

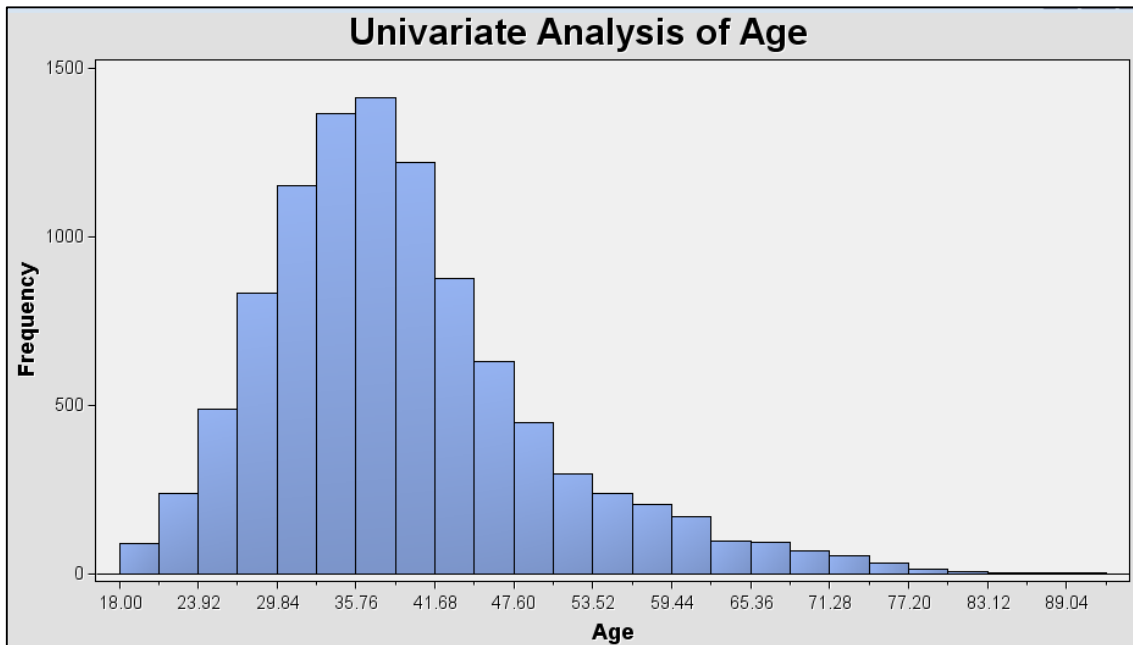


Figure 3.10: Univariate analysis of the variable Age

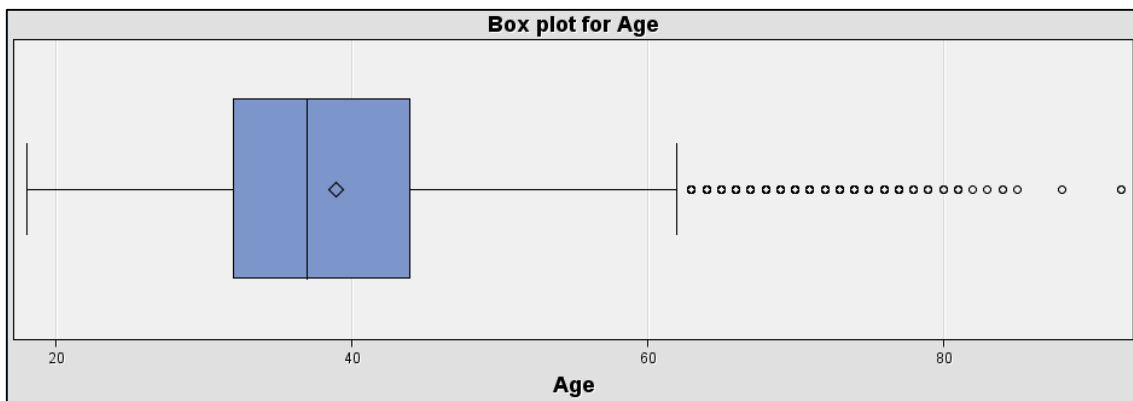


Figure 3.11: Box plot for variable Age

Figure 3.10 shows the histogram while Figure 3.11 shows the box plot from the univariate analysis for the variable Age. Based on Figure 3.9, there is no missing value identified in this variable. The variable has a range between 18 and 92. In addition, the mean age identified is 38.92 which is greater than the median age of 37, which indicates a skew in the data distribution. It is evident based on the figures, that the distribution is slightly positively skewed. This indicates that majority of customers have an age between 30 to 40. The box plot indicates outliers are present at the extreme end of the distribution.

3.3.2.2 Univariate Analysis for Numerical Variable – Balance

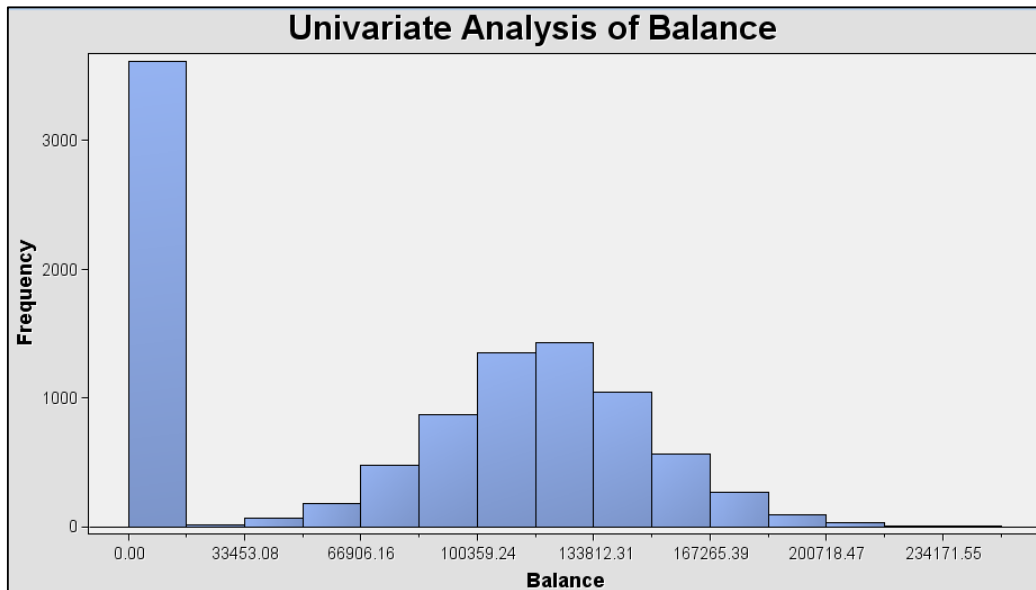


Figure 3.12: Univariate analysis of the variable Balance

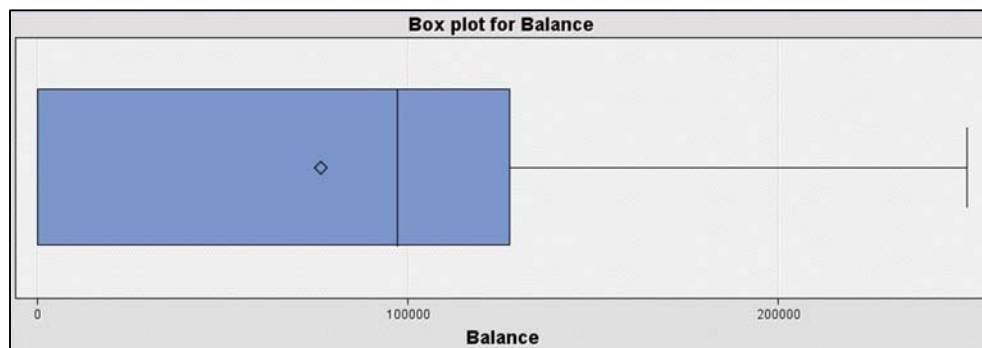


Figure 3.13: Box plot for variable Balance

Figure 3.12 shows the histogram while Figure 3.13 shows the box plot from the univariate analysis for the variable Balance. Based on Figure 3.9, there is no missing value identified in this variable. The variable has a range between 250,898.1. In addition, the mean balance identified is 76,485.89 which is lower than the median balance of 97,188.62, which indicates a skew in the data distribution. It is evident based on the figures, that the distribution is negatively skewed. This indicates that majority of customers have an account balance on the higher range. However, there exist a major number of customers with account balance of zero. This finding is worth further investigation to understand why many customers have a zero account balance. The box plot indicates no outliers are present.

3.3.2.3 Univariate Analysis for Numerical Variable – CreditScore

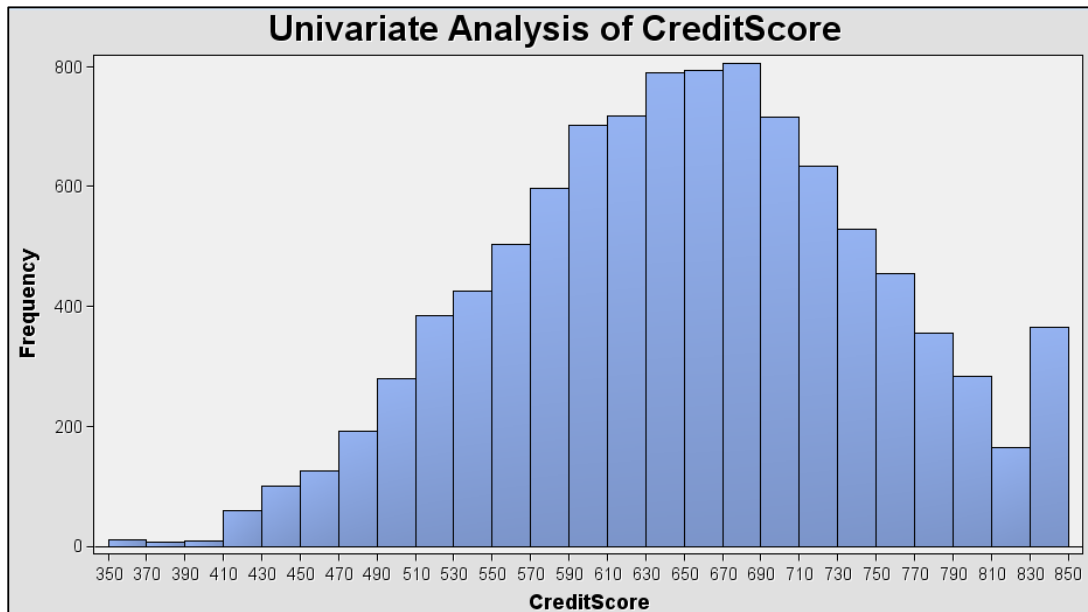


Figure 3.14: Univariate analysis of the variable CreditScore

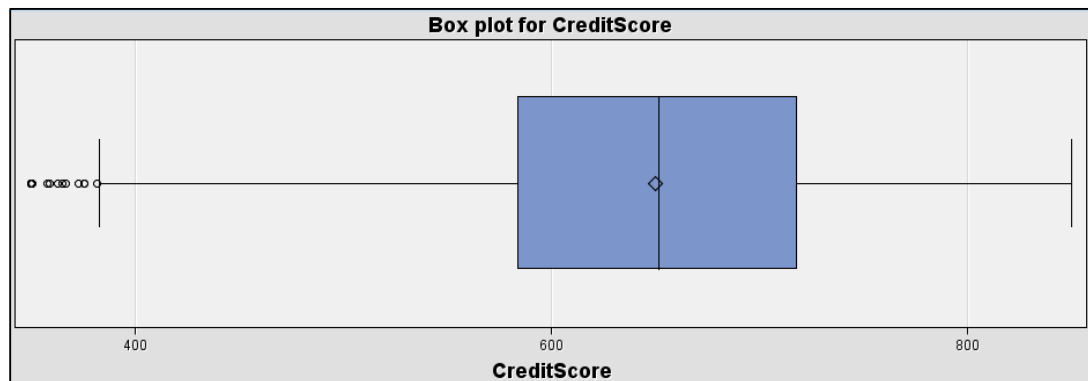


Figure 3.15: Box plot for variable CreditScore

Figure 3.14 shows the histogram while Figure 3.15 shows the box plot from the univariate analysis for the variable CreditScore. Based on Figure 3.9, there is no missing value identified in this variable. The variable has a range between 350 to 850. In addition, the mean credit score identified is 650.53 which is lower than the median credit score of 652, which indicates a skew in the data distribution. It is evident based on the figures, that the distribution is negatively skewed. This indicates that majority of customers have credit score on the higher range. However, there exist a high number of customers with a credit score between 830 and 850. The box plot indicates outliers are present at the extreme end of the distribution.

3.3.2.4 Univariate Analysis for Numerical Variable – EstimatedSalary

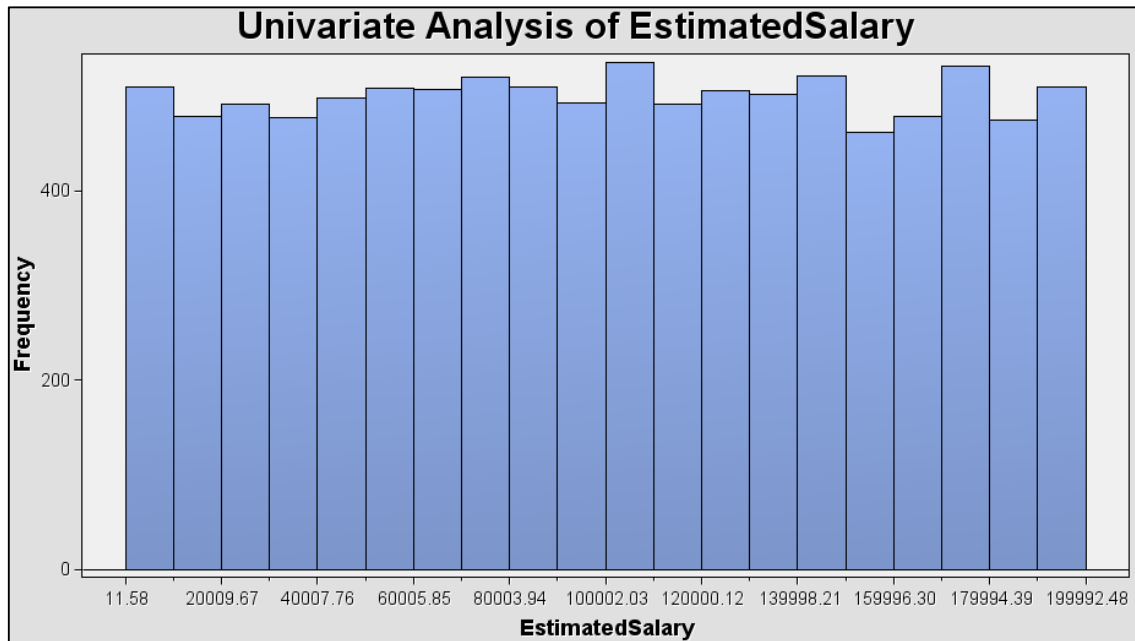


Figure 3.16: Univariate analysis of the variable EstimatedSalary

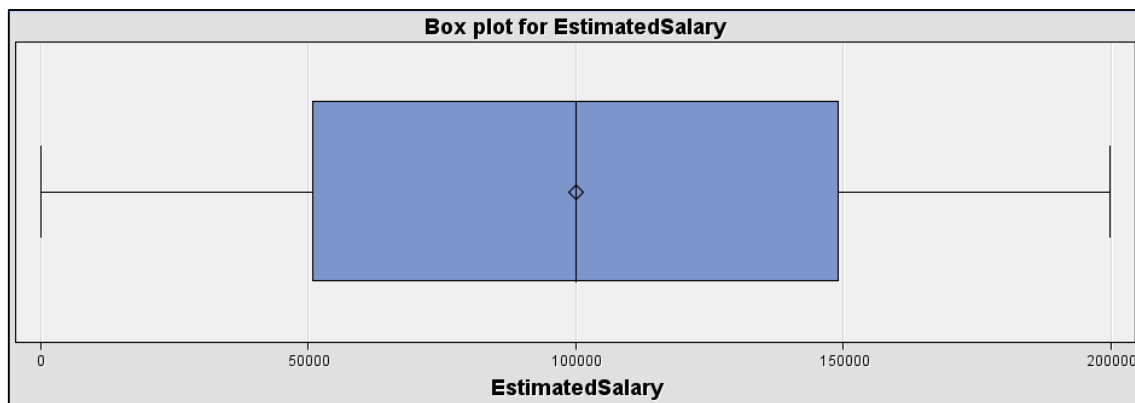


Figure 3.17: Box plot for variable EstimatedSalary

Figure 3.16 shows the histogram while Figure 3.17 shows the box plot from the univariate analysis for the variable EstimatedSalary. Based on Figure 3.9, there is no missing value identified in this variable. The variable has a range between 11.50 to 199,992.5. In addition, the mean salary identified is 100,090.2 which is lower than the median salary of 100,187.4, which indicates a skew in the data distribution. However, based on the figures, the data distribution exhibits a uniformly distributed manner. Thus, no pattern can be identified. The box plot indicates no outliers are present.

3.3.2.5 Univariate Analysis for Numerical Variable – Tenure

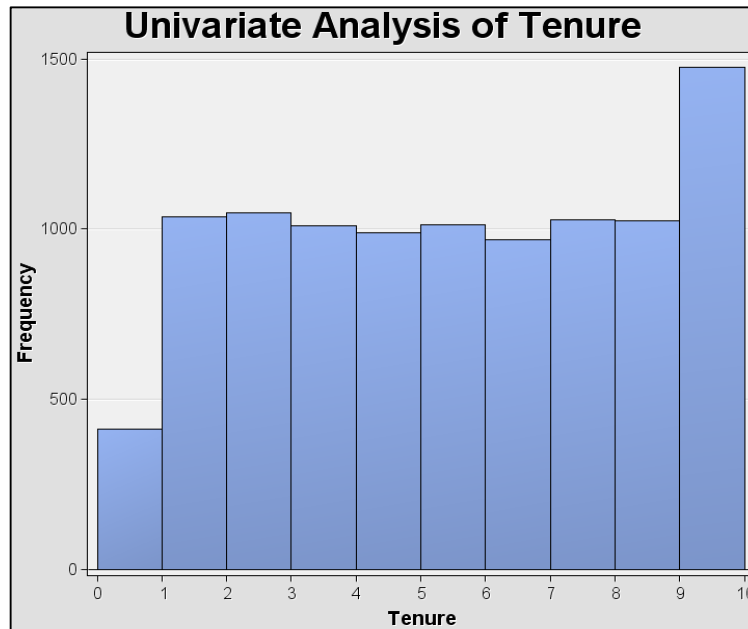


Figure 3.18: Univariate analysis of the variable Tenure

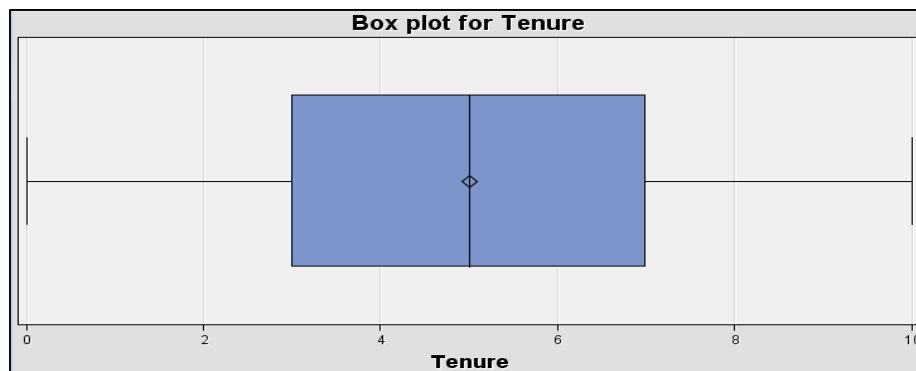


Figure 3.19: Box plot for variable Tenure

Figure 3.18 shows the histogram while Figure 3.19 shows the box plot from the univariate analysis for the variable Tenure. Based on Figure 3.9, there is no missing value identified in this variable. The variable has a range between 0 to 10. In addition, the mean tenure identified is 5.01 which is slightly lower than the median tenure of 5, which indicates a skew in the data distribution. However, based on the figures, the data distribution exhibits a uniformly distributed manner. Thus, no pattern can be identified. However, there exist a abnormally low number of customers with zero year tenure and an abnormally high number of customers with the ten year tenure. The box plot indicates no outliers are present.

3.3.3 Bivariate Analysis of Variables

This section analyzes five combinations of two variables found in the dataset. The analysis method used would include bivariate analysis. The five combinations of variables are chosen based on the unique trends observed that may provide additional insights to the study. It is chosen based on the previous descriptive analysis performed in part B of the assignment. Interpretation of the results to identify trends and anomalies will be performed.

In each analysis, a table output of summary statistics and graphical representation will be displayed. The table of summary statistics are adopted based on results generated from the StatExplore tool while the graphical representations are generated using the Graph Explore tool.

3.3.3.1 Bivariate Analysis for Variables – Exited and Geography

Table 3.3 shows the table summary statistics and Figure 3.20 shows the graphical representation for the variable combination of Exited and Geography.

Table 3.3: Summary statistics of variable Exited and Geography

Geography	Measures	Exited		
		0	1	Total
France	Frequency	4204	810	5014
	Row %	83.84%	16.16%	50.14%
	Column %	52.79%	39.76%	
Spain	Frequency	2064	413	2477
	Row %	83.32%	16.68%	24.77%
	Column %	25.92%	20.27%	
Germany	Frequency	1695	814	2509
	Row %	67.55%	32.45%	25.09%
	Column %	21.29%	39.97%	
Total	Frequency	7963	2037	10000
	Percentage	79.63%	20.37%	100%

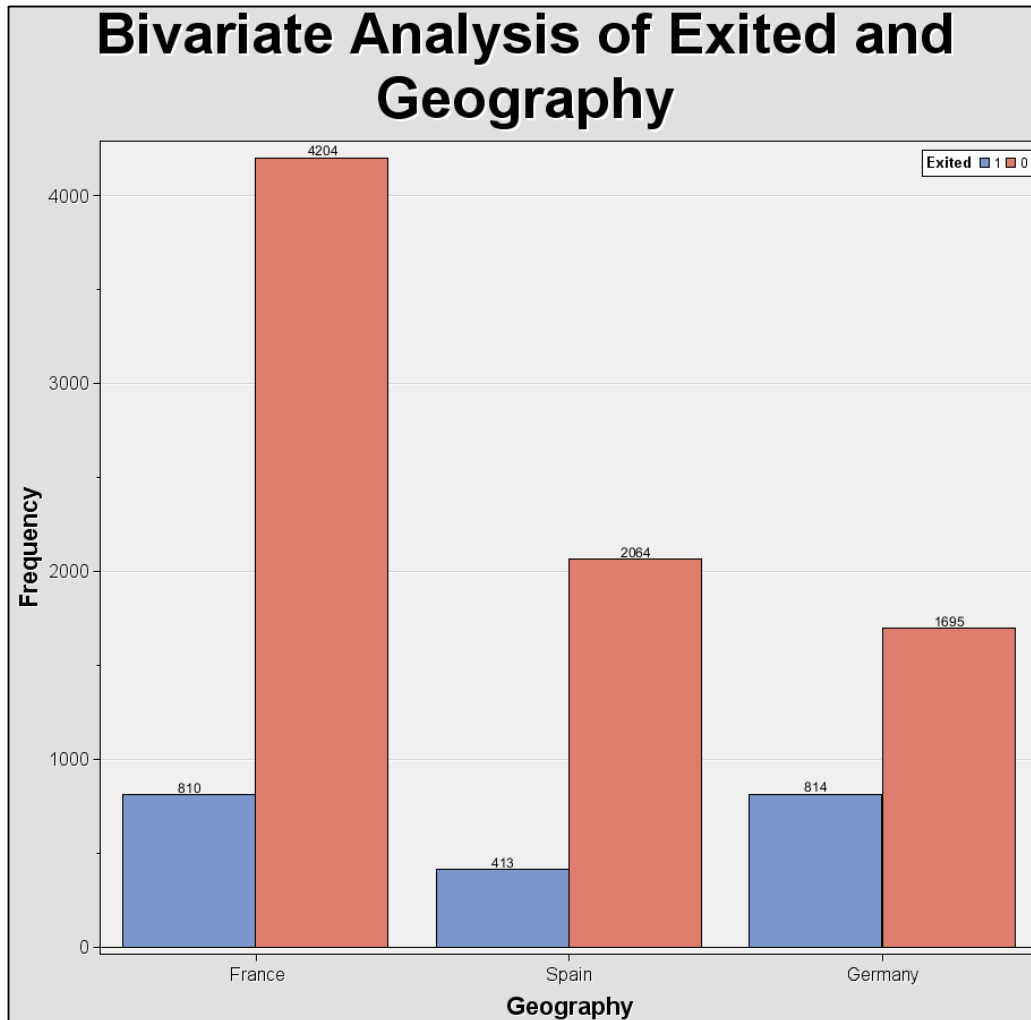


Figure 3.20: Bivariate analysis of the variables Exited and Geography

In overall, France has 5014 customers which is the highest as compared to Spain which has 2477 customers and Germany having 2509 customers. Looking at the number of customers churned, it can be identified that Germany has 814 customers churned which is the highest, followed by France which has 810 customers churned and Spain which has 413 customers churned. However, focused on each country individually, it can be identified that Germany has the highest proportion of customer churned (32.45%) as compared to France (16.16%) and Spain (16.68%). Further investigation should focus on Germany to identify the causes of high customer churn rate as compared to other countries. Probable reason of the higher churn rate observed in Germany is due to the highly competitive nature of the banking sector in Germany. Where more attractive offerings are provided constantly by different banks to attract customers. This results in customers jumping over to the competitors that provide better offerings thus a higher proportion of customer churned observed in Germany.

3.3.3.2 Bivariate Analysis for Variables – Exited and Gender

Table 3.4 shows the table summary statistics and Figure 3.21 shows the graphical representation for the variable combination of Exited and Gender.

Table 3.4: Summary statistics of variable Exited and Gender

Gender	Measures	Exited		
		0	1	Total
Female	Frequency	3404	1139	4543
	Row %	74.93%	25.07%	45.43%
	Column %	42.75%	55.92%	
Male	Frequency	4559	898	5457
	Row %	83.54%	16.46%	54.57%
	Column %	57.25%	44.08%	
Total	Frequency	7963	2037	10000
	Percentage	79.63%	20.37%	100%

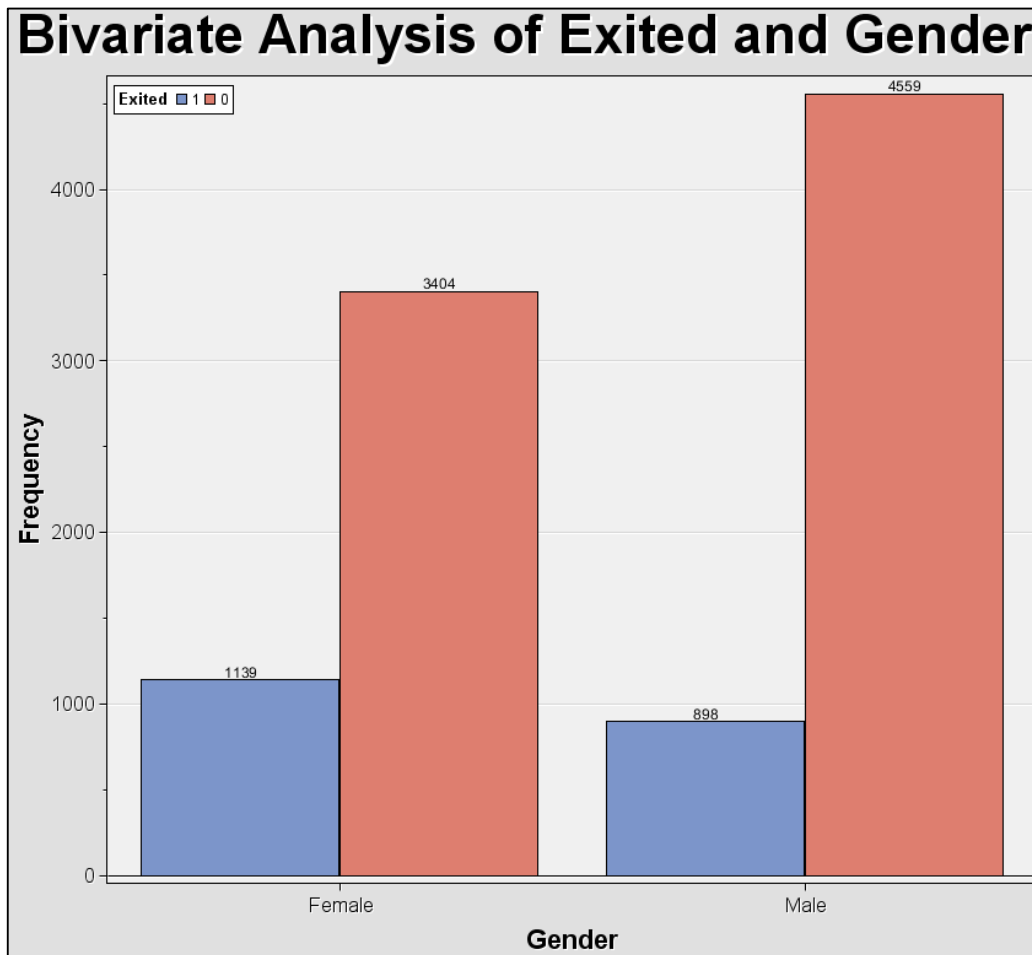


Figure 3.21: Bivariate analysis of the variables Exited and Gender

In overall, there are 4543 female customers, which is lower than the number of male customers which have 5457. Looking at the number of customers churned, it can be identified that the female category has 1139 customers churned, which is higher than the male category which has 898 customers churned. Focused into each gender individually, it can be identified that the female customers have a higher proportion of customer churned (25.07%) as compared to the male customers (16.46%). Further investigation should focus on the female customers to identify the causes of high customer churn rate as compared to the male customers. Probable reason of the higher churn rate observed in the female customers is due to the likelihood that females are a bigger spender as compared to males in specifically shopping. Banks are known to offer various reward points and cash back campaigns when using their bank offered cards for shopping transactions. The bank that offers the best rewards for shoppers may attract more customers. This results in female customers attracted and switching to the bank that offer more rewards for each shopping transaction.

3.3.3.3 Bivariate Analysis for Variables – Exited and IsActiveMember

Table 3.5 shows the table summary statistics and Figure 3.22 shows the graphical representation for the variable combination of Exited and IsActiveMember.

Table 3.5: Summary statistics of variable Exited and IsActiveMember

IsActiveMember	Measures	Exited		
		0	1	Total
0	Frequency	3547	1302	4849
	Row %	73.15%	26.85%	48.49%
	Column %	44.54%	63.92%	
1	Frequency	4416	735	5151
	Row %	85.73%	14.27%	51.51%
	Column %	55.46%	36.08%	
Total	Frequency	7963	2037	10000
	Percentage	79.63%	20.37%	100%

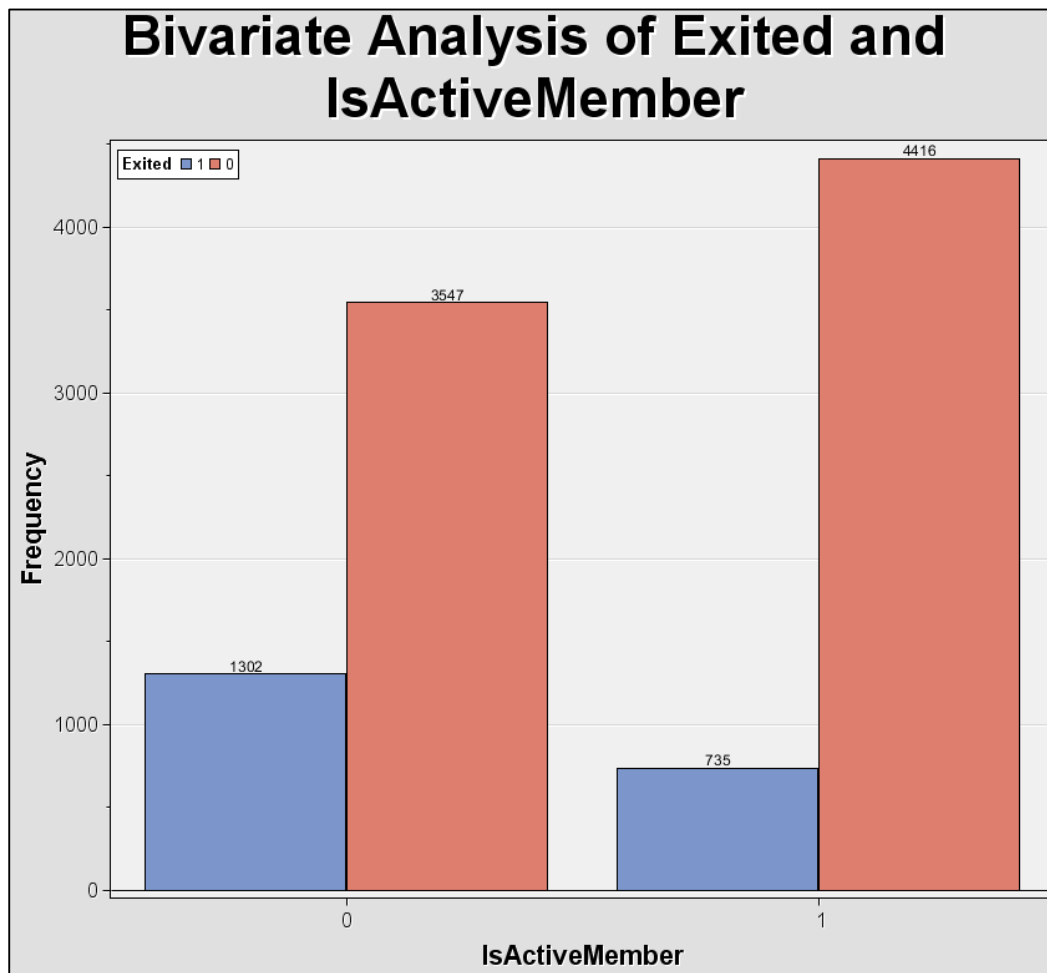


Figure 3.22: Bivariate analysis of the variables Exited and IsActiveMember

In overall, there are 4849 inactive customers, which is lower than the number of active customers which have 5151. Looking at the number of customers churned, it can be identified that the inactive members have 1302 customers churned, which is greater than the active members which have 735 customers churned. Focused into each activity status individually, it can be identified that the inactive members have a higher proportion of customer churned (26.85%) as compared to the active members (14.27%). Further investigation should focus on the inactive members to identify the causes of higher customer churn rate as compared to the active members. Probable reason of the higher churn rate observed in the inactive members is due to the likelihood that the customers no longer have interest in the bank which also explains the inactiveness of the members. This may be due to the services provided by the bank, not meeting the expectations of the customers thus leading to the customers reducing the likelihood to use the bank services. This results in customer churn where the customers are likely to switch to other banks that offer services at their expectation levels.

3.3.3.4 Bivariate Analysis for Variables – Exited and NumOfProd

Table 3.6 shows the table summary statistics and Figure 3.23 shows the graphical representation for the variable combination of Exited and NumOfProd.

Table 3.6: Summary statistics of variable Exited and NumOfProd

NumOfProducts	Measures	Exited		
		0	1	Total
1	Frequency	3675	1409	5084
	Row %	72.29%	27.71%	50.84%
	Column %	46.15%	69.17%	
2	Frequency	4242	348	4590
	Row %	92.42%	7.58%	45.90%
	Column %	53.27%	17.08%	
3	Frequency	46	220	266
	Row %	17.29%	82.71%	2.66%
	Column %	0.58%	10.80%	
4	Frequency	0	60	60
	Row %	0%	100%	0.60%
	Column %	0%	3.67%	
Total	Frequency	7963	2037	10000
	Percentage	79.63%	20.37%	100%

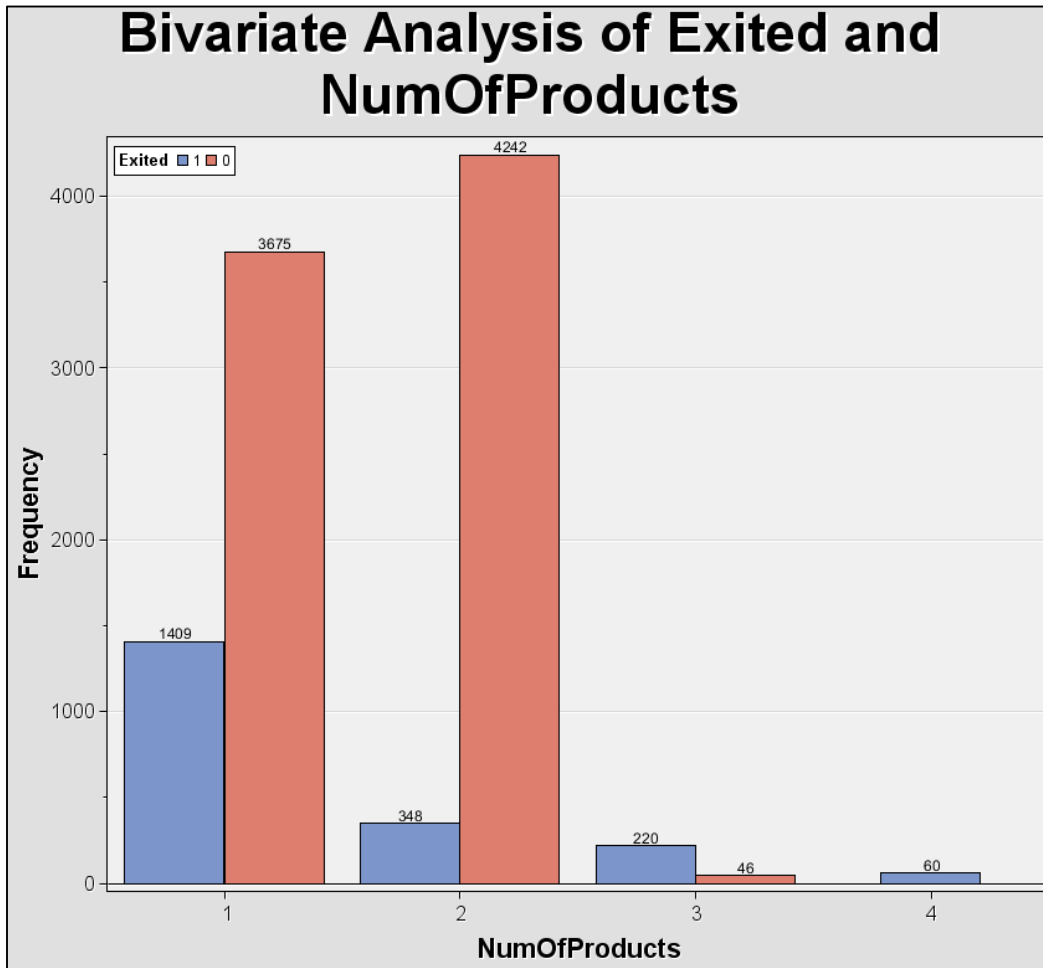


Figure 3.23: Bivariate analysis of the variables Exited and NumOfProducts

In overall, there are 5084 customers who purchased one product from the bank, 4590 customers who purchased two products from the bank, 266 customers who purchased three products from the bank, and 60 customers who purchased four products from the bank. Looking at the number of customers churned, it can be identified that category who purchased one product has 1409 customer churned (27.71%), category who purchased two products has 348 customers churned (7.58%), category who purchased three products has 220 customers churned (82.71%), and category who purchased four products have 60 customers churned (100%). It is observed that majority customers are purchasing one or two products from the bank, and very few customers would consider purchasing for the third or fourth time from the bank. In addition, customers who purchased for the third or fourth time from the bank exhibit very high churn rates. Probable reason of the high churn rate and low customer count observed in the third and fourth time purchasing from the bank is due to the likelihood that the products offered may not be accommodating the needs of the customers thus leading to low demand.

3.3.3.5 Bivariate Analysis for Variables – Exited and Age

Table 3.7 shows the table summary statistics and Figure 3.24 shows the graphical representation for the variable combination of Exited and Age.

Table 3.7: Summary statistics of variable Exited and Age

Exited	Variable: Age					
	Frequency	Min	Max	Mean	Median	Std. Dev.
0	7963	18	92	37.41	36	10.13
1	2037	18	84	44.84	45	9.76

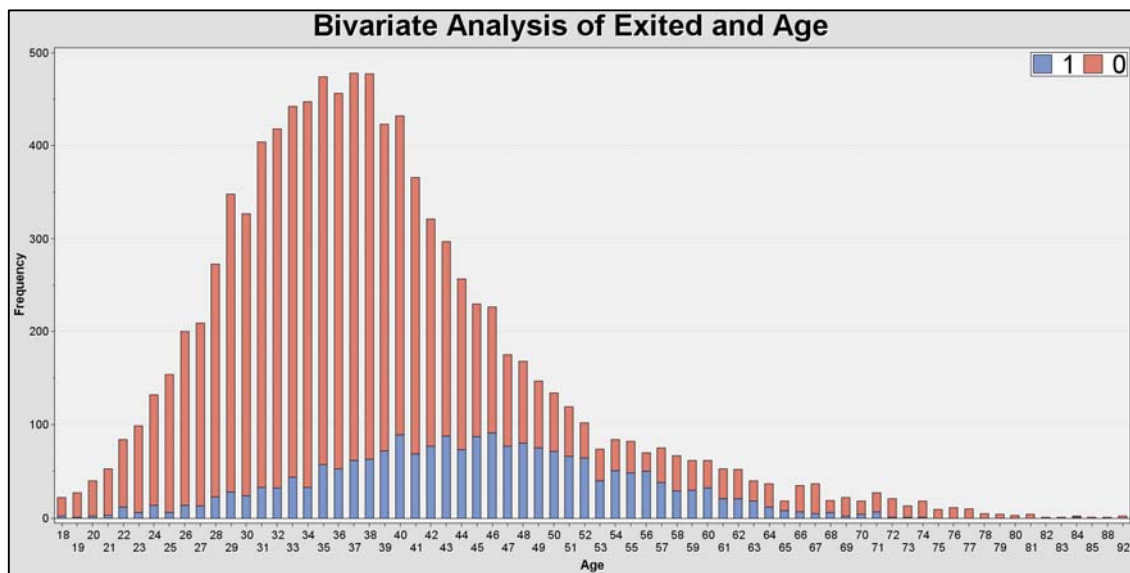


Figure 3.24: Bivariate analysis of the variables Exited and Age

In overall, there are 7963 customers retained while 2037 customers churned. Looking at the customers retained, the age of the customers in this category ranges between 18 to 92. The mean age identified is 37.41 which is slightly higher than the median age of 36. This indicates a skew in the data distribution which the figure shows a positive skew is identified. This results in more customers having ages in the lower range of the distribution. While for the customers churned, the age of the customers in this category ranges between 18 to 84. The mean age identified is 44.84 which is slightly lower than the median age of 45. This indicates a skew in the data distribution which the figure shows a negative skew is identified. This results in more customers having ages in the higher range of the distribution. Based on the findings, it can be identified that customers of higher age, are more likely to churn than customers of the lower age. Probable reason of the higher churn rate observed in the higher aged customers is due to

the likelihood that the bank is not offering services or products meeting the needs of customers of the higher age groups. Older customers tend to have different priorities as compared to younger customers in terms of banking services and products offered. The older customers tend to favor banks that offer services regarding pensions, inheritance, and taxes. Therefore, banks which are unable to provide such services would result in the higher age group customers switching to other banks that offer better services at competitive rates.

SECTION 4

IMPLEMENTATION

4.1 INTRODUCTION

This section outlines the development process of the predictive models which involves performing data pre-processing, model optimization, model validation, and comparison between the validated predictive models. In addition, the overall process flow diagram is shown in Figure 4.1.

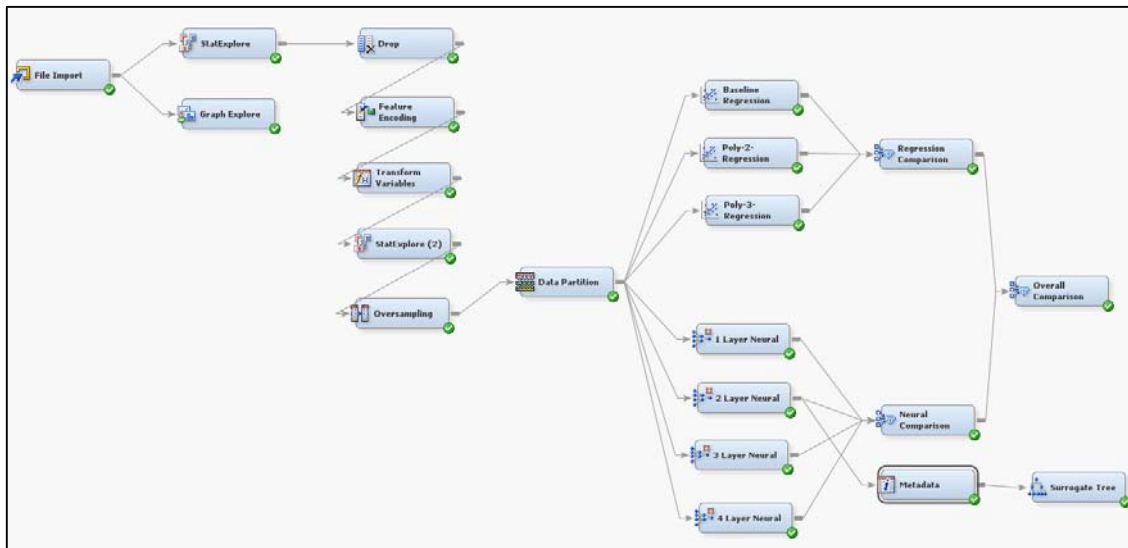


Figure 4.1: Overall process flow diagram

4.2 DATA PRE-PROCESSING

This section outlines the data pre-processing procedure performed on the dataset. The procedure would include feature selection, feature transformation, and data partitioning. Data pre-processing performed properly would facilitate the prediction outcomes of the prediction models.

4.2.1 Feature Selection

Feature selection involves the reduction of input variables to be used in the prediction models. The reduction in the number of input variables would improve the overall prediction accuracy and tendency of overfitting of the prediction models by eliminating redundant variables that may be a source of noise for the model. In addition, lesser number of variables indicate lesser

amount of data to be processed which translate to reduction in model training time. To identify the relevant variables to be retained, the variable worth from StatExplore tool will be utilized.

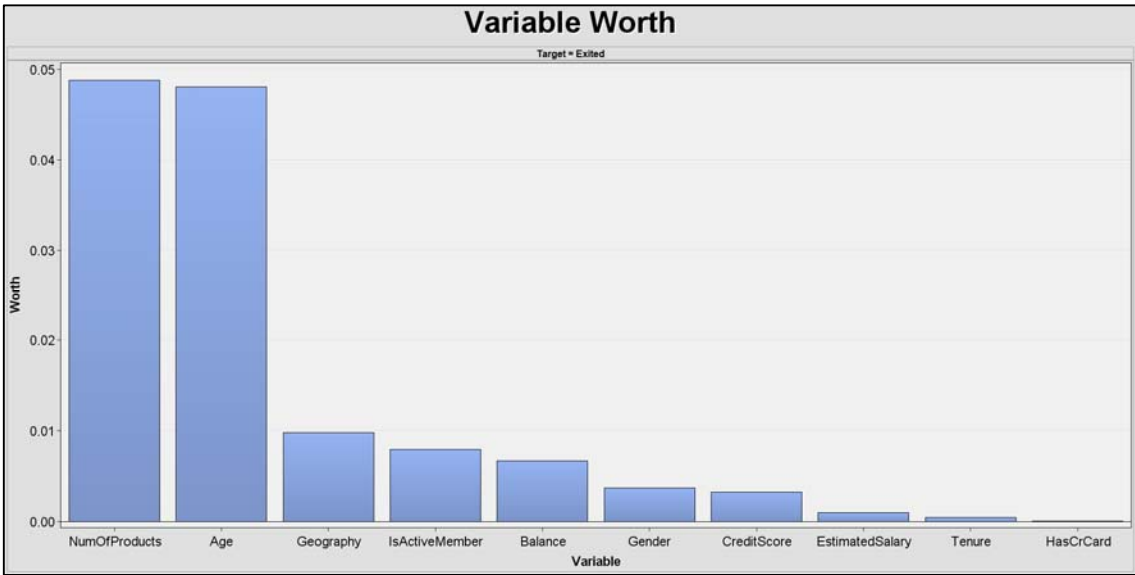


Figure 4.2: Variable worth

Table 4.1: Variable worth

Target	Variable	Importance	Worth
Exited	NumOfProducts	1	0.04878
Exited	Age	2	0.048023
Exited	Geography	3	0.009773
Exited	IsActiveMember	4	0.007908
Exited	Balance	5	0.00666
Exited	Gender	6	0.00368
Exited	CreditScore	7	0.003238
Exited	EstimatedSalary	8	.0009926
Exited	Tenure	9	0.000399
Exited	HasCrCard	10	1.653E-5

Figure 4.2 shows the variable worth in the graphical format while table 4.1 shows the variable worth in the table format. Based on the outputs, it can be identified that, three out of the ten variables have significantly low variable worth. This indicates that the variables are of low importance and exhibit low predictive power. The identified variables with low variable worth are “EstimatedSalary”, “Tenure”, and “HasCrCard”. These variables will be dropped from the dataset, while the remaining seven variables will be used for developing the prediction models.

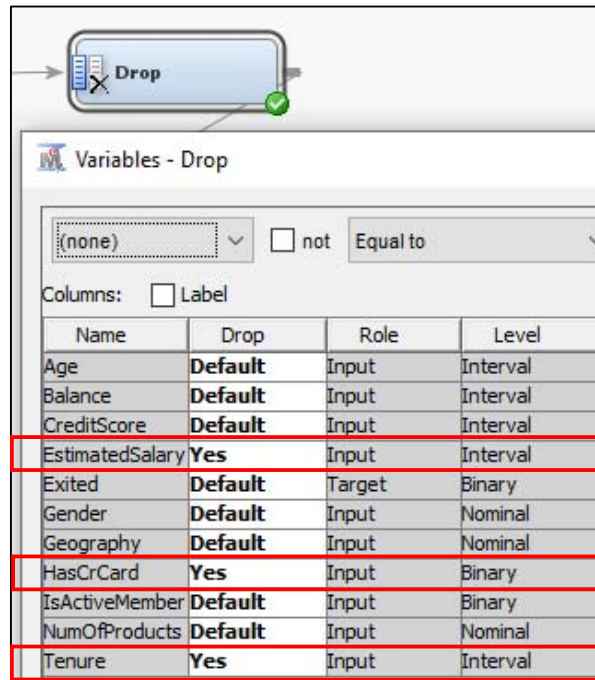



Figure 4.3: Variables dropping node settings

Figure 4.3 shows the settings in the Drop tool utilized in SAS Enterprise Miner. It is used to remove the three irrelevant variables from proceeding to the model development stage.

4.2.2 Feature Encoding

Feature encoding refers to the transformation of labels in categorical variables into a numerical representation. This process ensures the predictive models are able to work with the data, whereby not all predictive algorithms are able to work directly with character labels. The Replacement tool from SAS Enterprise Miner will be utilized to perform the feature encoding task.

It is identified that three categorical variables namely “Gender”, “Geography”, and “NumOfProducts”, require the labels to be replaced with a numerical value. A numerical representation starting from zero with increment of one will be provided for each label in each specified variable. Figure 4.4 shows the settings performed in the class variables replacement editor for the three variables mentioned.



Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
Exited	0		7963N			0
Exited	1		2037N			1
Exited	_UNKNOWN_	_DEFAULT_		N		.
Gender	Male	0	5457C		Male	.
Gender	Female	1	4543C		Female	.
Gender	_UNKNOWN_	_DEFAULT_		C		.
Geography	France	0	5014C		France	.
Geography	Germany	1	2509C		Germany	.
Geography	Spain	2	2477C		Spain	.
Geography	_UNKNOWN_	_DEFAULT_		C		.
IsActiveMember	1		5151N			1
IsActiveMember	0		4849N			0
IsActiveMember	_UNKNOWN_	_DEFAULT_		N		.
NumOfProducts	1	0	5084N			1
NumOfProducts	2	1	4590N			2
NumOfProducts	3	2	266N			3
NumOfProducts	4	3	60N			4
NumOfProducts	_UNKNOWN_	_DEFAULT_		N		.

Figure 4.4: Feature encoding node settings

4.2.3 Feature Transformation

Feature transformation refers to the application of mathematical functions to the features which transforms the data values. The application of feature transformation enhances the performance of prediction models, especially in the LR model which built on the assumption that data is normally distributed. Applying feature transformation would transform the data into a normal distribution which satisfy the assumption of LR, resulting in a better prediction model developed. In addition, feature transformation facilitates prediction algorithms to converge faster especially in the ANN model. The Transform Variables tool from SAS Enterprise Miner will be utilized to perform the feature transformation.

The log transformation will be utilized in this study to approximately conform the data into a normal distribution. Figure 4.5 shows the properties of the Transform Variables tool where the interval inputs will be applied the log transformation. Figure 4.6 shows the summary statistics after log transformation of the interval variables. It is identified that the variables now have a smaller and uniform range of values.

Property	Value
General	
Node ID	Trans
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Formulas	...
Interactions	...
SAS Code	...
Default Methods	
Interval Inputs	Log
Interval Targets	None
Class Inputs	None
Class Targets	None
Treat Missing as Level	No
Sample Properties	
Method	First N
Size	Default
Random Seed	12345
Optimal Binning	
Number of Bins	4
Missing Values	Use in Search
Grouping Method	
Cutoff Value	0.1

Figure 4.5: Feature transformation node settings

Interval Variable Summary Statistics (maximum 500 observations printed)										
Data Role=TRAIN										
Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
LOG_REP_Age	INPUT	3.653918	0.24957	10000	0	2.944439	3.637586	4.268091	0.150632	0.021859
LOG_REP_Balance	INPUT	7.441327	5.6064	10000	0	0	11.48442	12.43281	-0.57083	-1.66796
LOG_REP_CreditScore	INPUT	6.467919	0.152854	10000	0	5.890453	6.481577	6.746412	-0.44902	-0.04164

Figure 4.6: Interval variables after log transformation

4.2.4 Data Balancing

Data balancing refers to the sampling of data to achieve an even number of observations in each target class. It is typically performed for target variable that experiences an extremely uneven distribution of class labels. The use of imbalanced dataset to develop prediction models would yield a biased and inaccurate results. This problem can be solved by performing oversampling for the dataset. The Sample tool from SAS Enterprise Miner can be utilized to perform data balancing.

Figure 4.7 shows the properties of the Sample node set to perform oversampling. Figure 4.8 shows the results after oversampling. Where, the dataset now has equal number of observations in both target label “1” and “0” which is 2037 observations each.

.. Property	Value
General	
Node ID	Smpl
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Sample Method	Default
Random Seed	12345
<input checked="" type="checkbox"/> Size	
Type	Percentage
Observations	.
Percentage	100.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
<input checked="" type="checkbox"/> Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
<input checked="" type="checkbox"/> Level Based Options	
Level Selection	Rarest Level
Level Proportion	10.0
Sample Proportion	5.0

Figure 4.7: Sample node settings

Data=SAMPLE					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Exited	0	0	2037	50	
Exited	1	1	2037	50	

Figure 4.8: Data sample after sampling

4.2.5 Data Partitioning

Data partitioning refers to the division of the dataset into different subsets for different purposes. The dataset in this study will be partitioned into two subsets namely the training set and the validation set. The training set will be used for model fitting while the validation set will be used to validate the performance of the developed models. The partition ratio will follow the

70% for training set and 30% for validation set. The Data Partition tool from SAS Enterprise Miner will be utilized to perform the data partitioning. This resulted in 2851 observations for the training set and 1223 observations for the testing set. Figure 4.9 shows the properties of the Data Partition tool where the data split ratio is specified.

.. Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
[-] Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes

Figure 4.9: Data partition node settings

4.3 MODEL DEVELOPMENT

Two different types of prediction models will be developed in this study, namely the regression model and neural network model. The prediction task involves a classification to predict customer churned or retained. The churned is labeled by “1” which is the positive class, while the retained is labeled by “0” which is the negative class. The prediction models will undergo hyperparameter tuning to achieve optimal performance. The best model will be selected as the final model to predict customer churn.

4.3.1 Logistic Regression Model

The first model to be developed is the logistic regression model. After the feature selection phase, the remaining seven variables based on variable worth will be used by the regression to develop the prediction model. The regression models will adopt the Stepwise method for model selection. Several logistic regression models will be developed by increasing the polynomial degree. Figure 4.10 shows the partial process flow diagram of the developed logistic regression models.

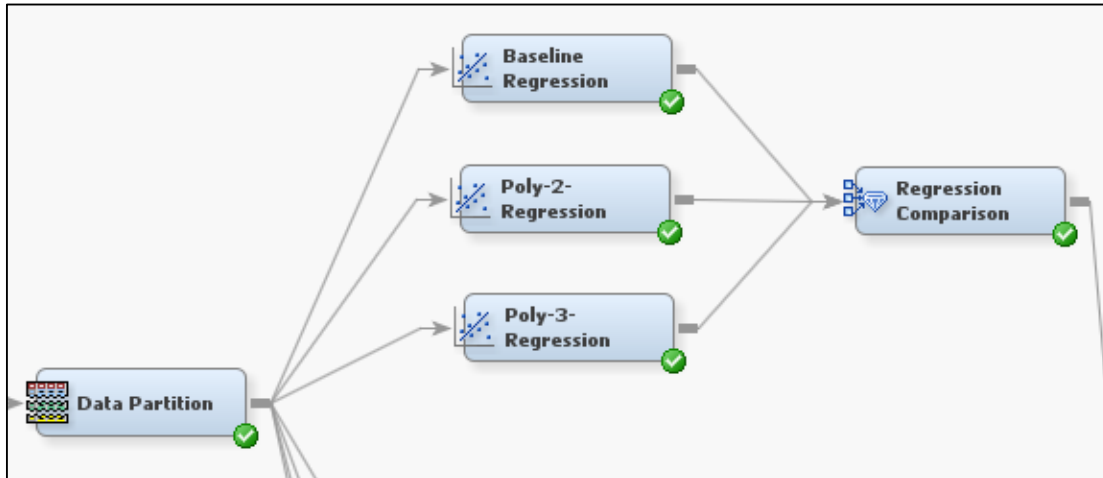


Figure 4.10: Partial process flow diagram for logistic regression model

4.3.1.1 Baseline Regression Model

The baseline regression model is developed using the Regression node from SAS Enterprise Miner. All properties of the regression node remain at default. Figure 4.11 shows the classification results for the model while Table 4.2 shows the evaluation metrics computed based on the classification results. Further results generated from SAS Enterprise Miner are attached under Appendix A.1.

Event Classification Table			
Data Role=TRAIN Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
340	1107	319	1085
Data Role=VALIDATE Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
165	463	148	447

Figure 4.11: Classification result regression baseline model

Table 4.2: Evaluation metrics for regression baseline model

	Training Set	Validation Set
Misclassification Rate	0.231147	0.255928
Precision	0.772792	0.751261
Recall	0.761404	0.730392

Based on the evaluation metrics, the model achieved a reasonable misclassification rate of 25.59% which indicates that of 100 predicted outcomes, only 25.59 outcomes are predicted incorrectly. While for the precision, the model achieved a rate of 75.13%. This indicates that the model has a reasonable accuracy for the predicted positive class. Which indicates that of 100 predicted customers that will churn, 75.13 of the predicted customers did churn. For the recall, the model achieved a rate of 73.04%. This indicates that the model has a reasonable accuracy for the actual positive class. Which indicates that of 100 actual customers that churned, the model predicted that 73.04 of the customers did churn.

4.3.1.2 Polynomial Degree 2 Regression Model

The polynomial degree 2 regression model is developed using the Regression node from SAS Enterprise Miner. The node is renamed to Poly-2-Regression for identification. Under the properties of the regression node, the Polynomial Terms is set to “Yes” and the Polynomial Degree is set to “2”. Figure 4.12 shows the classification results for the model while Table 4.3 shows the evaluation metrics computed based on the classification results. Further results generated from SAS Enterprise Miner are attached under Appendix A.2.

Event Classification Table			
Data Role=TRAIN Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
340	1110	316	1085
Data Role=VALIDATE Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
165	463	148	447

Figure 4.12: Classification result regression polynomial degree 2 model

Table 4.3: Evaluation metrics for regression polynomial degree 2 model

	Training Set	Validation Set
Misclassification Rate	0.230095	0.255928
Precision	0.774447	0.751261
Recall	0.761404	0.730392

Based on the evaluation metrics, the model achieved a reasonable misclassification rate of 25.59% which indicates that of 100 predicted outcomes, only 25.59 outcomes are predicted incorrectly. While for the precision, the model achieved a rate of 75.13%. This indicates that the model has a reasonable accuracy for the predicted positive class. Which indicates that of 100 predicted customers that will churn, 75.13 of the predicted customers did churn. For the recall, the model achieved a rate of 73.04%. This indicates that the model has a reasonable accuracy for the actual positive class. Which indicates that of 100 actual customers that churned, the model predicted that 73.04 of the customers did churn.

4.3.1.3 Polynomial Degree 3 Regression Model

The polynomial degree 3 regression model is developed using the Regression node from SAS Enterprise Miner. The node is renamed to Poly-3-Regression for identification. Under the properties of the regression node, the Polynomial Terms is set to “Yes” and the Polynomial Degree is set to “3”. Figure 4.13 shows the classification results for the model while Table 4.4 shows the evaluation metrics computed based on the classification results. Further results generated from SAS Enterprise Miner are attached under Appendix A.3.

Event Classification Table			
Data Role=TRAIN Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
333	1106	320	1092
Data Role=VALIDATE Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
168	465	146	444

Figure 4.13: Classification result regression polynomial degree 3 model

Table 4.4: Evaluation metrics for regression polynomial degree 3 model

	Training Set	Validation Set
Misclassification Rate	0.229042	0.256746
Precision	0.773371	0.752542
Recall	0.766316	0.725490

Based on the evaluation metrics, the model achieved a reasonable misclassification rate of 25.67% which indicates that of 100 predicted outcomes, only 25.67 outcomes are predicted incorrectly. While for the precision, the model achieved a rate of 75.25%. This indicates that the model has a reasonable accuracy for the predicted positive class. Which indicates that of 100 predicted customers that will churn, 75.25 of the predicted customers did churn. For the recall, the model achieved a rate of 72.55%. This indicates that the model has a reasonable accuracy for the actual positive class. Which indicates that of 100 actual customers that churned, the model predicted that 72.55 of the customers did churn.

4.3.1.4 Comparison of Regression Models

Figure 4.13 shows the receiver operating characteristic (ROC) curve of the three regression models. While the following table, Table 4.5 shows the ROC index for each regression model.

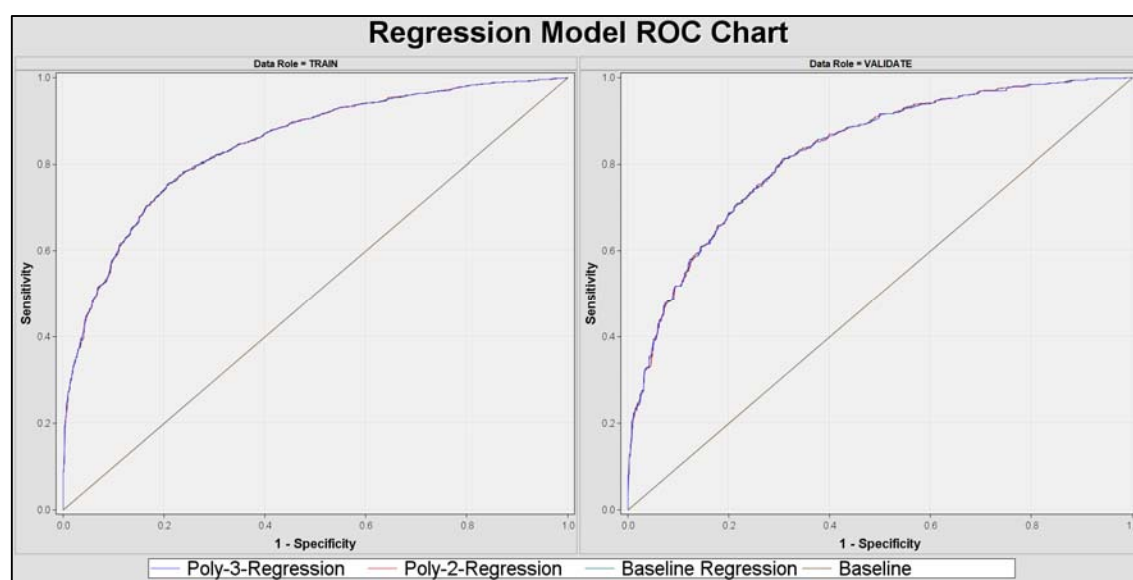


Figure 4.14: ROC chart for regression models

Table 4.5: ROC index for regression models

Models	ROC Index	
	Training Set	Validation Set
Baseline Regression Model	0.844	0.827
Polynomial Degree 2 Regression Model	0.844	0.827
Polynomial Degree 3 Regression Model	0.844	0.826

Table 4.6: Validation evaluation metrics for regression models

Model	Misclassification	Precision	Recall
Baseline Regression Model	0.255928	0.751261	0.730392
Polynomial Degree 2 Regression Model	0.255928	0.751261	0.730392
Polynomial Degree 3 Regression Model	0.256746	0.752542	0.725490

Based on Table 4.5, it is identified that both baseline and polynomial degree 2 regression model achieved the best ROC index of 0.827 which indicates the models have high performance in distinguishing between positive and negative classes. While based on Table 4.6, the baseline model and the polynomial degree 2 model achieved the same and best performance in terms of all evaluation metrics. However, the baseline model will be chosen due to the lower model complexity as compared to the polynomial degree 2 model.

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-5.0291	9.7228	0.27	0.6050		0.007
IsActiveMember	0	0.5487	0.0476	132.86	<.0001		1.731
LOG_REP_Age	1	3.5572	0.2082	291.92	<.0001	0.4915	35.065
LOG_REP_CreditScore	1	-0.7546	0.3006	6.30	0.0121	-0.0652	0.470
REP_Gender	0	-0.2758	0.0468	34.80	<.0001		0.759
REP_Geography	0	-0.3838	0.0626	37.57	<.0001		0.681
REP_Geography	1	0.6410	0.0705	82.76	<.0001		1.898
REP_NumOfProducts	0	-2.9125	9.5004	0.09	0.7592		0.054
REP_NumOfProducts	1	-4.3706	9.5005	0.21	0.6455		0.013
REP_NumOfProducts	2	-0.5025	9.5035	0.00	0.9578		0.605

Figure 4.15: Maximum likelihood estimates for baseline regression model

Figure 4.15 shows the maximum likelihood estimates for the baseline regression model. Based on the outputs, it can be identified that six parameters exhibit significant contribution to the prediction of customer churn. The following interpret each parameter with respect to customer churn:

- When feature IsActiveMember class 0 increases by 1 unit, the odds of customer churn increase by 0.731 times.
- When feature LOG_REP_Age increases by 1 unit, the odds of customer churn increase by 34.065 times.
- When feature LOG_REP_CreditScore increases by 1 unit, the odds of customer churn decrease by 0.530 times.
- When feature REP_Gender class 0 increases by 1 unit, the odds of customer churn decrease by 0.241.
- When feature REP_Geography class 0 increases by 1 unit, the odds of customer churn decrease by 0.319.
- When feature REP_Geography class 1 increases by 1 unit, the odds of customer churn increase by 0.898.

4.3.2 Neural Network Model

The second model to be developed is the neural network model. The neural network model will undergo optimization to identify the best performing model by trial and error using various combination of the number of hidden layers and hidden neurons hyperparameters. Figure 4.16 shows the partial process flow diagram of the developed neural network models.

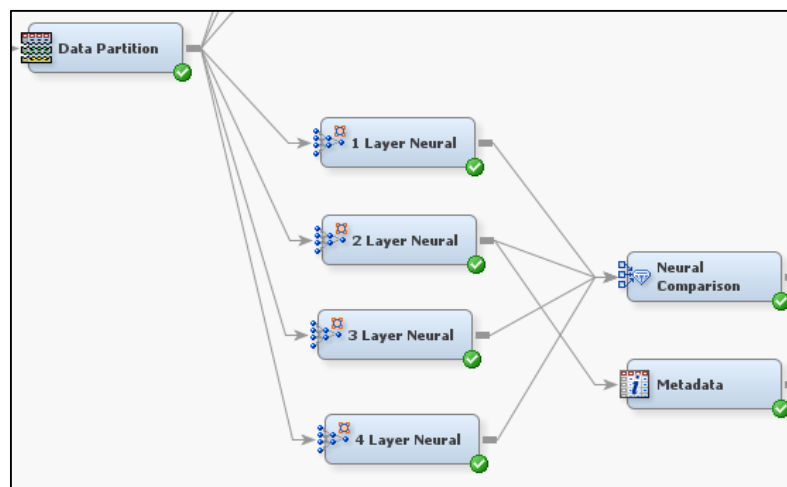


Figure 4.16: Partial process flow diagram for neural network model

4.3.2.1 1 Hidden Layer Neural Network

The 1 hidden layer neural network model is developed using the HP Neural node from SAS Enterprise Miner. The node is renamed to 1 Layer Neural for identification. Under the properties of the node, the maximum iterations are set to 1000 and the architecture is set to “One Layer”. While for the number of hidden neurons, several values have been utilized to determine the optimal performing model. Table 4.7 shows the misclassification rate for the different number of hidden neurons used.

Table 4.7: Misclassification of 1 hidden layer neural network models

Hidden Layer	Hidden Neuron	Misclassification Rate	
		Train	Validate
1 Layer	12	0.203087	0.248569
1 Layer	10	0.210452	0.244481
1 Layer	8	0.208348	0.249387
1 Layer	6	0.210102	0.255928
1 Layer	4	0.203087	0.247751
1 Layer	2	0.226587	0.246934

Based on Table 4.7, it is identified that 1 hidden layer neural network model performed best with 10 hidden neurons which yielded the lowest misclassification rate for the validation set. Figure 4.17 shows the classification results for the optimal model while Table 4.8 shows the evaluation metrics computed based on the classification results. Further results generated from SAS Enterprise Miner are attached under Appendix A.4.

Event Classification Table			
Data Role=TRAIN Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
347	1173	253	1078
Data Role=VALIDATE Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
174	486	125	438

Figure 4.17: Classification result of optimal 1 hidden layer neural network model

Table 4.8: Evaluation metrics for optimal 1 hidden layer neural network model

	Training Set	Validation Set
Misclassification Rate	0.210452	0.244481
Precision	0.809917	0.777975
Recall	0.756491	0.715686

Based on the evaluation metrics, the model achieved a reasonable misclassification rate of 24.45% which indicates that of 100 predicted outcomes, only 24.45 outcomes are predicted incorrectly. While for the precision, the model achieved a rate of 77.80%. This indicates that the model has a reasonable accuracy for the predicted positive class. Which indicates that of 100 predicted customers that will churn, 77.80 of the predicted customers did churn. For the recall, the model achieved a rate of 71.57%. This indicates that the model has a reasonable accuracy for the actual positive class. Which indicates that of 100 actual customers that churned, the model predicted that 71.57 of the customers did churn.

4.3.2.2 2 Hidden Layer Neural Network

The 2 hidden layer neural network model is developed using the HP Neural node from SAS Enterprise Miner. The node is renamed to 2 Layer Neural for identification. Under the properties of the node, the maximum iterations are set to 1000 and the architecture is set to “User-Defined”. While for the number of hidden neurons, several values have been utilized by changing the settings in the hidden layer options to determine the optimal performing model. Table 4.9 shows the misclassification rate for the different combination of number of hidden neurons used.

Table 4.9: Misclassification of 2 hidden layer neural network models

Hidden Layer	Hidden Neuron	Misclassification Rate	
		Train	Validate
2 Layer	12, 10	0.202034	0.256746
2 Layer	10, 8	0.217468	0.246934
2 Layer	8, 6	0.206594	0.251840
2 Layer	6, 4	0.207296	0.246116
2 Layer	4, 2	0.210452	0.249387
2 Layer	12, 12	0.207646	0.244481
2 Layer	10, 10	0.212908	0.252657
2 Layer	8, 8	0.216065	0.242845
2 Layer	6, 6	0.213609	0.239575
2 Layer	4, 4	0.212206	0.251022

2 Layer	2, 2	0.222027	0.251840
---------	------	----------	----------

Based on Table 4.9, it is identified that 2 hidden layer neural network model performed best with the 6, 6 hidden neurons configuration which yielded the lowest misclassification rate for the validation set. Figure 4.18 shows the classification results for the optimal model while Table 4.10 shows the evaluation metrics computed based on the classification results. Further results generated from SAS Enterprise Miner are attached under Appendix A.5.

Event Classification Table			
Data Role=TRAIN Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
328	1145	281	1097
Data Role=VALIDATE Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
155	473	138	457

Figure 4.18: Classification result of optimal 2 hidden layer neural network model

Table 4.10: Evaluation metrics for optimal 2 hidden layer neural network model

	Training Set	Validation Set
Misclassification Rate	0.213609	0.239575
Precision	0.796081	0.768067
Recall	0.769824	0.746732

Based on the evaluation metrics, the model achieved a reasonable misclassification rate of 23.96% which indicates that of 100 predicted outcomes, only 23.96 outcomes are predicted incorrectly. While for the precision, the model achieved a rate of 76.81%. This indicates that the model has a reasonable accuracy for the predicted positive class. Which indicates that of 100 predicted customers that will churn, 76.81 of the predicted customers did churn. For the recall, the model achieved a rate of 74.67%. This indicates that the model has a reasonable accuracy for the actual positive class. Which indicates that of 100 actual customers that churned, the model predicted that 74.67 of the customers did churn.

4.3.2.3 3 Hidden Layer Neural Network

The 3 hidden layer neural network model is developed using the HP Neural node from SAS Enterprise Miner. The node is renamed to 3 Layer Neural for identification. Under the properties of the node, the maximum iterations are set to 1000 and the architecture is set to “User-Defined”. While for the number of hidden neurons, several values have been utilized by changing the settings in the hidden layer options to determine the optimal performing model. Table 4.11 shows the misclassification rate for the different combination of number of hidden neurons used.

Table 4.11: Misclassification of 3 hidden layer neural network models

Hidden Layer	Hidden Neuron	Misclassification Rate	
		Train	Validate
3 Layer	12, 10, 8	0.203437	0.247751
3 Layer	10, 8, 6	0.216766	0.250204
3 Layer	8, 6, 4	0.212206	0.264922
3 Layer	6, 4, 2	0.202385	0.252657
3 Layer	12, 12, 12	0.199930	0.241210
3 Layer	10, 10, 10	0.205893	0.241210
3 Layer	8, 8, 8	0.204840	0.237939
3 Layer	6, 6, 6	0.211505	0.246116
3 Layer	4, 4, 4	0.210803	0.259199
3 Layer	2, 2, 2	0.223080	0.245298

Based on Table 4.11, it is identified that 3 hidden layer neural network model performed best with the 8, 8, 8 hidden neurons configuration which yielded the lowest misclassification rate for the validation set. Figure 4.19 shows the classification results for the optimal model while Table 4.12 shows the evaluation metrics computed based on the classification results. Further results generated from SAS Enterprise Miner are attached under Appendix A.6.

Event Classification Table			
Data Role=TRAIN Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
322	1164	262	1103
Data Role=VALIDATE Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
166	486	125	446

Figure 4.19: Classification result of optimal 3 hidden layer neural network model

Table 4.12: Evaluation metrics for optimal 3 hidden layer neural network model

	Training Set	Validation Set
Misclassification Rate	0.204840	0.237939
Precision	0.808059	0.781086
Recall	0.774035	0.728758

Based on the evaluation metrics, the model achieved a reasonable misclassification rate of 23.79% which indicates that of 100 predicted outcomes, only 23.79 outcomes are predicted incorrectly. While for the precision, the model achieved a rate of 78.11%. This indicates that the model has a reasonable accuracy for the predicted positive class. Which indicates that of 100 predicted customers that will churn, 78.11 of the predicted customers did churn. For the recall, the model achieved a rate of 72.88%. This indicates that the model has a reasonable accuracy for the actual positive class. Which indicates that of 100 actual customers that churned, the model predicted that 72.88 of the customers did churn.

4.3.2.4 4 Hidden Layer Neural Network

The 4 hidden layer neural network model is developed using the HP Neural node from SAS Enterprise Miner. The node is renamed to 4 Layer Neural for identification. Under the properties of the node, the maximum iterations are set to 1000 and the architecture is set to “User-Defined”. While for the number of hidden neurons, several values have been utilized by changing the settings in the hidden layer options to determine the optimal performing model.

Table 4.13 shows the misclassification rate for the different combination of number of hidden neurons used.

Table 4.13: Misclassification of 4 hidden layer neural network models

Hidden Layer	Hidden Neuron	Misclassification Rate	
		Train	Validate
4 Layer	12, 10, 8, 6	0.203788	0.242028
4 Layer	10, 8, 6, 4	0.200982	0.240398
4 Layer	8, 6, 4, 2	0.212557	0.245298
4 Layer	12, 12, 12, 12	0.203788	0.250204
4 Layer	10, 10, 10, 10	0.204139	0.246934
4 Layer	8, 8, 8, 8	0.210102	0.243663
4 Layer	6, 6, 6, 6	0.215363	0.249387
4 Layer	4, 4, 4, 4	0.215363	0.242028
4 Layer	2, 2, 2, 2	0.218520	0.248569

Based on Table 4.13, it is identified that 4 hidden layer neural network model performed best with the 10, 8, 6, 4 hidden neurons configuration which yielded the lowest misclassification rate for the validation set. Figure 4.20 shows the classification results for the optimal model while Table 4.14 shows the evaluation metrics computed based on the classification results. Further results generated from SAS Enterprise Miner are attached under Appendix A.7.

Event Classification Table			
Data Role=TRAIN Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
314	1167	259	1111
Data Role=VALIDATE Target=Exited Target Label=' '			
False Negative	True Negative	False Positive	True Positive
163	480	131	449

Figure 4.20: Classification result of optimal 4 hidden layer neural network model

Table 4.14: Evaluation metrics for optimal 4 hidden layer neural network model

	Training Set	Validation Set
Misclassification Rate	0.200982	0.240392
Precision	0.810949	0.774138
Recall	0.779649	0.733660

Based on the evaluation metrics, the model achieved a reasonable misclassification rate of 24.04% which indicates that of 100 predicted outcomes, only 24.04 outcomes are predicted incorrectly. While for the precision, the model achieved a rate of 77.41%. This indicates that the model has a reasonable accuracy for the predicted positive class. Which indicates that of 100 predicted customers that will churn, 77.41 of the predicted customers did churn. For the recall, the model achieved a rate of 73.37%. This indicates that the model has a reasonable accuracy for the actual positive class. Which indicates that of 100 actual customers that churned, the model predicted that 73.37 of the customers did churn.

4.3.2.5 Comparison of Neural Network Models

Figure 4.21 shows the ROC curve of the four neural network models. While the following table, Table 4.15 shows the ROC index for each neural network model.

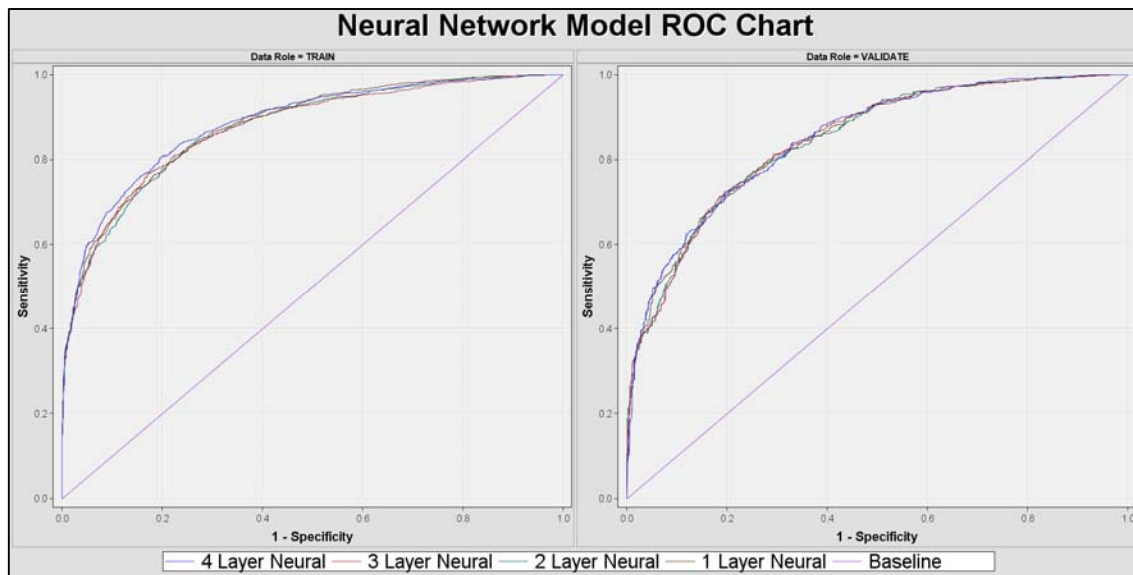


Figure 4.21: ROC chart for neural network models

Table 4.15: ROC index for optimal neural network models

Models	ROC Index	
	Training Set	Validation Set
1 Hidden Layer Neural Network	0.879	0.846
2 Hidden Layer Neural Network	0.873	0.844
3 Hidden Layer Neural Network	0.872	0.844
4 Hidden Layer Neural Network	0.884	0.848

Table 4.16: Validation evaluation metrics for optimal neural network models

Model	Misclassification	Precision	Recall
1 Hidden Layer Neural Network	0.244481	0.777975	0.715686
2 Hidden Layer Neural Network	0.239575	0.768067	0.746732
3 Hidden Layer Neural Network	0.237939	0.781086	0.728758
4 Hidden Layer Neural Network	0.240392	0.774138	0.733660

Based on Table 4.15, it is identified that the 4 hidden layer neural network model achieved the best ROC index of 0.848 which indicates the model has higher performance in distinguishing between positive and negative classes. However, the ROC index does not differ much among the other models thus would require other metrics for evaluation. While based on Table 4.16, the 3 hidden layer neural network model achieved the best performance in terms of misclassification and precision. While the 2 hidden layer neural network model achieved the best performance in terms of recall.

However, based on the scenario of this study which is to improve customer retention rate. A higher recall rate is more favorable, as the false negatives should be minimized. This is to ensure customers that are actually going to churn is accurately identified by the model so that retention campaigns are targeted towards more of the right audiences. Therefore, the 2 hidden layer neural network model should be chosen based on higher recall rate which is more suitable for the use case.

A surrogate model with underlying decision tree is used to understand the prediction output by the 2 hidden layer neural network model. The surrogate model has undergone tree pruning to reduce overfitting, with the final number of leaves to remain set to 17. Figure 4.22 shows the misclassification rate graph of the surrogate model after applying pruning.

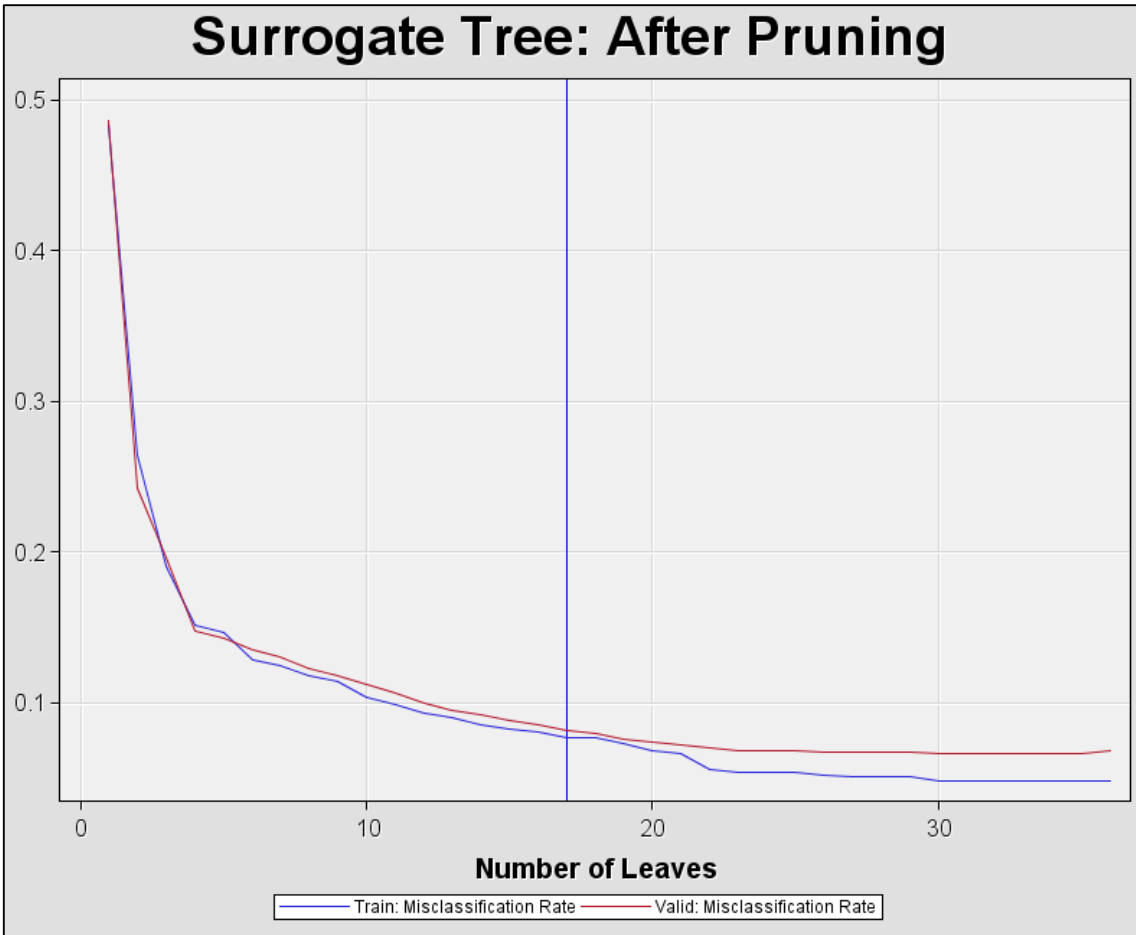


Figure 4.22: Misclassification rate after pruning for surrogate model

Based on the node rules generated by the surrogate model, the most relevant and significant node identified is the node number 15 which represents the highest number of observations used and has a prediction accuracy of 98% for predicting customers that will churn. Figure 4.23 shows the node rule number 15 generated from the surrogate model. The complete list of node rules can be found in Appendix A.8.

```

*-----*
Node = 15
*-----*
if Transformed: Replacement: Age >= 3.67622 or MISSING
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND IsActiveMember IS ONE OF: 0 or MISSING
then
Tree Node Identifier    = 15
Number of Observations = 680
Predicted: U_Exited=1 = 0.98
Predicted: U_Exited=0 = 0.03

```

Figure 4.23: Node rule number 15 from surrogate model

The age feature shown in the node rule has undergone log transformation, the formula to revert the age feature back to original value is attached under Appendix A.8. Which after conversion, the age represented in the node rule is 38.5.

Based on the node rule, three features are identified to be significant. It is identified that if the customer age is greater than or equal 38.5 and have purchased 1, 3, or 4 products from the bank and is an inactive member. Such customers have a 98% chance of churning.

4.4 DISCUSSION

One optimal model was identified from each type of prediction algorithms used. However, upon comparison of their performances, the 2 hidden layer neural network model is chosen as the final model. As this model provides a reasonable misclassification and precision rate while producing the highest recall rate among all other models. The higher recall rate is favorable in this use case, as minimizing the false negatives would allow the model to more precisely identify customers that are likely to churn. This allows targeted measures to be provided to the right audiences to minimize the likelihood of customer churn.

Significant features identified from the chosen model include customers with age over 38, customers who have purchased 1 or 3 or 4 products from the bank, and customers with an inactive member status. Based on the previous descriptive analytics performed, these variables are indeed showing high customer churn rate which enforces the findings. In addition to the identified features, the customer churn prediction model allows the early detection of customers that are likely to churn. Which the bank can intervene at the early stages to reduce the likelihood of customer churning.

Based on the findings, the identified customer group is of older age who have purchased from the bank and remained inactive. The following measures are suggested to increase the retention rate of the identified customer group which suffers from high churn rate:

1. First suggestion involves knowing your customers. Surveys and questionnaires should be conducted on the targeted audiences to obtain feedbacks and reviews of the products and services offered. This allows the bank to understand the expectations and demands of the customers which would lead to better development of personalized products and services for the customers. Personalized offerings can significantly improve the customer retention rate where based on a survey conducted by Accenture, 91% of the customers are likely to return and conduct the purchase again if an offering provided is meaningful and relevant to them (Zoghby et al., 2018). As for the identified customer group which consist of the older age customers, personalized offerings such as life insurance policy, wills, and estate planning will be more favorable to attract and retain the specified customers.
2. Second suggestion involves increasing customer satisfaction. It is evident that happier and satisfied customers are less likely to churn. Low customer satisfaction typically arises from poor services offered by the bank which can significantly increases the customer churn rate. A study shows that poor services offered by the bank can lead to 40% of the customers to switch to the competitors who is offering better services (Kamalaratnam, 2022). As for the identified customer group which consist of older age customers, they are typically more appreciative of face-to-face engagement and are likely less tech-savvy. The rapid digital transformation age where all services are moving online, causes the resources to be focused on the digital platform and neglected the on-premises services and engagement offered. This resulted in customers conducting walk-in are less satisfied. On-premises services and customer engagement can be improved to ensure the targeted customer group experience increase satisfaction. An example can be a simple gesture of sending a personalized hand-written note to acknowledge the loyalty and continuous support from the customers which can significantly improve the customer relationship and appreciation leading to greater customer retention.

3. Third suggestion involves the use of a customer churn prediction model. The prediction model acts as an early detection system to identify customers that are likely to churn. This allows the bank to have an early engagement with the specified customers to understand the needs and situation of the customers and likely coming up with the right strategy to enhance the retention of such customers. As for the identified customers, regular customer contact and follow up would ensure the expectations and satisfactions of customers are met. The early intervention would likely change the minds of customers that are considering switching to the competitors thus reducing customer churn.

SECTION 5

CONCLUSION

Customer churning is a profound issue faced by many banks and various measures are taken to ensure the retention of their customers. This is due to the cost of acquiring new customers are typically higher than retaining the existing pool of customers. In this study, two customer churn prediction models using different prediction algorithms namely logistic regression and neural network were produced. The 2 hidden layer neural network model was chosen due to the higher recall rate of the model which is more suitable for the current scenario. Based on the model, it was identified that age, number of products purchased, and member active status played a significant role in identifying the likelihood of a customer churning. Which these features are pointing towards the older aged customers with history of purchasing from the bank and gone inactive. Several suggestions to increase customer retention based on the identified features were proposed which include knowing your customers, increasing customer satisfaction, and use of prediction model as an early detection system. However, further improvement to the study can be made in terms of increasing size of dataset to include more customers which the model can better learn the data patterns resulting in improved prediction accuracy. In addition, experimentation of different prediction algorithms is highly encouraged as there may be algorithms which may produce higher prediction accuracies such as decision trees or ensembles. Moreover, customer churn dataset is typically imbalanced and contains outliers. It is highly suggested to perform a thorough data pre-processing procedure to ensure prediction model is truly reflecting the dataset.

REFERENCES

- Broby, D. (2022). The use of predictive analytics in finance. *The Journal of Finance and Data Science*, 8, 145-161. doi:<https://doi.org/10.1016/j.jfds.2022.05.003>
- de Lima Lemos, R. A., Silva, T. C., & Tabak, B. M. (2022). Propension to customer churn in a financial institution: a machine learning approach. *Neural Computing and Applications*. doi:10.1007/s00521-022-07067-x
- Kamalaratnam, J. (2022). 5+ Ways Banks Can Retain Customers. Retrieved from <https://www.xerago.com/blog/5-ways-banks-can-retain-customers/>
- Tang, Q., Xia, G., Zhang, X., & Li, Y. (2020). *A Feature Interaction Network for Customer Churn Prediction*. Paper presented at the Proceedings of the 2020 12th International Conference on Machine Learning and Computing, Shenzhen, China. doi:10.1145/3383972.3384046
- Wang, X., Nguyen, K., & Nguyen, B. P. (2020). *Churn Prediction using Ensemble Learning*. Paper presented at the Proceedings of the 4th International Conference on Machine Learning and Soft Computing, Haiphong City, Vietnam. doi:10.1145/3380688.3380710
- Zoghby, J., Tieman, S., & Moino, J. P. (2018). Making It Personal. Retrieved from https://www.accenture.com/_acnmedia/PDF-77/Accenture-Pulse-Survey.pdf

APPENDIX A

MODEL STATISTICS & OUTPUTS

A.1 Baseline Regression Model Outputs

This section outlines the output results generated from SAS Enterprise Miner for the baseline regression model, which consist of the results for analysis of effects, maximum likelihood estimates, and fit statistics.

Type 3 Analysis of Effects			
Effect	DF	Wald	Pr > ChiSq
		Chi-Square	
IsActiveMember	1	132.8618	<.0001
LOG_REP_Age	1	291.9153	<.0001
LOG_REP_CreditScore	1	6.3007	0.0121
REP_Gender	1	34.8006	<.0001
REP_Geography	2	90.1176	<.0001
REP_NumOfProducts	3	276.2459	<.0001

Figure A.1: Analysis of effects for baseline regression model

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-5.0291	9.7228	0.27	0.6050		0.007
IsActiveMember	0 1	0.5487	0.0476	132.86	<.0001		1.731
LOG_REP_Age	1	3.5572	0.2082	291.92	<.0001	0.4915	35.065
LOG_REP_CreditScore	1	-0.7546	0.3006	6.30	0.0121	-0.0652	0.470
REP_Gender	0 1	-0.2758	0.0468	34.80	<.0001		0.759
REP_Geography	0 1	-0.3838	0.0626	37.57	<.0001		0.681
REP_Geography	1 1	0.6410	0.0705	82.76	<.0001		1.898
REP_NumOfProducts	0 1	-2.9125	9.5004	0.09	0.7592		0.054
REP_NumOfProducts	1 1	-4.3706	9.5005	0.21	0.6455		0.013
REP_NumOfProducts	2 1	-0.5025	9.5035	0.00	0.9578		0.605

Figure A.2: Maximum likelihood estimates for baseline regression model

Fit Statistics			
Target=Exited Target Label= ' '			
Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	2812.73	.
ASE	Average Squared Error	0.16	0.17
AVERR	Average Error Function	0.49	0.51
DFE	Degrees of Freedom for Error	2841.00	.
DFM	Model Degrees of Freedom	10.00	.
DFT	Total Degrees of Freedom	2851.00	.
DIV	Divisor for ASE	5702.00	2446.00
ERR	Error Function	2792.73	1250.48
FPE	Final Prediction Error	0.16	.
MAX	Maximum Absolute Error	0.99	0.98
MSE	Mean Square Error	0.16	0.17
NOBS	Sum of Frequencies	2851.00	1223.00
NW	Number of Estimate Weights	10.00	.
RASE	Root Average Sum of Squares	0.40	0.41
RFPE	Root Final Prediction Error	0.40	.
RMSE	Root Mean Squared Error	0.40	0.41
SBC	Schwarz's Bayesian Criterion	2872.29	.
SSE	Sum of Squared Errors	917.25	416.58
SUNW	Sum of Case Weights Times Freq	5702.00	2446.00
MISC	Misclassification Rate	0.23	0.26

Figure A.3: Fit statistics for baseline regression model

A.2 Poly-2-Regression Model Outputs

This section outlines the output results generated from SAS Enterprise Miner for the polynomial degree 2 regression model, which consist of the results for analysis of effects, maximum likelihood estimates, and fit statistics.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
IsActiveMember	1	132.8149	<.0001
LOG_REP_Age	1	74.2523	<.0001
REP_Gender	1	34.7645	<.0001
REP_Geography	2	90.1164	<.0001
REP_NumOfProducts	3	276.2780	<.0001
LOG_REP_Age*LOG_REP_CreditScore	1	6.5835	0.0103

Figure A.4: Analysis of effects for Poly-2-Regression model

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-9.9075	9.5314	1.08	0.2986		0.000
IsActiveMember	0 1	0.5486	0.0476	132.81	<.0001		1.731
LOG_REP_Age	1	4.8944	0.5680	74.25	<.0001	0.6763	133.534
REP_Gender	0 1	-0.2757	0.0468	34.76	<.0001		0.759
REP_Geography	0 1	-0.3840	0.0626	37.60	<.0001		0.681
REP_Geography	1 1	0.6409	0.0705	82.73	<.0001		1.898
REP_NumOfProducts	0 1	-2.9125	9.5004	0.09	0.7592		0.054
REP_NumOfProducts	1 1	-4.3712	9.5005	0.21	0.6454		0.013
REP_NumOfProducts	2 1	-0.5027	9.5035	0.00	0.9578		0.605
LOG_REP_Age*LOG_REP_CreditScore	1	-0.2068	0.0806	6.58	0.0103		0.813

Figure A.5: Maximum likelihood estimates for Poly-2-Regression model

Fit Statistics			
Target=Exited Target Label=' '			
Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	2812.44	.
ASE	Average Squared Error	0.16	0.17
AVERR	Average Error Function	0.49	0.51
DFE	Degrees of Freedom for Error	2841.00	.
DFM	Model Degrees of Freedom	10.00	.
DFT	Total Degrees of Freedom	2851.00	.
DIV	Divisor for ASE	5702.00	2446.00
ERR	Error Function	2792.44	1250.85
FPE	Final Prediction Error	0.16	.
MAX	Maximum Absolute Error	0.99	0.98
MSE	Mean Square Error	0.16	0.17
NOBS	Sum of Frequencies	2851.00	1223.00
NW	Number of Estimate Weights	10.00	.
RASE	Root Average Sum of Squares	0.40	0.41
RFPE	Root Final Prediction Error	0.40	.
RMSE	Root Mean Squared Error	0.40	0.41
SBC	Schwarz's Bayesian Criterion	2872.00	.
SSE	Sum of Squared Errors	917.11	416.69
SUMW	Sum of Case Weights Times Freq	5702.00	2446.00
MISC	Misclassification Rate	0.23	0.26

Figure A.6: Fit statistics for Poly-2-Regression model

A.3 Poly-3-Regression Model Outputs

This section outlines the output results generated from SAS Enterprise Miner for the polynomial degree 3 regression model, which consist of the results for analysis of effects, maximum likelihood estimates, and fit statistics.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
IsActiveMember	1	131.8177	<.0001
LOG_REP_Age	1	37.5927	<.0001
REP_Gender	1	34.7468	<.0001
REP_Geography	2	90.1189	<.0001
REP_NumOfProducts	3	275.7586	<.0001
LOG_REP_Age*LOG_REP_Age*LOG_REP_CreditScore	1	7.7261	0.0054

Figure A.7: Analysis of effects for Poly-3-Regression model

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept	1	-15.1025	9.7251	2.41	0.1204		0.000
IsActiveMember	0 1	0.5465	0.0476	131.82	<.0001		1.727
LOG_REP_Age	1	6.3589	1.0371	37.59	<.0001	0.8786	577.637
REP_Gender	0 1	-0.2757	0.0468	34.75	<.0001		0.759
REP_Geography	0 1	-0.3837	0.0626	37.58	<.0001		0.681
REP_Geography	1 1	0.6412	0.0705	82.74	<.0001		1.899
REP_NumOfProducts	0 1	-2.9136	9.5076	0.09	0.7593		0.054
REP_NumOfProducts	1 1	-4.3703	9.5077	0.21	0.6458		0.013
REP_NumOfProducts	2 1	-0.4960	9.5107	0.00	0.9584		0.609
LOG_REP_Age*LOG_REP_Age*LOG_REP_CreditScore	1	-0.0582	0.0209	7.73	0.0054		0.943

Figure A.8: Maximum likelihood estimates for Poly-3-Regression model

Fit Statistics			
Target=Exited Target Label=' '			
Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	2811.29	.
ASE	Average Squared Error	0.16	0.17
AVERR	Average Error Function	0.49	0.51
DFE	Degrees of Freedom for Error	2841.00	.
DFM	Model Degrees of Freedom	10.00	.
DFT	Total Degrees of Freedom	2851.00	.
DIV	Divisor for ASE	5702.00	2446.00
ERR	Error Function	2791.29	1252.28
FPE	Final Prediction Error	0.16	.
MAX	Maximum Absolute Error	0.99	0.98
MSE	Mean Square Error	0.16	0.17
NOBS	Sum of Frequencies	2851.00	1223.00
NW	Number of Estimate Weights	10.00	.
RASE	Root Average Sum of Squares	0.40	0.41
RFPE	Root Final Prediction Error	0.40	.
RMSE	Root Mean Squared Error	0.40	0.41
SBC	Schwarz's Bayesian Criterion	2870.85	.
SSE	Sum of Squared Errors	916.03	417.04
SUMW	Sum of Case Weights Times Freq	5702.00	2446.00
MISC	Misclassification Rate	0.23	0.26

Figure A.9: Fit statistics for Poly-3-Regression model

A.4 1 Hidden Layer Neural Network Model Outputs

Below shows the output of the optimal 1 hidden layer neural network model found with the hidden neuron configuration of 10 hidden neurons. The outputs include fit statistics and misclassification rate curve.

Fit Statistics			
Target=Exited Target Label=' '			
Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.14	0.16
DIV	Divisor for ASE	5702.00	2446.00
MAX	Maximum Absolute Error	0.98	1.00
NOBS	Sum of Frequencies	2851.00	1223.00
RASE	Root Average Squared Error	0.38	0.40
SSE	Sum of Squared Errors	804.61	394.89
DISF	Frequency of Classified Cases	2851.00	1223.00
MISC	Misclassification Rate	0.21	0.24
WRONG	Number of Wrong Classifications	600.00	299.00

Figure A.10: Fit statistics for 1 hidden layer neural network model

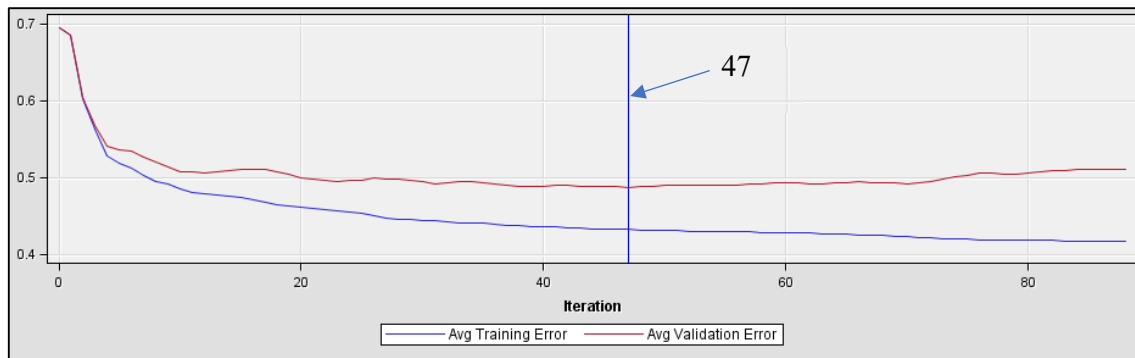


Figure A.11: Misclassification rate for 1 hidden layer neural network model

A.5 2 Hidden Layer Neural Network Model Outputs

Below shows the output of the optimal 2 hidden layer neural network model found with the hidden neuron configuration of 6, 6 hidden neurons. The outputs include fit statistics and misclassification rate curve.

Fit Statistics				
Target=Exited Target Label=' '				
Fit Statistics	Statistics Label	Train	Validation	
ASE	Average Squared Error	0.14	0.16	
DIV	Divisor for ASE	5702.00	2446.00	
MAX	Maximum Absolute Error	0.99	1.00	
NOBS	Sum of Frequencies	2851.00	1223.00	
RASE	Root Average Squared Error	0.38	0.40	
SSE	Sum of Squared Errors	822.82	397.54	
DISF	Frequency of Classified Cases	2851.00	1223.00	
MISC	Misclassification Rate	0.21	0.24	
WRONG	Number of Wrong Classifications	609.00	293.00	

Figure A.12: Fit statistics for 2 hidden layer neural network model

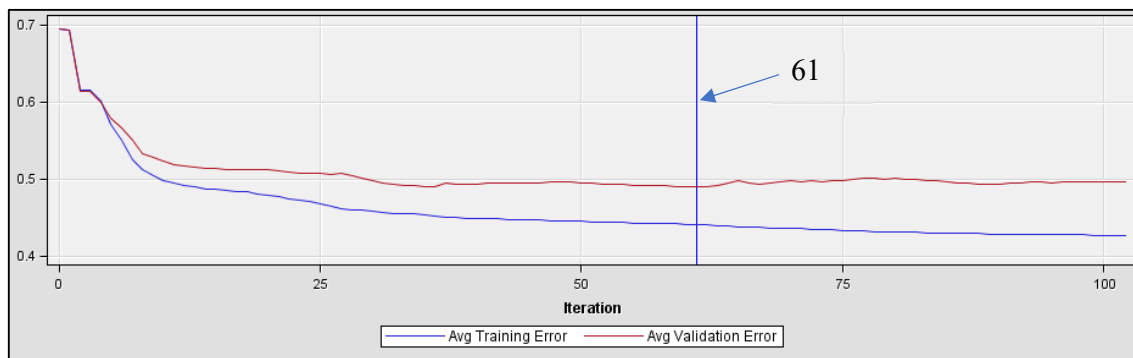


Figure A.13: Misclassification rate for 2 hidden layer neural network model

A.6 3 Hidden Layer Neural Network Model Outputs

Below shows the output of the optimal 3 hidden layer neural network model found with the hidden neuron configuration of 8, 8, 8 hidden neurons. The outputs include fit statistics and misclassification rate curve.

Fit Statistics				
Target=Exited Target Label=' '				
Fit Statistics	Statistics Label	Train	Validation	
ASE	Average Squared Error	0.14	0.16	
DIV	Divisor for ASE	5702.00	2446.00	
MAX	Maximum Absolute Error	0.99	1.00	
NOBS	Sum of Frequencies	2851.00	1223.00	
RASE	Root Average Squared Error	0.38	0.40	
SSE	Sum of Squared Errors	822.78	397.17	
DISF	Frequency of Classified Cases	2851.00	1223.00	
MISC	Misclassification Rate	0.20	0.24	
WRONG	Number of Wrong Classifications	584.00	291.00	

Figure A.14: Fit statistics for 3 hidden layer neural network model

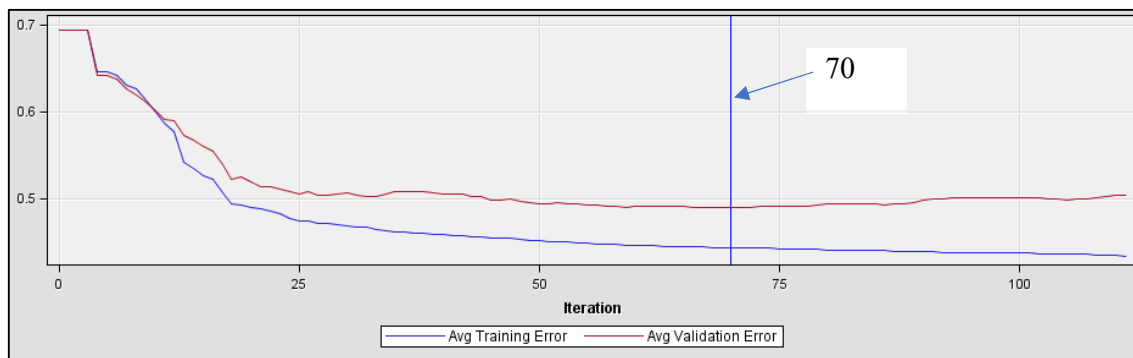


Figure A.15: Misclassification rate for 3 hidden layer neural network model

A.7 4 Hidden Layer Neural Network Model Outputs

Below shows the output of the optimal 4 hidden layer neural network model found with the hidden neuron configuration of 10, 8, 6, 4 hidden neurons. The outputs include fit statistics and misclassification rate curve.

Fit Statistics				
Target=Exited Target Label=' '				
Fit				
Statistics	Statistics Label	Train	Validation	
ASE	Average Squared Error	0.14	0.16	
DIV	Divisor for ASE	5702.00	2446.00	
MAX	Maximum Absolute Error	0.99	0.99	
NOBS	Sum of Frequencies	2851.00	1223.00	
RASE	Root Average Squared Error	0.37	0.40	
SSE	Sum of Squared Errors	783.35	393.38	
DISF	Frequency of Classified Cases	2851.00	1223.00	
MISC	Misclassification Rate	0.20	0.24	
WRONG	Number of Wrong Classifications	573.00	294.00	

Figure A.16: Fit statistics for 4 hidden layer neural network model

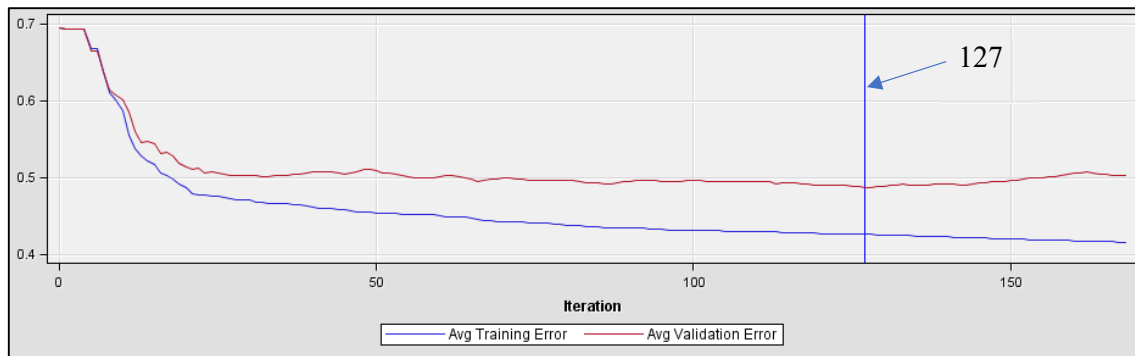


Figure A.17: Misclassification rate for 4 hidden layer neural network model

A.8 Surrogate Tree Complete List of Node Rules

This section outlines every node rule generated by the surrogate model for the 2 hidden layer neural network model. In addition, at the end shows the formula used for converting the log transformed variable back to the original state.

```
*-----*
Node = 4
*-----*
if Transformed: Replacement: Age < 3.79543 or MISSING
AND Replacement: NumOfProducts IS ONE OF: 1
then
Tree Node Identifier    = 4
Number of Observations = 731
Predicted: U_Exited=1 = 0.03
Predicted: U_Exited=0 = 0.97

*-----*
Node = 11
*-----*
if Transformed: Replacement: Age >= 3.79543
AND Replacement: NumOfProducts IS ONE OF: 1
AND IsActiveMember IS ONE OF: 1 or MISSING
then
Tree Node Identifier    = 11
Number of Observations = 153
Predicted: U_Exited=1 = 0.20
Predicted: U_Exited=0 = 0.80

*-----*
Node = 15
*-----*
if Transformed: Replacement: Age >= 3.67622 or MISSING
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND IsActiveMember IS ONE OF: 0 or MISSING
then
Tree Node Identifier    = 15
Number of Observations = 680
Predicted: U_Exited=1 = 0.98
Predicted: U_Exited=0 = 0.03

*-----*
Node = 18
*-----*
if Transformed: Replacement: Balance < 5.36551
AND Transformed: Replacement: Age >= 3.79543
AND Replacement: NumOfProducts IS ONE OF: 1
AND IsActiveMember IS ONE OF: 0
then
Tree Node Identifier    = 18
Number of Observations = 37
Predicted: U_Exited=1 = 0.38
Predicted: U_Exited=0 = 0.62
```

```

*-----*
Node = 19
*-----*
if Transformed: Replacement: Balance >= 5.36551 or MISSING
AND Transformed: Replacement: Age >= 3.79543
AND Replacement: NumOfProducts IS ONE OF: 1
AND IsActiveMember IS ONE OF: 0
then
  Tree Node Identifier    = 19
  Number of Observations = 77
  Predicted: U_Exited=1 = 0.99
  Predicted: U_Exited=0 = 0.01

*-----*
Node = 23
*-----*
if Transformed: Replacement: Age < 3.67622
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 1
AND Replacement: Gender IS ONE OF: 1 or MISSING
then
  Tree Node Identifier    = 23
  Number of Observations = 101
  Predicted: U_Exited=1 = 0.98
  Predicted: U_Exited=0 = 0.02

*-----*
Node = 24
*-----*
if Transformed: Replacement: Balance < 10.8144
AND Transformed: Replacement: Age < 3.67622
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 0, 2 or MISSING
then
  Tree Node Identifier    = 24
  Number of Observations = 143
  Predicted: U_Exited=1 = 0.55
  Predicted: U_Exited=0 = 0.45

*-----*
Node = 27
*-----*
if Transformed: Replacement: Age >= 3.67622 or MISSING
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 1
AND IsActiveMember IS ONE OF: 1
then
  Tree Node Identifier    = 27
  Number of Observations = 154
  Predicted: U_Exited=1 = 1.00
  Predicted: U_Exited=0 = 0.00

```

```

*-----*
Node = 42
*-----*
if Transformed: Replacement: Balance >= 10.8144 or MISSING
AND Transformed: Replacement: Age < 3.67622
AND Replacement: NumOfProducts IS ONE OF: 2, 3
AND Replacement: Geography IS ONE OF: 0, 2 or MISSING
then
Tree Node Identifier   = 42
Number of Observations = 17
Predicted: U_Exited=1 = 1.00
Predicted: U_Exited=0 = 0.00

*-----*
Node = 43
*-----*
if Transformed: Replacement: Balance >= 10.8144 or MISSING
AND Transformed: Replacement: Age < 3.67622
AND Replacement: NumOfProducts IS ONE OF: 0 or MISSING
AND Replacement: Geography IS ONE OF: 0, 2 or MISSING
then
Tree Node Identifier   = 43
Number of Observations = 370
Predicted: U_Exited=1 = 0.02
Predicted: U_Exited=0 = 0.98

*-----*
Node = 44
*-----*
if Transformed: Replacement: Balance < 10.7354
AND Transformed: Replacement: Age >= 3.67622 or MISSING
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 0, 2 or MISSING
AND IsActiveMember IS ONE OF: 1
then
Tree Node Identifier   = 44
Number of Observations = 110
Predicted: U_Exited=1 = 0.90
Predicted: U_Exited=0 = 0.10

*-----*
Node = 56
*-----*
if Transformed: Replacement: Age < 3.34975
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 1
AND Replacement: Gender IS ONE OF: 0
AND IsActiveMember IS ONE OF: 0
then
Tree Node Identifier   = 56
Number of Observations = 5
Predicted: U_Exited=1 = 0.00
Predicted: U_Exited=0 = 1.00

```

```

*-----*
Node = 57
*-----*
if Transformed: Replacement: Age < 3.67622 AND Transformed: Replacement: Age >= 3.34975 or MISSING
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 1
AND Replacement: Gender IS ONE OF: 0
AND IsActiveMember IS ONE OF: 0
then
Tree Node Identifier = 57
Number of Observations = 42
Predicted: U_Exited=1 = 1.00
Predicted: U_Exited=0 = 0.00

*-----*
Node = 58
*-----*
if Transformed: Replacement: Age < 3.62425 or MISSING
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 1
AND Replacement: Gender IS ONE OF: 0
AND IsActiveMember IS ONE OF: 1 or MISSING
then
Tree Node Identifier = 58
Number of Observations = 38
Predicted: U_Exited=1 = 0.08
Predicted: U_Exited=0 = 0.92

*-----*
Node = 59
*-----*
if Transformed: Replacement: Age < 3.67622 AND Transformed: Replacement: Age >= 3.62425
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 1
AND Replacement: Gender IS ONE OF: 0
AND IsActiveMember IS ONE OF: 1 or MISSING
then
Tree Node Identifier = 59
Number of Observations = 12
Predicted: U_Exited=1 = 0.92
Predicted: U_Exited=0 = 0.08

*-----*
Node = 66
*-----*
if Transformed: Replacement: Balance >= 10.7354 or MISSING
AND Transformed: Replacement: Age < 3.88151 AND Transformed: Replacement: Age >= 3.67622 or MISSING
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 0, 2 or MISSING
AND IsActiveMember IS ONE OF: 1
then
Tree Node Identifier = 66
Number of Observations = 92
Predicted: U_Exited=1 = 0.12
Predicted: U_Exited=0 = 0.88

```

```

*-----*
Node = 67
*-----*
if Transformed: Replacement: Balance >= 10.7354 or MISSING
AND Transformed: Replacement: Age >= 3.88151
AND Replacement: NumOfProducts IS ONE OF: 0, 2, 3 or MISSING
AND Replacement: Geography IS ONE OF: 0, 2 or MISSING
AND IsActiveMember IS ONE OF: 1
then
  Tree Node Identifier   = 67
  Number of Observations = 89
  Predicted: U_Exited=1 = 0.60
  Predicted: U_Exited=0 = 0.40

```

Computed Transformations (maximum 500 observations printed)					
Input Name	Role	Input Level	Name	Level	Formula
REP_Age	INPUT	INTERVAL	LOG_REP_Age	INTERVAL	$\log(\text{REP_Age} + 1)$
REP_Balance	INPUT	INTERVAL	LOG_REP_Balance	INTERVAL	$\log(\text{REP_Balance} + 1)$
REP_CreditScore	INPUT	INTERVAL	LOG_REP_CreditScore	INTERVAL	$\log(\text{REP_CreditScore} + 1)$

Figure A.18: Formula for log transformed variables

The following shows the equation used to convert log transformed age feature back to original scale.

$$\text{LOG_REP_Age} = \log(\text{REP_Age} + 1)$$

$$e^{\text{LOG_REP_Age}} = \text{REP_Age} + 1$$

$$\text{REP_Age} = e^{\text{LOG_REP_Age}} - 1$$