

**UNIVERSITY  
OF MALAYA**

**MASTER OF DATA SCIENCE (SEMESTER 2 – 2022/2023)**

**FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY**

**WQD7005 DATA MINING**

**GROUP ASSIGNMENT**

## **BANK CUSTOMER CHURN PREDICTION**

Group Member	Matric Number
Xian Longdi	S2172650

## Table of Contents

1. Introduction .....	3
2. Dataset Description.....	4
3. Business Understanding .....	5
3.1 Analysis Goal .....	5
3.2 Analysis Data .....	5
4. Methodology.....	5
4.1 SEMMA Description .....	5
4.2 SEMMA Process .....	5
5. Results .....	6
5.1 Sample.....	6
5.1.1 Metadata .....	7
5.1.2 Reclassification of the Role and Level of the Variables.....	8
5.2 Explore .....	9
5.2.1 Summary Statistics .....	9
5.2.2 Univariate Analysis.....	11
5.2.3 Bivariate Analysis – Variable Association.....	17
5.2.4 Multivariate Analysis.....	20
5.2.5 Interesting Visualization.....	25
5.2.6 Model: Decision Tree .....	25
6. Conclusion.....	32
Appendix .....	34

# 1. Introduction

The banking sector has increasingly used data-driven marketing techniques to improve customer acquisition and retention strategies. Bank marketing involves promoting bank products and services to potential customers through various channels such as direct mail, email, and telemarketing. The data for bank marketing is typically stored in a database or spreadsheet, with each row representing a single customer and each column containing customer attributes such as age, job, marital status, education, financial information, and contact details.

To implement the data-driven environment in banking sector, many tools can be used. One of the tool examples is SAS (Statistical Analysis System). SAS (Statistical Analysis System) is a popular statistical software package that provides a range of data mining and predictive modeling tools. One popular approach to data mining in SAS is the SEMMA methodology, which stands for *Sample, Explore, Modify, Model, and Assess*. SEMMA provides a structured framework for conducting data mining projects, starting with selecting and preparing a representative sample of the data.

In the *Sample* stage, researchers first identify a representative sample of the data for analysis. This sample should be selected using a random sampling technique and should be large enough to provide reliable estimates of model parameters. Once a sample has been selected, researchers can move on to the Explore stage.

In the *Explore* stage, researchers use various data visualization and summary techniques to gain insights into the data. This involves visualizing and summarizing the data to identify patterns and trends and identifying any outliers or data quality issues that need to be addressed before conducting further analysis.

In the *Modify* stage, researchers make changes to the data, such as transforming variables or creating new variables based on existing ones, to improve the accuracy of the models. For example, researchers may create a new variable that combines information about a customer's job and education to predict better their likelihood of subscribing to a bank product.

Thereafter, researchers get into the *Model* stage to model customer behavior and predict customer outcomes. This step is possible as SAS itself provides various modeling techniques, including logistic regression, decision trees, and random forests.

Finally, in the *Assess* stage, researchers evaluate the performance of the models by comparing predicted outcomes to actual outcomes. This allows researchers to identify the most accurate and reliable models, which can then be used to guide marketing and business decisions.

Overall, the SEMMA methodology provides a structured approach to data mining in SAS, which can be particularly useful for conducting large-scale marketing research projects in the banking sector. By following the SEMMA framework, researchers enable to ensure that the analysis is reliable, accurate, and actionable, leading to better marketing and business outcomes for banks and financial institutions.

## 2. Dataset Description

This analysis will use a dataset that contains information about bank customers' behavior in using various products. It consists of 14 columns, where '*Customer Id*' is a unique identifier assigned to each customer. The dataset also includes demographic information such as "*surname*", "*gender*", "*age*", and "*geographic location*" which describe the location of customers. The '*Credit Rating*' list displays the credit rating of each customer. The "*Usage Period*" column displays the duration of the customer's use of banking services.

The dataset also provides information about customer account activity, such as the '*Balance*' column, which displays the customer's average balance. The '*NumOfProducts*' list shows the number of bank products the customer uses. The "*HasCrCard*" list shows whether the customer has a credit card, and the "*IsActiveMember*" column shows whether the customer is an active bank member. The "*Estimated Salary*" list shows the estimated salary of the customer, and the target variable "*Exited*" indicates whether the customer has stirred up (i.e., closed their account). Overall, this dataset provides valuable insights into the behavior and characteristics of bank customers and can be used to predict which customers are most likely to lose in the future.

Table 1: Variables' description

Variables	Description
age	customer age (numeric)
CustomerId	CustomerID is given (numeric)
Surname	Surname of the customer (numeric)
CreditScore	Credit Score of customers (numeric)
Geography	location of customer (categorical: 'France', 'Spain'...)
housing	Gender whether male or female (categorical: 'male', 'female')
Tenure	From how many years customer is in bank (numeric)
balance	Average balance of customer (numeric)
NumOfProducts	Number of bank product facilities customer is using (numeric)
HasCrCard	Has the credit card been activated (Binary:'0','1')
IsActiveMember	Has the membership card been activated (Binary:'0','1')
EstimatedSalary	Customer's expected salary (numeric)
Exited	Is the customer still present (Binary:'0','1')

## 3. Business Understanding

### 3.1 Analysis Goal

The goal of this analysis is to predict whether a bank customer will churn or not based on their demographic and financial information. The term "churn" refers to customers who stop using a bank's services or close their account altogether. By identifying customers who are likely to churn, the bank can take proactive measures to retain them and prevent loss of business. In further detail, the analysis goals can be pointed as follows:

- i. Understand the basic situation of bank customers.
- ii. Determine the reason for customer churn based on the relationship between all attributes of the bank and all factors of the customer.
- iii. Based on the analysis above, evaluate the predicting model of whether customers will churn or not.

### 3.2 Analysis Data

The source of this data was downloaded from the UCI (UC Irvine) (the UC Irvine Machine Learning Repository). The dataset is publicly available. The data relates to direct marketing activities of a Portuguese banking institution. The marketing activities are based on telephone calls. Usually, several contacts with the same customer are required to find out if the product (bank term deposit) will be subscribed.

## 4. Methodology

### 4.1 SEMMA Description

SEMMA is an abbreviation for the data mining process developed by SAS (Statistical Analysis System). It represents sampling, exploration, modification, modeling, and evaluation. Each stage of SEMMA aims to help users navigate the data mining process and provide a systematic approach to understanding data. SEMMA consists of five main steps, namely sampling, exploration, modification, modeling, and evaluation, all of which are crucial for successful data mining projects. These steps are easily available in the SAS Enterprise Miner tool.

### 4.2 SEMMA Process

In this project, the first two methods (Sample and Explore) are used for implementation, with a focus on analysis. Figure 4.1 shows the process of SEMMA.

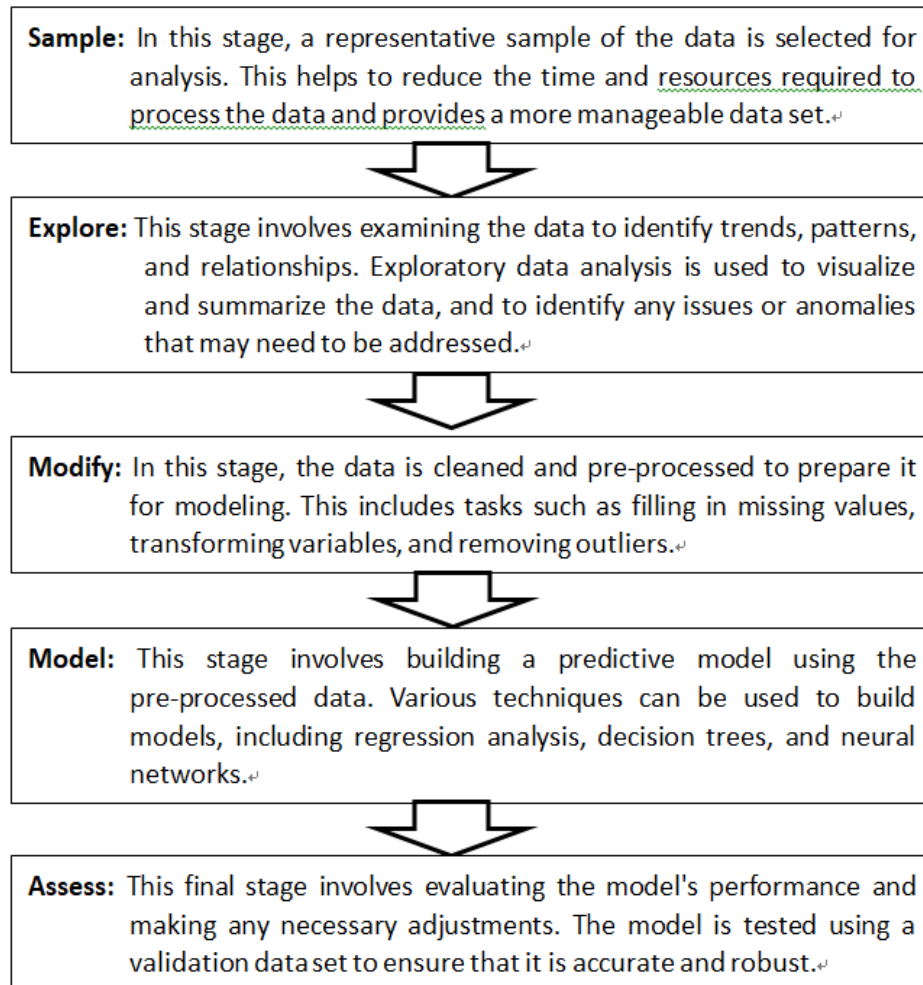


Figure 4.1 SEMMA process

## 5. Results

### 5.1 Sample

The collected data is saved in CSV format. For data exploration, the CSV file has been imported and saved as an SAS file. Containing 4,522 records and 17 variables, the collected dataset includes a set of activity data with customer level information, including demographic data, usage, and income activities of consumers before and after the event.

Collected variables are:

- i. customer age (age)
- ii. customerID is given (CustomerId)
- iii. surname of the customer (Surname)
- iv. credit Score of customers (CreditScor)
- v. location of customer (Geography)
- vi. gender whether male or female (housing)

- vii. From how many years customer is in bank (Tenure)
- viii. Average balance of customer (balance)
- ix. Number of bank product facilities customer is using (NumOfProducts))
- x. Has the credit card been activated (HasCrCard)
- xi. Has the membership card been activated (isActiveMember)
- xii. Customer's expected salary (EstimatedSalary)
- xiii. Is the customer still present (Exited)

### 5.1.1 Metadata

There are basic and complex parameters available in SAS Enterprise Miner to specify the data formats for the variable. Figure 5.1 displays the metadata for the columns, and Figure 5.2 displays the data types in their most basic configurations. Figure 5.3 displays the metadata for the columns, and Figure 5.4 displays the data types in their default advanced settings.

Name	Role	Level	Report	O...	Drop	Lower	Upper	Type	Format	Informat	Length
Age	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Balance	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
CreditScore	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
CustomerId	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
EstimatedSalary	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Exited	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Gender	Input	Nominal	No		No	.	.	Character	\$6.	\$6.	6
Geography	Input	Nominal	No		No	.	.	Character	\$7.	\$7.	7
HasCrCard	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
IsActiveMember	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
NumOfProducts	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
RowNumber	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Surname	Input	Nominal	No		No	.	.	Character	\$17.	\$17.	17
Tenure	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8

Figure 5.1: column metadata (Basic setting)

Metadata Completed.		
Library:	GROUPDAT	
Data Source:	basicdata	
Role:	Raw	
<b>Role</b>	<b>Level</b>	<b>Count</b>
Input	Interval	11
Input	Nominal	3

Figure 5.2: Data type summary (Basic setting)

According to Figures 5.1 and 5.2, we observe that the data types are divided into nominal and interval as input roles for the default basic system settings. Based on potential values for variables, the system automatically detects measurement levels. Characters are classed as nominal by default, whereas numerical data is classified as interval.

Because most data types are still inaccurate, the basic settings are not suitable for implementation. After selecting advanced settings (based on Figures 5.3 and 5.4), the data

types were redefined as binary, interval, and nominal. The system automatically detects the role of variables as input and reject roles. Therefore, there are still errors where the data type cannot reflect the correct data type of the dataset. Therefore, before entering the next stage (exploration), manual adjustments are needed to modify the data type.

Name	Role	Level	Report	...	Drop	Lower	Upper	Type	Format	Informat	Length
Age	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Balance	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
CreditScore	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
CustomerId	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
EstimatedSalary	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Exited	Input	Binary	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Gender	Input	Binary	No		No	.	.	Character	\$6.	\$6.	6
Geography	Input	Nominal	No		No	.	.	Character	\$7.	\$7.	7
HasCrCard	Input	Binary	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
IsActiveMember	Input	Binary	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
NumOfProducts	Input	Nominal	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
RowNumber	Input	Interval	No		No	.	.	Numeric	BEST12.0	BEST32.0	8
Surname	Rejected	Nominal	No		No	.	.	Character	\$17.	\$17.	17
Tenure	Input	Nominal	No		No	.	.	Numeric	BEST12.0	BEST32.0	8

Figure 5.3: column metadata (Advanced setting)


Metadata Completed.		
Library:	GROUPDAT	
Data Source:	Groupdataset	
Role:	Raw	
Role	Level	Count
Input	Binary	4
Input	Interval	6
Input	Nominal	3
Rejected	Nominal	1

Figure5.4: Data type summary (Advanced setting)

### 5.1.2 Reclassification of the Role and Level of the Variables

Reclassification is to create appropriate charts for these variables. Figure 5.5 shows the manually adjusted data type modification, and Figure 5.6 is the final data type summary.

Name /	Role	Level
Age	Input	Interval
Balance	Input	Interval
CreditScore	Input	Interval
CustomerId	Input	Interval
EstimatedSalary	Input	Interval
Exited	Input	Binary
Gender	Input	Binary
Geography	Input	Nominal
HasCrCard	Input	Binary
IsActiveMember	Input	Binary
NumOfProducts	Input	Nominal
RowNumber	Input	Interval
Surname	Rejected	Nominal
Tenure	Input	Nominal



Name	Role	Level
Age	Input	Interval
Balance	Input	Interval
CreditScore	Input	Interval
CustomerId	Input	Interval
EstimatedSalary	Input	Interval
Exited	Target	Binary
Gender	Input	Binary
Geography	Input	Nominal
HasCrCard	Input	Binary
IsActiveMember	Input	Binary
NumOfProducts	Input	Nominal
RowNumber	Input	Interval
Surname	Input	Nominal
Tenure	Input	Nominal

Figure 5.5: Comparison between role and data level on advance settings and manual reclassification



Library:	GROUPDAT	
Data Source:	group	
Role:	Raw	
Role	Level	Count
Input	Binary	3
Input	Interval	6
Input	Nominal	4
Target	Binary	1

Figure 5.6: Data type summary

According to Figure 5.5, the rejected variables are changed into input variables because all data should be included before any analysis is conducted to demonstrate the reasonableness of variable deletion. The target variable is manually added. Assuming the unary and binary variables of nominal data are converted, the remaining variables remain unchanged.

In addition, we observed that some variables, such as Age, Estimated Salary, and Balance, are interval variables. This is because their values range from 0 to positive infinity. The data type is displayed as a nominal value. Once data cleaning is performed during the modification phase, the data types of these variables will be more accurate.

## 5.2 Explore

Data was explored to identify the relationships and anomalies via univariate, bivariate and multivariate analysis with several graphs.

### 5.2.1 Summary Statistics

A statistical data summary will be produced after accessing and examining the dataset. Data patterns including minimum and maximum values, mean values, missing values, and standard deviations are outlined by summary statistics. The summary statistical data for interval variables are shown in Figure 5.7, and the summary statistical data for class variables are shown in Figure 5.8.

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
CustomerId	INPUT	15690941	71936.19	10000	0	15565701	15690733	15815690	0.001149	-1.19584
EstimatedSalary	INPUT	100090.2	57510.49	10000	0	11.58	100187.4	199992.5	0.002085	-1.18152
Exited	INPUT	0.2037	0.402769	10000	0	0	0	1	1.471611	0.165671
HasCrCard	INPUT	0.7055	0.45584	10000	0	0	1	1	-0.90181	-1.18697
IsActiveMember	INPUT	0.5151	0.499797	10000	0	0	1	1	-0.06044	-1.99675
NumOfProducts	INPUT	1.5302	0.581654	10000	0	1	1	4	0.745568	0.582981
RowNumber	INPUT	5000.5	2886.896	10000	0	1	5000	10000	0	-1.2
Tenure	INPUT	5.0128	2.892174	10000	0	0	5	10	0.010991	-1.16523

Figure 5.7: Interval Variable Summary Statistics

According to Figure 5.7, there are 14 variables without missing values, and summary

statistics are regular. All data is within the valid range, so this dataset meets the research requirements.

Class Variable Summary Statistics  
(maximum 500 observations printed)

Data Role=TRAIN

Data	Variable	Number	Mode	Mode2
Role	Name	of Levels	Percentage	Percentage
TRAIN	Gender	2	Male 54.57	Female 45.43
TRAIN	Geography	3	France 50.14	Germany 25.09
TRAIN	Surname	513	He 0.78	Shih 0.78

Figure 5.8: Class Variable Summary Statistics

Based on Figure 5.8, there were no missing values among the class variables. However, part of the nominal variables such as surname contained an extremely high number of levels which was abnormal.

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737688	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	15792365	He	501	France	Male	44	4	142051.07	2	0	1	74840.5	0
10	15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0
11	15767821	Bearce	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
12	15737173	Andrews	497	Spain	Male	24	3	0	2	1	0	76390.01	0
13	15632264	Kay	478	France	Female	34	10	0	2	1	0	26260.98	0
14	15691483	Chin	549	France	Female	25	5	0	2	0	0	190857.79	0
15	15600892	Scott	635	Spain	Female	35	7	0	2	1	1	85951.65	0
16	15643966	Coforth	616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0
17	15737452	Romeo	653	Germany	Male	58	1	132602.88	1	1	0	5097.67	1
18	15788218	Henderson	549	Spain	Female	24	9	0	2	1	1	14406.41	0
19	1561507	Muldrow	587	Spain	Male	45	6	0	1	0	0	158984.81	0
20	15568982	Hao	726	France	Female	24	6	0	2	1	1	54724.03	0
21	15577857	McDonald	732	France	Male	41	8	0	2	1	1	170886.17	0

Figure 5.9: Overview of data

Based on Figure 5.9, the dataset being studied is already comprehensive and includes all the necessary variables to predict customer churn. The dataset includes a unique identifier for each customer, "CustomerId". Furthermore, the dataset includes important financial information such as the "CreditScore" and "Balance" columns, which provide insight into the creditworthiness and financial status of the customer. The "Geography" column provides information about the customer's location, which can be useful for analyzing regional trends in customer behavior.

The demographic information provided in the "Gender" and "Age" columns can also be valuable for understanding customer behavior patterns. The "Tenure" column provides information about how long the customer has been with the bank, which can be a crucial factor in predicting customer churn. The "NumOfProducts" column indicates the number of bank products the customer is currently using, while the "HasCrCard" and "IsActiveMember" columns provide information about their current account activity. Finally, the "EstimatedSalary" column provides an estimate of the customer's salary, which can be relevant for predicting churn.

## 5.2.2 Univariate Analysis

Univariate analysis, the simplest form of statistical data analysis, is used to explore each variable in a data set. It does not deal with causality or causal relationship, but only discovers patterns from each variable. Graphs such as histograms, block diagrams, and pie charts are best suited for univariate analysis to examine pattern distribution, outliers, noise data, and missing values.

Firstly, using a histogram to visualize the distribution and missing values of 7 interval variables. Figure 5.9 shows an overview of all charts, while Table 5.1 summarizes the patterns, analyses, and anomalies.

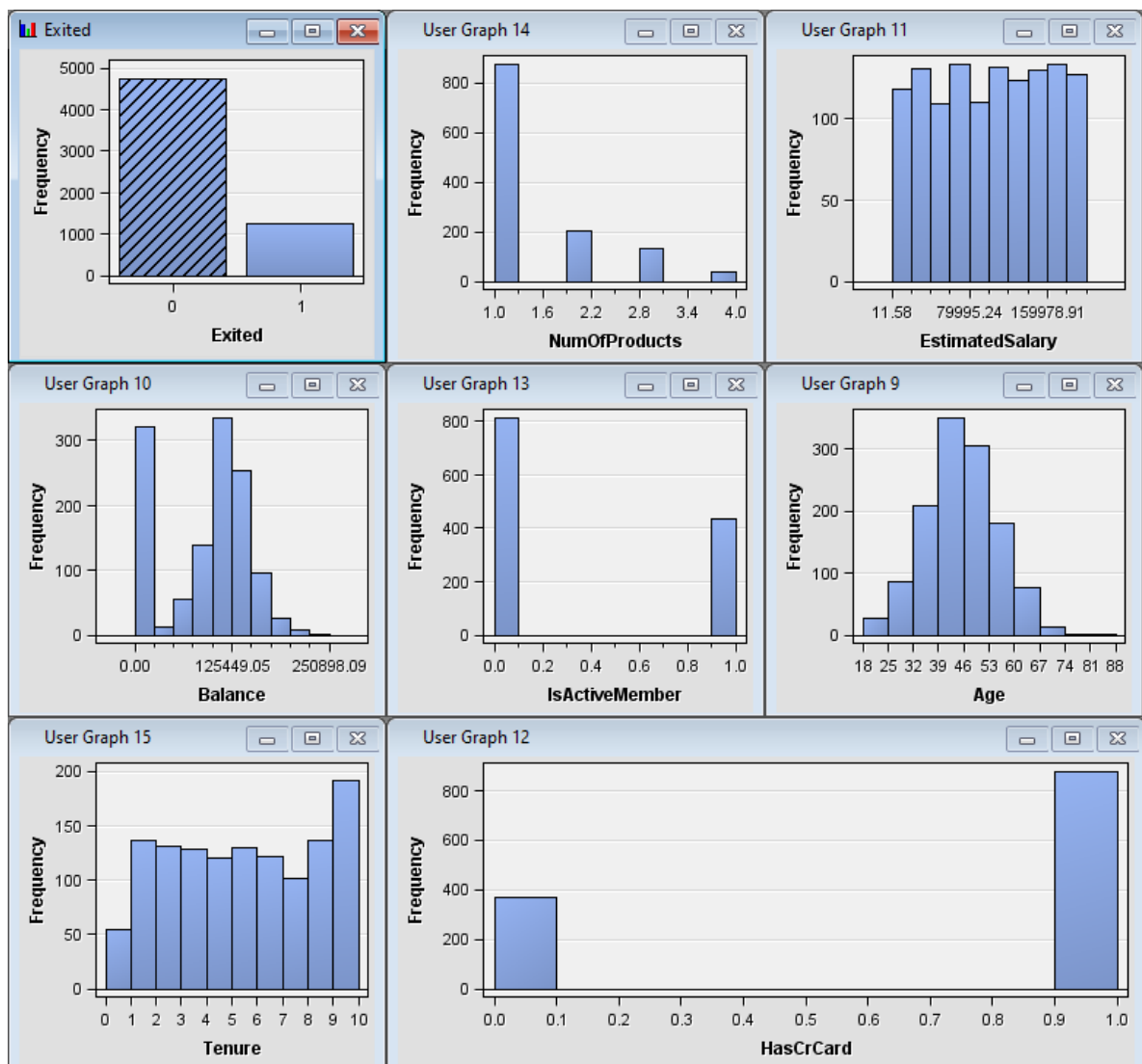
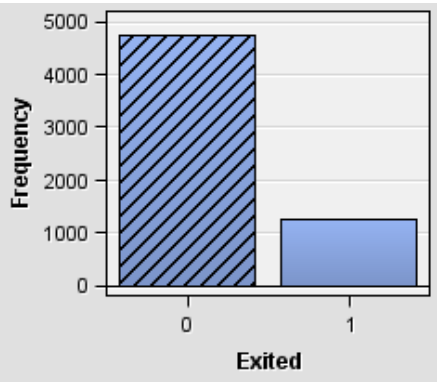
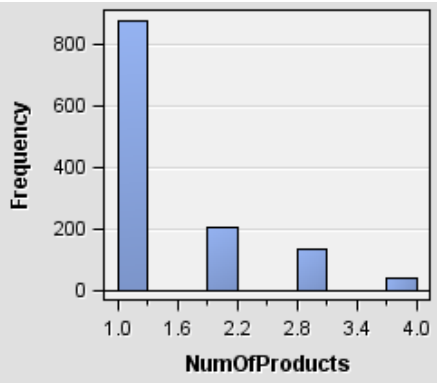
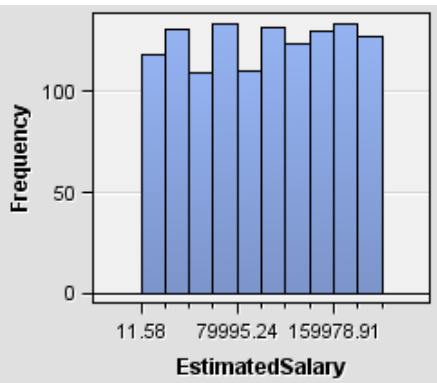
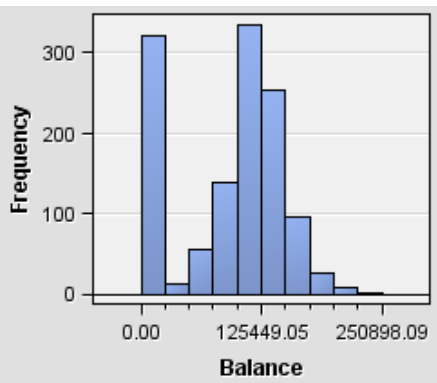
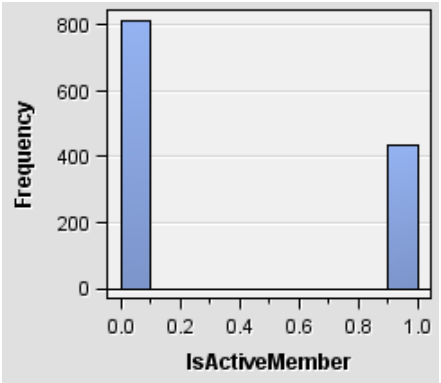
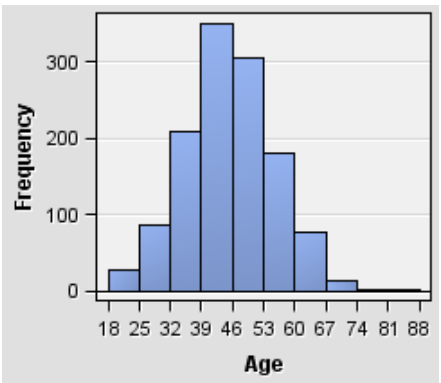
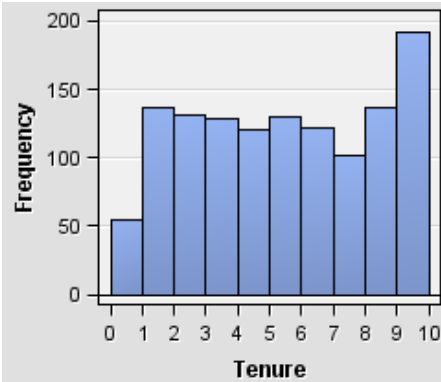
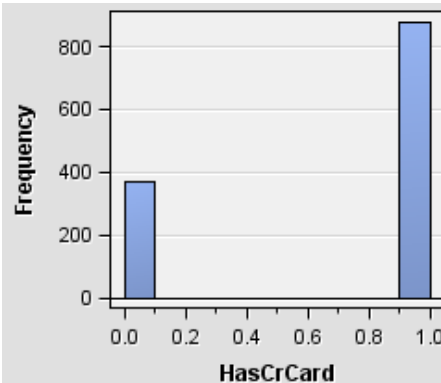


Figure 5.10: Histogram of interval variable

Table 5.1: Variables' pattern description

1	 <p>A bar chart titled 'Exited' with 'Frequency' on the y-axis (0 to 5000) and 'Exited' on the x-axis (0 and 1). The bar for '0' is hatched and reaches a frequency of approximately 4800. The bar for '1' is solid blue and reaches a frequency of approximately 1200.</p>	<p>There are far more no “exited” users than active users, and the number of “exited” users is about 4800, indicating that banks should carry out more marketing activities to increase user activity.</p> <p>No outliers were detected.</p>
2	 <p>A bar chart titled 'NumOfProducts' with 'Frequency' on the y-axis (0 to 800) and 'NumOfProducts' on the x-axis (1.0, 1.6, 2.2, 2.8, 3.4, 4.0). The bar for '1.0' is the tallest, reaching a frequency of approximately 850. Other bars are at 2.2 (~200), 2.8 (~150), and 4.0 (~50).</p>	<p>The number of banking products customers use is "1" at most, and over 800.</p> <p>No outliers were detected.</p>
3	 <p>A bar chart titled 'EstimatedSalary' with 'Frequency' on the y-axis (0 to 100) and 'EstimatedSalary' on the x-axis (11.58, 79995.24, 159978.91, and others). The bars are relatively uniform in height, mostly between 100 and 120.</p>	<p>"EstimatedSalary" - Customer's expected salary most evenly balanced distribution, indicating that the client's salary level is more balanced.</p> <p>No outliers were detected.</p>
4	 <p>A bar chart titled 'Balance' with 'Frequency' on the y-axis (0 to 300) and 'Balance' on the x-axis (0.00, 125449.05, 250898.09, and others). The distribution is skewed to the right, with a peak frequency of approximately 320 at the 125449.05 category.</p>	<p>The average balance of customer presents a normal distribution, but there are also groups with a lower average balance.</p> <p>No outliers were detected.</p>

5	 <p>A histogram showing the frequency of users based on their 'IsActiveMember' status. The x-axis is labeled 'IsActiveMember' and ranges from 0.0 to 1.0. The y-axis is labeled 'Frequency' and ranges from 0 to 800. There are two bars: one at 0.0 with a frequency of approximately 800, and another at 1.0 with a frequency of approximately 450.</p>	<p>Most of the bank customers are inactive users, and the inactive users are like no “exited”, about 800 people, indicating that there may be more potential lost users.</p> <p>No outliers were detected.</p>
6	 <p>A histogram showing the frequency of users by age. The x-axis is labeled 'Age' and has categories: 18, 25, 32, 39, 46, 53, 60, 67, 74, 81, 88. The y-axis is labeled 'Frequency' and ranges from 0 to 300. The distribution is roughly bell-shaped, peaking at age 46 with a frequency of approximately 350.</p>	<p>The age of users presents a normal distribution, and the oldest group is between 39-46 years old.</p> <p>No outliers were detected.</p>
7	 <p>A histogram showing the frequency of users by tenure in years. The x-axis is labeled 'Tenure' and ranges from 0 to 10. The y-axis is labeled 'Frequency' and ranges from 0 to 200. The distribution is roughly bell-shaped, peaking at tenure 9 with a frequency of approximately 190.</p>	<p>From how many years a customer is in the bank, the distribution of users with a tenure of 1-7 years is average, and the number of users with a tenure of 9-10 years is large.</p> <p>No outliers were detected.</p>
8	 <p>A histogram showing the frequency of users based on whether they have a credit card. The x-axis is labeled 'HasCrCard' and ranges from 0.0 to 1.0. The y-axis is labeled 'Frequency' and ranges from 0 to 800. There are two bars: one at 0.0 with a frequency of approximately 350, and another at 1.0 with a frequency of approximately 850.</p>	<p>Customers with credit cards are not easily lost, while customers without credit cards are easily lost.</p> <p>No outliers were detected.</p>

In addition, a boxplot is one of the good tools for visualizing outliers. Similarly, a box chart was created for 9 interval variables to verify the existence of an outlier, as shown in Figure 5.10. Table 3 summarizes the findings of the outlier.

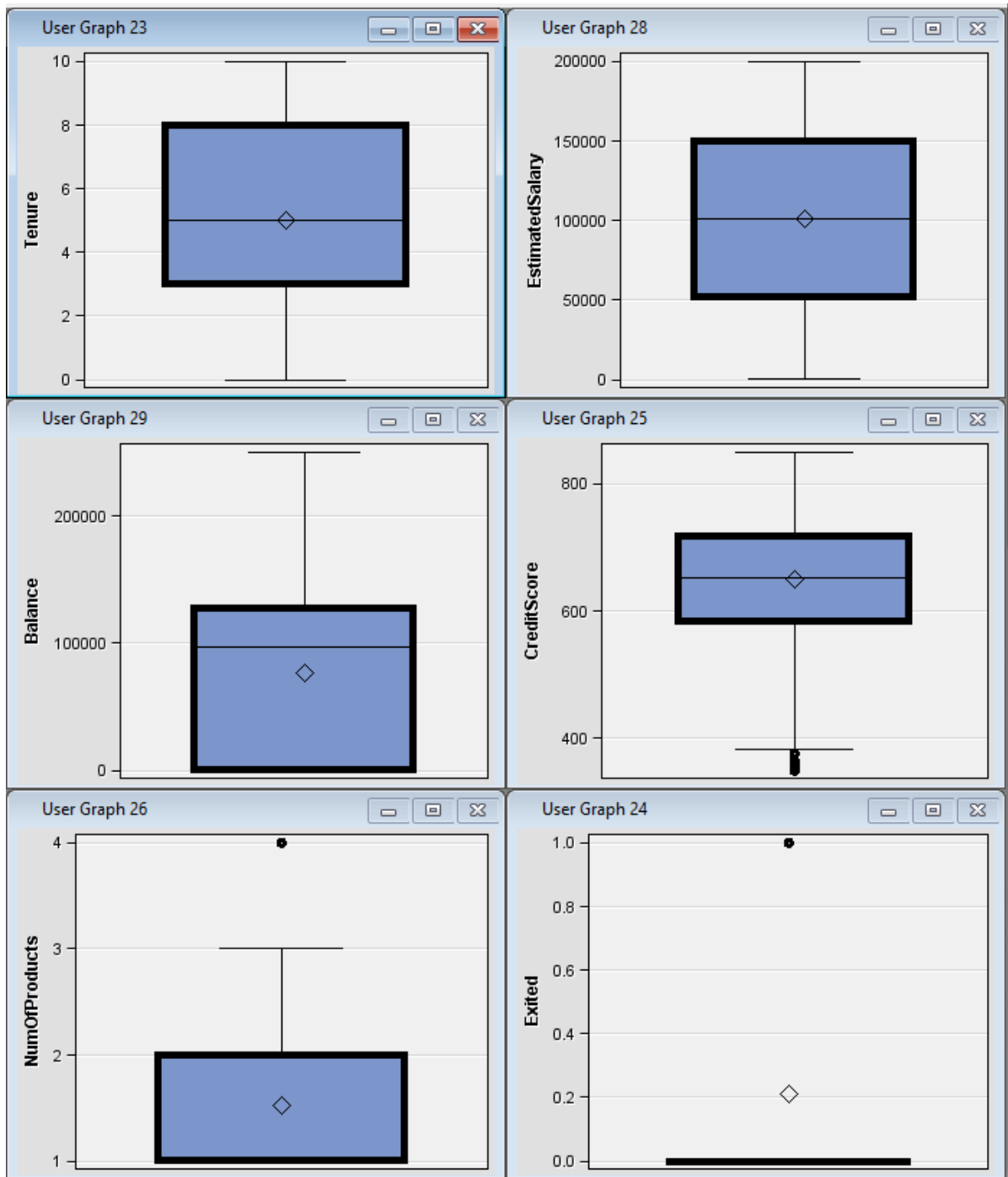
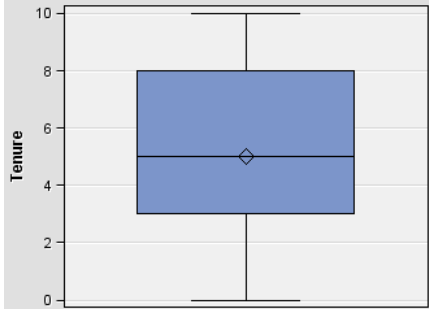
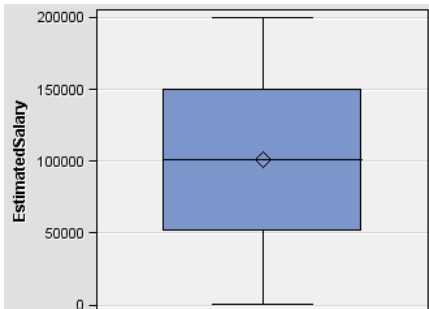
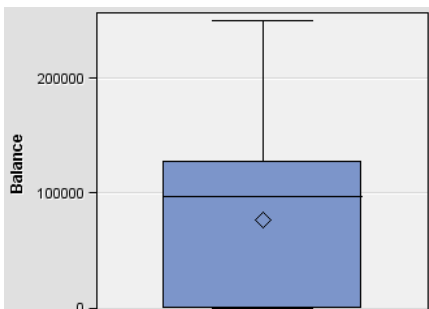
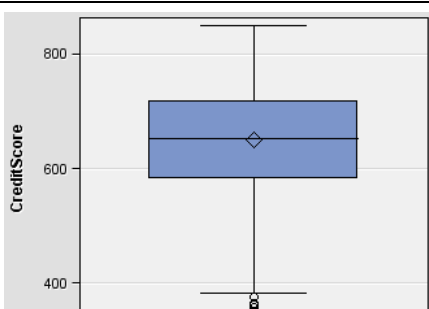
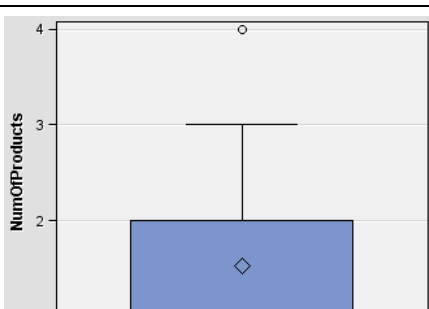


Figure 5.11: Box plots of interval variables

Table 5.2: Variables' boxplot description

1	 A boxplot for the variable 'Tenure'. The y-axis is labeled 'Tenure' and ranges from 0 to 10. The box is blue, with a median line at 5. The interquartile range (IQR) is from 3 to 8. Whiskers extend from 0 to 10. A diamond marker is at the median (5).	The average level of “Tenure” is 5. In 3-8 means the bin contains 50% of the data. Therefore, the width of the box reflects the degree of fluctuation of the data in the range of 3-8.
2	 A boxplot for the variable 'EstimatedSalary'. The y-axis is labeled 'EstimatedSalary' and ranges from 0 to 200,000. The box is blue, with a median line at 100,000. The IQR is from 50,000 to 150,000. Whiskers extend from 0 to 200,000. A diamond marker is at the median (100,000).	The average level of “EstimatedSalary” is 100000. In 5000-150000 indicates a wide range of salary levels, and marketing activities can be personalized and recommended according to different salary level stages.
3	 A boxplot for the variable 'Balance'. The y-axis is labeled 'Balance' and ranges from 0 to 200,000. The box is blue, with a median line at 100,000. The IQR is from 0 to 125,000. Whiskers extend from 0 to 200,000. A diamond marker is at the median (100,000).	The average balance of customer indicates the average level is 100000. Individual "balance" will be exceptionally large, and the overall distribution is expected to be relatively low. Targeted marketing is possible for large users.
4	 A boxplot for the variable 'CreditScore'. The y-axis is labeled 'CreditScore' and ranges from 400 to 800. The box is blue, with a median line at 650. The IQR is from 580 to 720. Whiskers extend from 380 to 850. A diamond marker is at the median (650). There are several outliers below 400.	The average level of Credit Score of customers is almost 650. The whole is concentrated, and the data span is not large. But overall, it is closer to the upper quartile.
5	 A boxplot for the variable 'NumOfProducts'. The y-axis is labeled 'NumOfProducts' and ranges from 1 to 4. The box is blue, with a median line at 1.5. The IQR is from 1 to 2. Whiskers extend from 1 to 3. A diamond marker is at the median (1.5). There is one outlier at 4.	The number of bank product facilities customers use overall is at 1-2, and one outlier at 4.

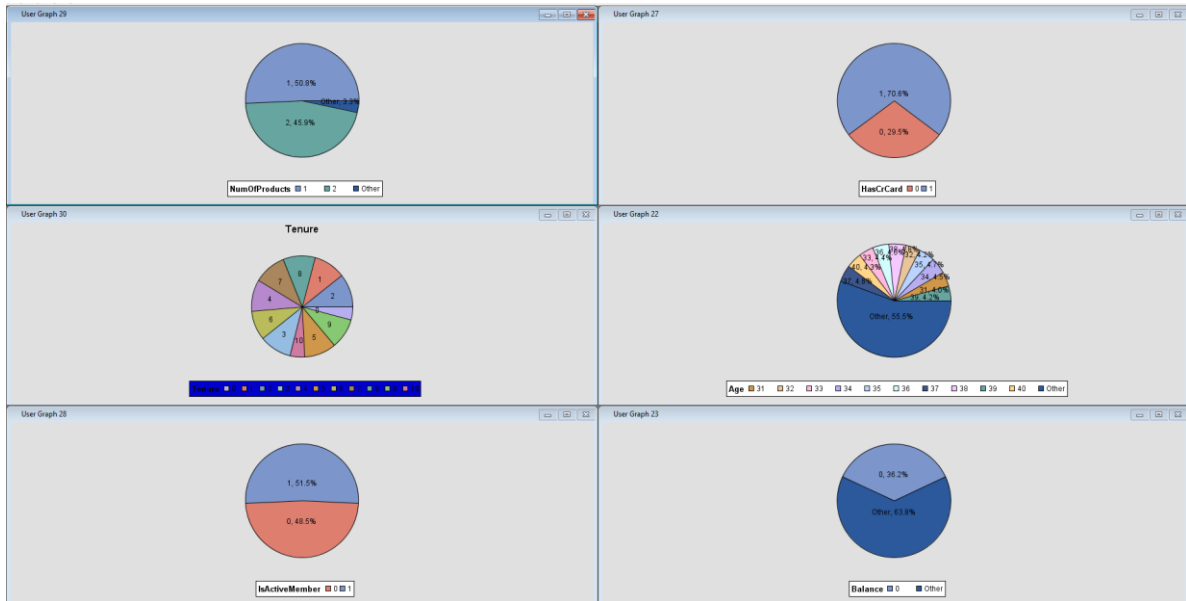
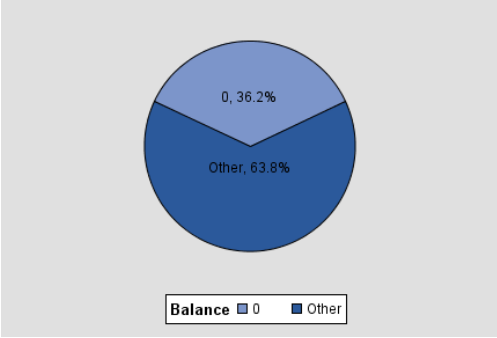
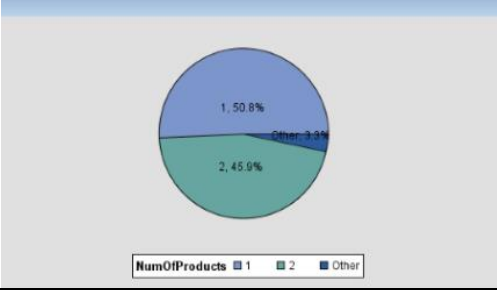
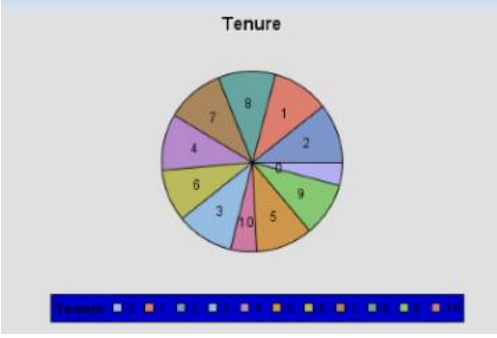


Figure 5.12: Pie plots of normal variables

Table 5.3: Pie chart finding's description

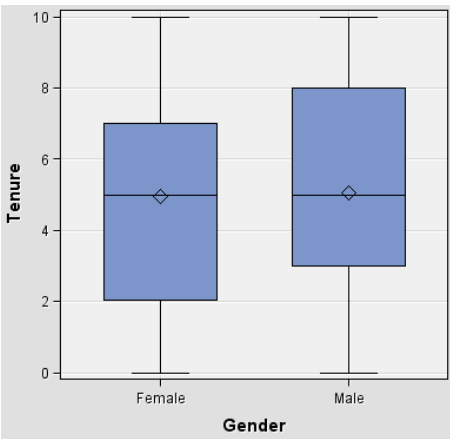
1	<p>HasCrCard 0 1</p>	<p>More than 70.6% of consumers have a credit card, and only 29.5% of consumers do not have a credit card. The number of people with a credit card is significantly higher than those without one. Credit plays a significant role in customer churn and profitability.</p>
2	<p>IsActiveMember 0 1</p>	<p>51.5% of people activated their membership card, while 0.48% of people did not activate their membership card. The difference between the two is not significant. This factor is not the main reason for customer churn.</p>
3	<p>Age 31 32 33 34 35 36 37 38 39 40 Other</p>	<p>The proportion of people over 40 years old is 55.5%, while the proportion of all other people is only 44.5%, indicating that the elderly population is more likely to lose compared to the younger population.</p>



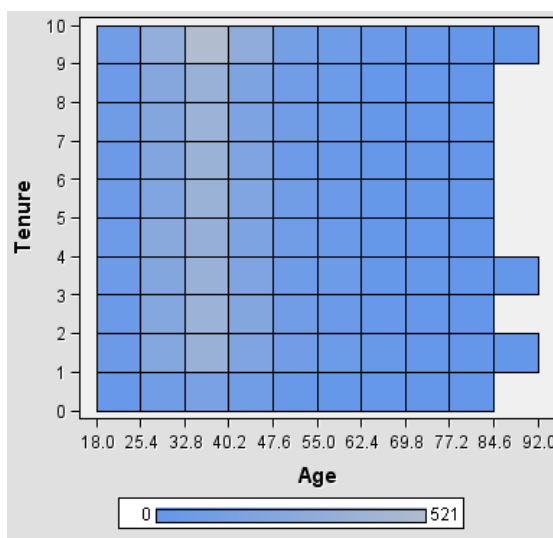
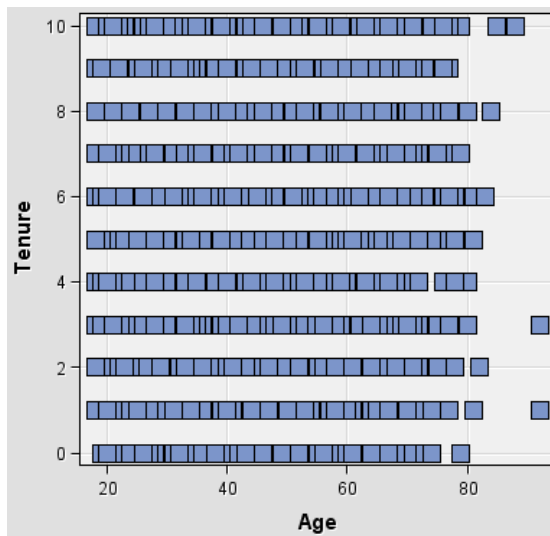
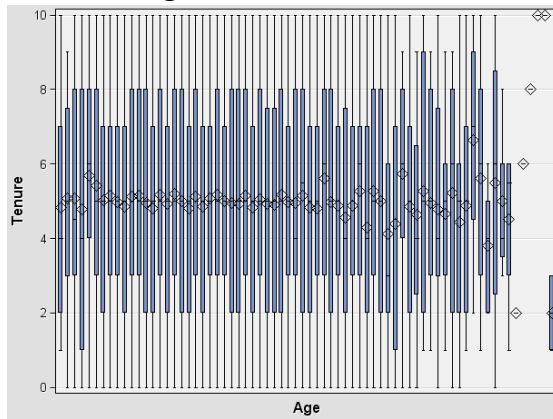
4		The balance of 36.2% of the population turned out to be 0, indicating that many people no longer deposit into the bank
5		More than half of the population believe that the convenience frequency can only be 1, which clearly proves that banks are losing population.
6		The number of people in the bank in the year 0 is extremely small. Most concentrated in 5, 8, 7, 4, and 9 years

### 5.2.3 Bivariate Analysis – Variable Association

Bivariate analysis indicates the relationship between two variables. Charts with simple insights, including scatter plots, bar plots, and box plots, can be quite helpful. The results between the variables in the dataset were shown in Table 5.4.

1	<p><b>Tenure Vs Gender</b></p> 	The distribution between males and females is terribly similar and there is slight difference in average age.
---	--	---

2

**Tenure Vs Age**

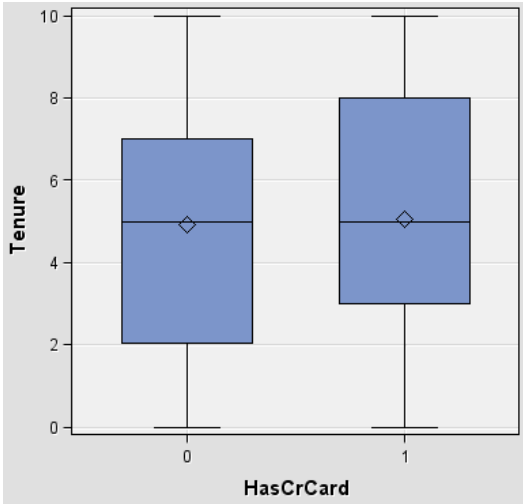
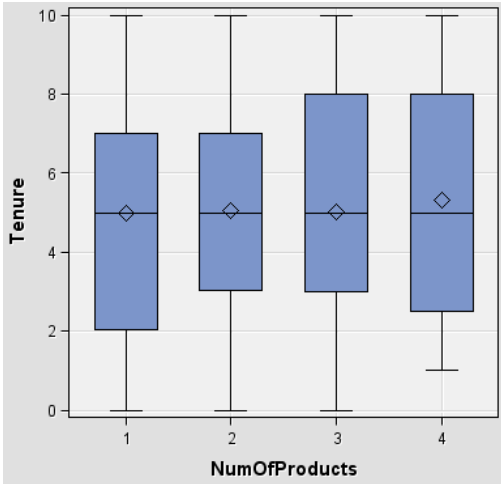
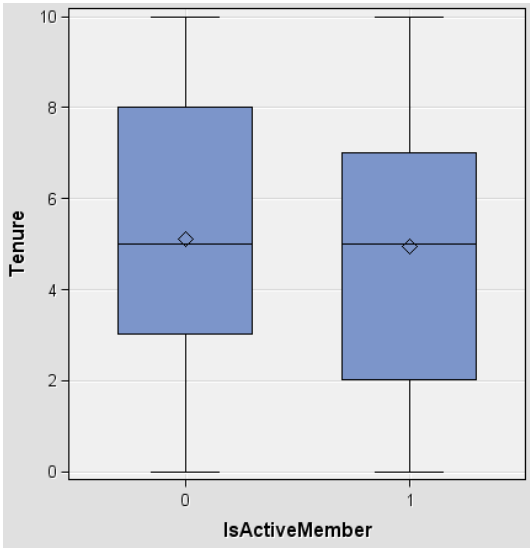
The younger the age, the shorter the tenure.

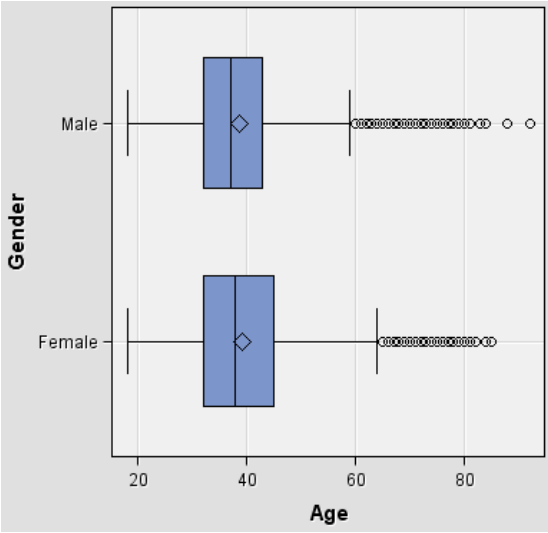
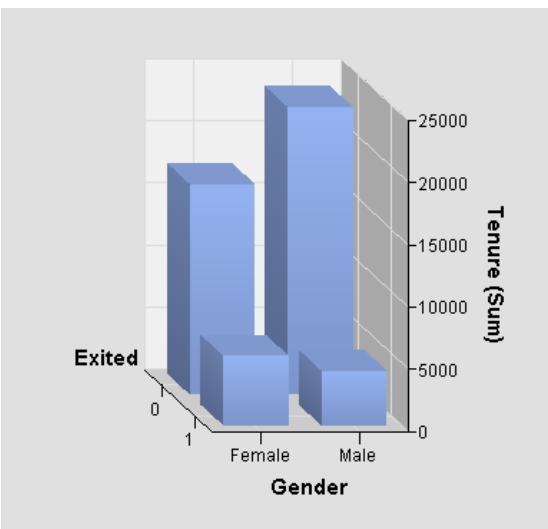
tenure is focused on 4-6 years.

3

**Tenure Vs HasCrCard**

“HasCrCard” category “1” is more widely distributed and has a longer tenure, up to 8 years.

	 <p>A box plot showing the distribution of Tenure (Y-axis, 0 to 10) for two categories of HasCrCard (X-axis: 0 and 1). For HasCrCard = 0, the median tenure is approximately 5 years, with a box from 2 to 7 and whiskers from 0 to 10. For HasCrCard = 1, the median tenure is approximately 5 years, with a box from 3 to 8 and whiskers from 0 to 10.</p>	<p>HasCrCard category "0" has a shorter tenure, with the shortest being 2 years.</p>
4	<p><b>Tenure Vs NumOfProducts</b></p>  <p>A box plot showing the distribution of Tenure (Y-axis, 0 to 10) for four categories of NumOfProducts (X-axis: 1, 2, 3, 4). For NumOfProducts = 1, the median tenure is approximately 5 years, with a box from 2 to 7 and whiskers from 0 to 10. For NumOfProducts = 2, the median tenure is approximately 5 years, with a box from 3 to 7 and whiskers from 0 to 10. For NumOfProducts = 3, the median tenure is approximately 5 years, with a box from 3 to 8 and whiskers from 0 to 10. For NumOfProducts = 4, the median tenure is approximately 5 years, with a box from 2.5 to 8 and whiskers from 1 to 10.</p>	<p>Of the four categories, Product 4 is the one that covers the longest term. Consider continuing to develop Product 4 subsequently.</p>
5	<p><b>Tenure vs isActiveMember</b></p>  <p>A box plot showing the distribution of Tenure (Y-axis, 0 to 10) for two categories of isActiveMember (X-axis: 0 and 1). For isActiveMember = 0, the median tenure is approximately 5 years, with a box from 3 to 8 and whiskers from 0 to 10. For isActiveMember = 1, the median tenure is approximately 5 years, with a box from 2 to 7 and whiskers from 0 to 10.</p>	<p>Most of the active users who belong to "1" have a short tenure Most of the customers with longer tenure are inactive users</p>
6	<p><b>Gender Vs Age</b></p>	<p>In the relationship between gender and age of clients, the number of genders is evenly distributed, with more females</p>

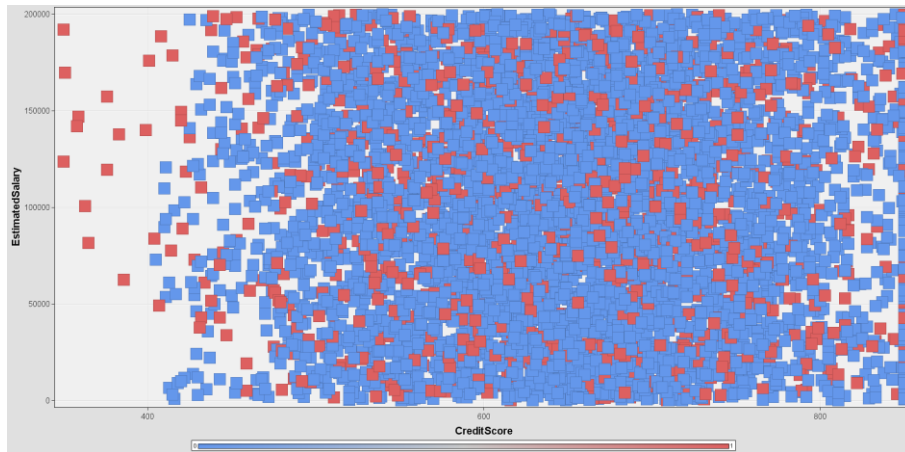
		<p>overall.</p> <p>The age of men and women was concentrated around 40 years old.</p>
7	<p><b>Tenure, Gender Vs Exited</b></p> 	<p>The majority of non-retention users indicate serious customer churn, with more male than female customers lost. Among the retained users, there are more females than males.</p> <p>This suggests that banks can personalize their marketing activities to target non-retention customers based on customer characteristics such as gender, age, etc. For retained customers, regular promotions can be used to increase customer stickiness.</p>

**Table 5.4: Variable Association's findings**

### 5.2.4 Multivariate Analysis

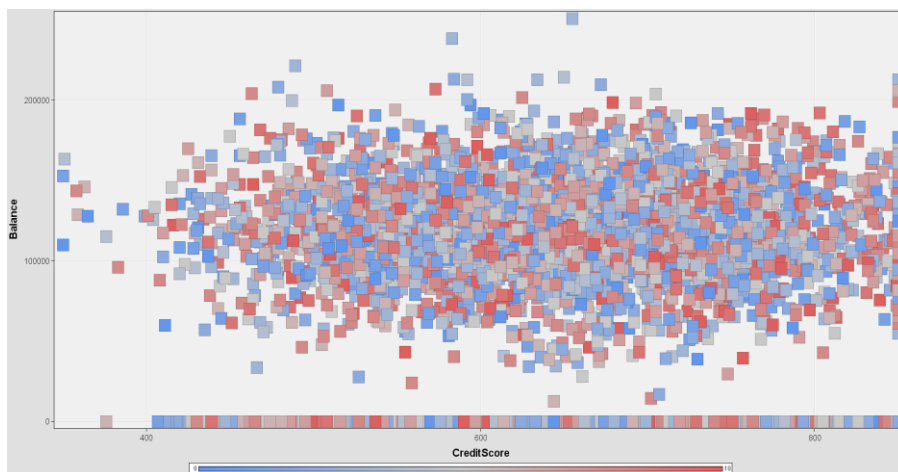
Multivariate analysis is a statistical process involving multiple dependent variables in producing an outcome. In other words, more than two dependent variables are analyzed simultaneously with all other variables.

Figure 5.13 illustrates the credit scores and estimated wages in this scatter chart. The chart shows the relationship between credit score and estimated salary under exited. And most of these units are in the middle of the value. And just little in small value parts. and small parts in huge value area.

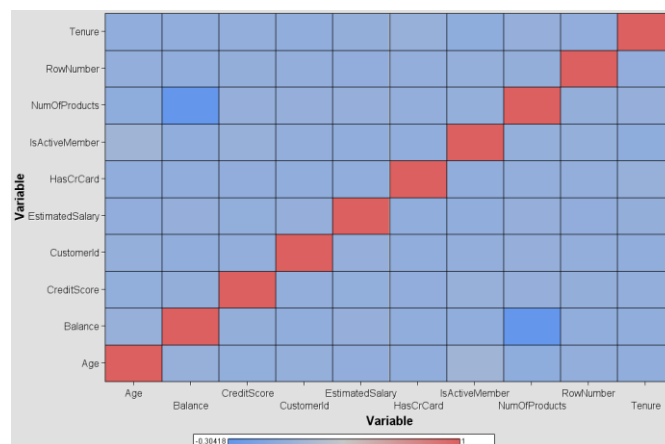


**Figure 5.13: Findings of creditscore and estimate salary whin exited**

Figure 5.14 illustrates the credit scores and balance in this scatter chart. The chart shows the relationship between credit score and balance under exited. And most of these units are in the middle of the value. And just little in small value parts. and small parts in huge value area. And the biggest value of credit score has unity parts.



**Figure 5.14: Findings of creditscore and balance whin exited**



**Figure 5.15: Correlation Matrix**

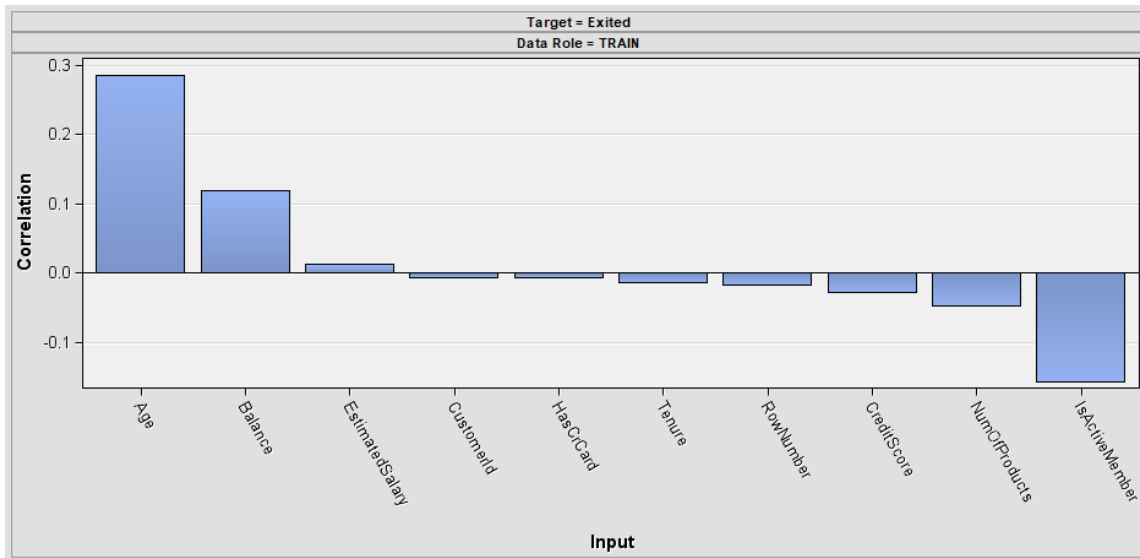
Variable	Variable	Correlation
Age	Age	1
Balance	Age	0.028308
CreditScore	Age	-0.00396
CustomerId	Age	0.009497
EstimatedSalary	Age	-0.0072
HasCrCard	Age	-0.01172
IsActiveMember	Age	0.085472
NumOfProducts	Age	-0.03068
RowNumber	Age	.0007826
Tenure	Age	-0.01
Age	Balance	0.028308
Balance	Balance	1
CreditScore	Balance	0.006268
CustomerId	Balance	-0.01242
EstimatedSalary	Balance	0.012797
HasCrCard	Balance	-0.01486
IsActiveMember	Balance	-0.01008
NumOfProducts	Balance	-0.30418
RowNumber	Balance	-0.00907
Tenure	Balance	-0.01225
Age	CreditScore	-0.00396
Balance	CreditScore	0.006268
CreditScore	CreditScore	1
CustomerId	CreditScore	0.005308
EstimatedSalary	CreditScore	-0.00138
HasCrCard	CreditScore	-0.00546
IsActiveMember	CreditScore	0.025651
NumOfProducts	CreditScore	0.012238
RowNumber	CreditScore	0.00584
Tenure	CreditScore	.0008419
Age	CustomerId	0.009497
Balance	CustomerId	-0.01242
CreditScore	CustomerId	0.005308
CustomerId	CustomerId	1
EstimatedSalary	CustomerId	0.015271
HasCrCard	CustomerId	-0.01403
IsActiveMember	CustomerId	0.001665
NumOfProducts	CustomerId	0.016972
RowNumber	CustomerId	0.004202
Tenure	CustomerId	-0.01488
Age	EstimatedSalary	-0.0072
Balance	EstimatedSalary	0.012797
CreditScore	EstimatedSalary	-0.00138
CustomerId	EstimatedSalary	0.015271
EstimatedSalary	EstimatedSalary	1
HasCrCard	EstimatedSalary	-0.00993
IsActiveMember	EstimatedSalary	-0.01142
NumOfProducts	EstimatedSalary	0.014204
RowNumber	EstimatedSalary	-0.00599
Tenure	EstimatedSalary	0.007784

Age	HasCrCard	-0.01172
Balance	HasCrCard	-0.01486
CreditScore	HasCrCard	-0.00546
CustomerId	HasCrCard	-0.01403
EstimatedSalary	HasCrCard	-0.00993
HasCrCard	HasCrCard	1
IsActiveMember	HasCrCard	-0.01187
NumOfProducts	HasCrCard	0.003183
RowNumber	HasCrCard	.0005987
Tenure	HasCrCard	0.022583
Age	IsActiveMember	0.085472
Balance	IsActiveMember	-0.01008
CreditScore	IsActiveMember	0.025651
CustomerId	IsActiveMember	0.001665
EstimatedSalary	IsActiveMember	-0.01142
HasCrCard	IsActiveMember	-0.01187
IsActiveMember	IsActiveMember	1
NumOfProducts	IsActiveMember	0.009612
RowNumber	IsActiveMember	0.012044
Tenure	IsActiveMember	-0.02836
Age	NumOfProducts	-0.03068
Balance	NumOfProducts	-0.30418
CreditScore	NumOfProducts	0.012238
CustomerId	NumOfProducts	0.016972
EstimatedSalary	NumOfProducts	0.014204
HasCrCard	NumOfProducts	0.003183
IsActiveMember	NumOfProducts	0.009612
NumOfProducts	NumOfProducts	1
RowNumber	NumOfProducts	0.007246
Tenure	NumOfProducts	0.013444
Age	RowNumber	.0007826
Balance	RowNumber	-0.00907
CreditScore	RowNumber	0.00584
CustomerId	RowNumber	0.004202
EstimatedSalary	RowNumber	-0.00599
HasCrCard	RowNumber	.0005987
IsActiveMember	RowNumber	0.012044
NumOfProducts	RowNumber	0.007246
RowNumber	RowNumber	1
Tenure	RowNumber	-0.00649
Age	Tenure	-0.01
Balance	Tenure	-0.01225
CreditScore	Tenure	.0008419
CustomerId	Tenure	-0.01488
EstimatedSalary	Tenure	0.007784
HasCrCard	Tenure	0.022583
IsActiveMember	Tenure	-0.02836
NumOfProducts	Tenure	0.013444
RowNumber	Tenure	-0.00649
Tenure	Tenure	1

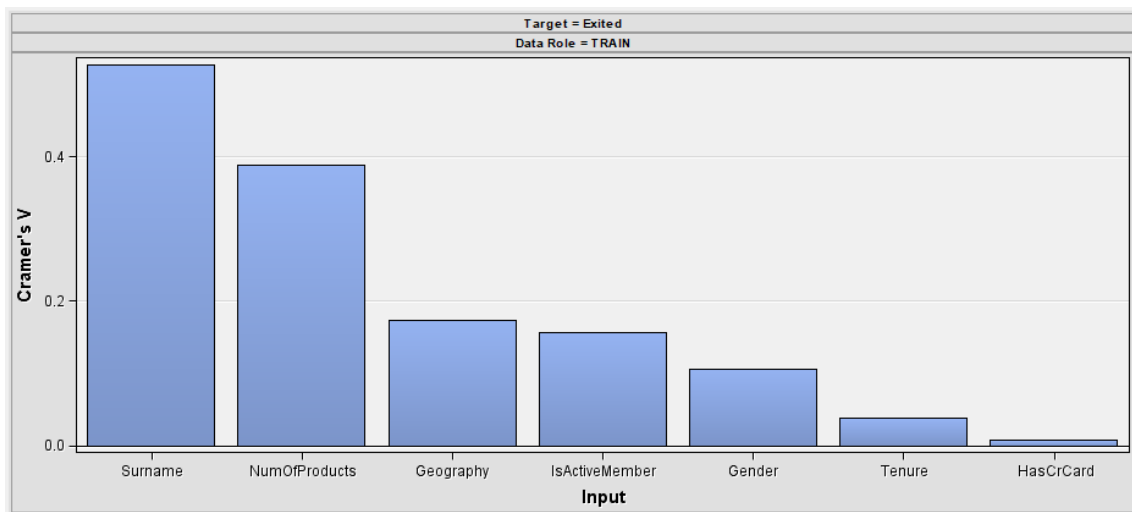
**Figure 5.16: Correction Table**

Based on the correlation matrix and correlation table visualized in Figure 5.15 and Figure 5.16, the variables do not have correlation values greater than 0.9. Therefore, no variables were removed.

For categorical variables, Cramer's V statistics were conducted to measure the association between categorical variables. Figure 5.19 below shows the findings of association between target variable (exited) and nominal variables.



**Figure 5.17: correction for target variable**



**Figure 5.18: Cramer's V statistics with respect to target variable**

In Cramer's V measure, the coefficient ranges from 0 to 0.6, where 0 indicates no association and 1 indicates a perfect association between the variables. A coefficient of 0.1 was used as a threshold to indicate that there is a relationship between two variables.

Based on Figure 5.17 and Figure 5.19, age and balance showed a perfect association. the volume name of hascr card , isactivemember showed a perfect association whereas GENDER showed almost no association. By taking a threshold of 0.1, age and balance were the only variables that showed association with target variable (excited).



## 5.2.5 Interesting Visualization

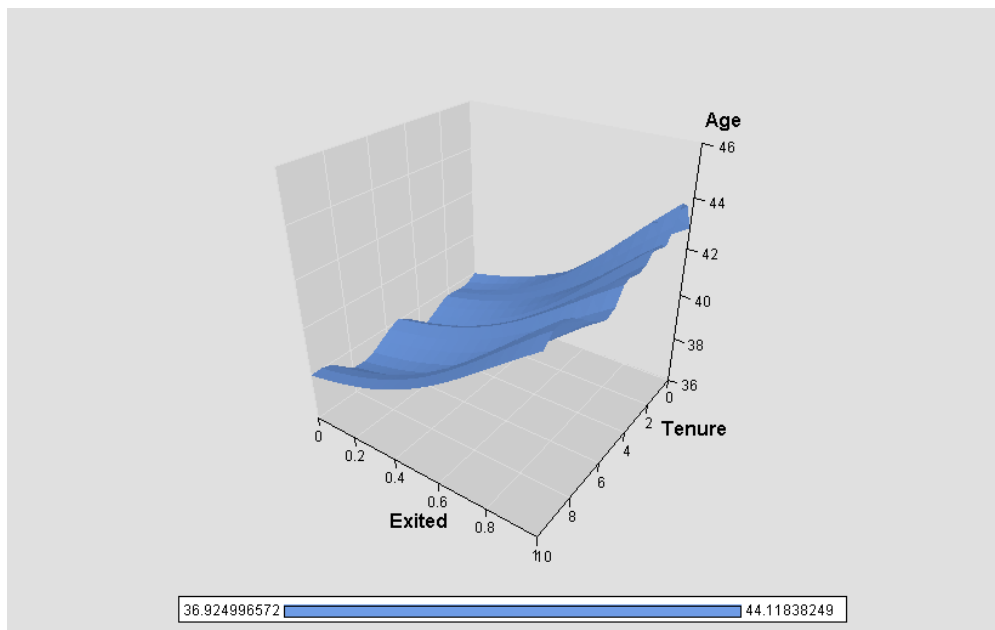


Figure 5.19 Finding of Relationship within Exited, Tenure and Age

## 5.3 Model

### 5.3.1 Modifying and Correcting Source Data

Quality issues are identified for this bank customer dataset. Data quality is assessed to identify any issues with the source data. This may include missing data, incorrect data types, data inconsistencies, and outliers. These include:

- (1) Cleaning and transforming data: Data cleaning and transformation techniques are used to correct any identified problems. This may involve removing duplicate data, filling in missing data, converting data types, and normalizing data values.
- (2) Compensating for missing data: If missing data is found, use imputation techniques to estimate missing values. This may involve average imputation, regression imputation, or other methods.
- (3) Correcting errors: If any errors are found in the source data, correct them as needed. This may involve manual correction or automatic correction using algorithms or rules.
- (4) Validate changes to the data: Once changes have been made, validate these changes to ensure that they do not introduce new problems or inconsistencies in the data.

By following these steps, it is possible to ensure that bank customer data is of high quality and suitable for subsequent modeling and analysis. This will improve the accuracy and validity of any data mining models developed.

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
CustomerId	INPUT	15690941	71936.19	10000	0	15565701	15690733	15815690	0.001149	-1.19584
EstimatedSalary	INPUT	100090.2	57510.49	10000	0	11.58	100187.4	199992.5	0.002085	-1.18152
Exited	INPUT	0.2037	0.402769	10000	0	0	0	1	1.471611	0.165671
HasCrCard	INPUT	0.7055	0.45584	10000	0	0	1	1	-0.90181	-1.18697
IsActiveMember	INPUT	0.5151	0.499797	10000	0	0	1	1	-0.06044	-1.99675
NumOfProducts	INPUT	1.5302	0.581654	10000	0	1	1	4	0.745568	0.582981
RowNumber	INPUT	5000.5	2886.896	10000	0	1	5000	10000	0	-1.2
Tenure	INPUT	5.0128	2.892174	10000	0	0	5	10	0.010991	-1.16523

Figure 5.20 check missing data

After the data exploration section, re-check the data for import. It was found that there was no missing data. This indicates that no data cleaning or other imputation methods are required to fill in the data. Referring to figure 5.20.

Name	Role	Level
Age	Input	Interval
Balance	Input	Interval
CreditScore	Input	Interval
CustomerId	Input	Interval
EstimatedSalary	Input	Interval
Exited	Target	Binary
Gender	Input	Binary
Geography	Input	Nominal
HasCrCard	Input	Binary
IsActiveMember	Input	Binary
NumOfProducts	Input	Nominal
RowNumber	Input	Interval
Surname	Input	Nominal
Tenure	Input	Nominal

Figure 5.21 Data Roles

The format of most of the data was checked to meet the format requirements for subsequent modeling, and only "Exited" needed to be converted to target values for subsequent decision tree modeling. Referring to figure 5.21.

### 5.3.2 Examining Exported Data

The processed data is exported, selected variables with high correlation and checked for errors, and then prepared for the decision tree model.

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
CustomerId	INPUT	15690941	71936.19	10000	0	15565701	15690733	15815690	0.001149	-1.19584
EstimatedSalary	INPUT	100090.2	57510.49	10000	0	11.58	100187.4	199992.5	0.002085	-1.18152
HasCrCard	INPUT	0.7055	0.45584	10000	0	0	1	1	-0.90181	-1.18697
RowNumber	INPUT	5000.5	2886.896	10000	0	1	5000	10000	0	-1.2

Figure 5.21 Interval Variable Summary Statistics of Input

Data	Variable		Level	Frequency	
	Role	Name		Count	Percent
TRAIN	Exited	TARGET	0	7963	79.63
TRAIN	Exited	TARGET	1	2037	20.37

**Figure 5.22 Interval Variable Summary Statistics of Target**

From Figures 5.21 and 5.22, it can be seen that there is no missing data in all data of the input and target values. And it is easy to see from the two graphs that all the data is in order and meets the conditions for building the model.

### 5.3.3 Creating Training and Validation Data

Exported Data	...
Notes	...
<b>Train</b>	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
<input checked="" type="checkbox"/> Data Set Allocation	
Training	70.0
Validation	30.0
Test	0.0
<b>Report</b>	
Interval Targets	Yes
Class Targets	Yes
<b>Status</b>	
Create Time	5/11/23 6:03 AM

**Figure 5.23 Setting the training and validation**

Partition Summary		
Type	Data Set	Number of Observations
DATA	EMWS1.Smpl_DATA	1000
TRAIN	EMWS1.Part_TRAIN	699
VALIDATE	EMWS1.Part_VALIDATE	301

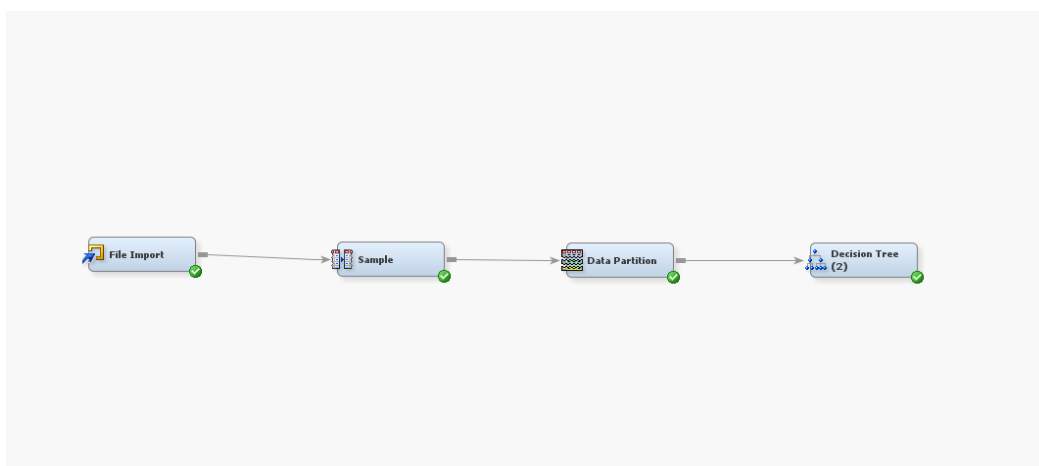
**Figure 5.24 Partition Summary**

Summary Statistics for Class Targets					
Data=DATA					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Exited	0	0	796	79.6	
Exited	1	1	204	20.4	
Data=TRAIN					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Exited	0	0	557	79.6853	
Exited	1	1	142	20.3147	
Data=VALIDATE					
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Exited	0	0	239	79.4020	
Exited	1	1	62	20.5980	

**Figure 5.25 Summary Statistics for Class Targets**

As shown in Figures 5.23, 5.24, and 5.25, this study divided the dataset into a training set (70%) and a validation set (30%). All aggregated data meet the criteria for building the model, including ensuring that the data represent the population under study, that there are no missing values or outlier, and that the variables used in the model are relevant and meaningful. Overall, the text indicates that the data partitioning scheme used in this study is appropriate and the data meets the necessary criteria for establishing a predictive model.

### 5.3.4 Constructing a Decision Tree Predictive Model



**Figure 5.26 Constructing a Decision Tree Predictive Model**

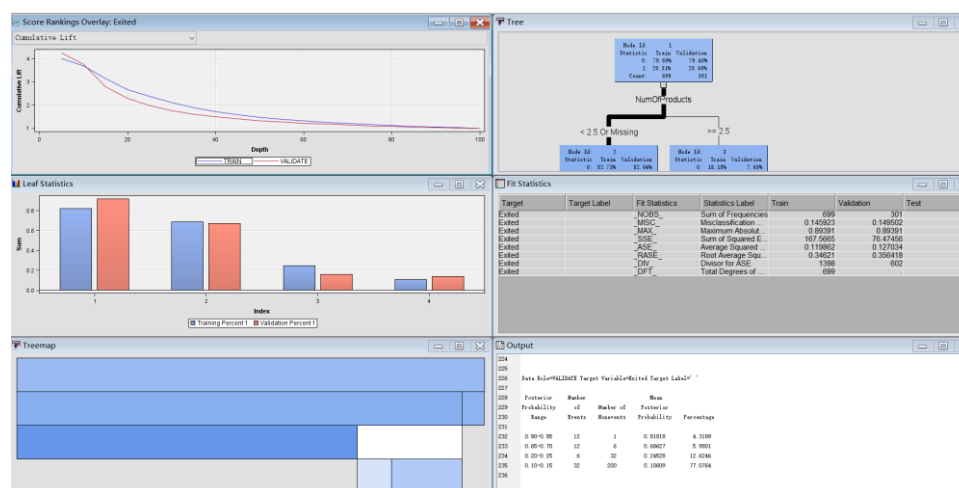
As shown in Figure 5.26, the construction of the model includes a total of 4 stages. The first stage is to introduce data, the second stage is to explore data for samples, the third step is to partition data, and the last step is to run the model.

Name	Use	Report	Role	Level
Age	Default	No	Input	Interval
Balance	Default	No	Input	Interval
CreditScore	Default	No	Input	Interval
CustomerId	Default	No	Input	Interval
EstimatedSalary	Default	No	Input	Interval
Exited	Yes	No	Target	Binary
Gender	Default	No	Input	Binary
Geography	Default	No	Input	Nominal
HasCrCard	Default	No	Input	Interval
IsActiveMember	Default	No	Input	Binary
NumOfProducts	Default	No	Input	Nominal
RowNumber	Default	No	Input	Interval
Surname	Default	No	Input	Nominal
Tenure	Default	No	Input	Nominal
_dataobs_		No	ID	Interval

**Figure 5.27 Constructing a Decision Tree Predictive Model's data**

As shown in Figure 5.27, various distributions of the data used in this model construction are included, which included 6 intervals, 4 normal, and 3 binaries, providing conditions for data prediction.

### 5.3.5 Assessing a Decision Tree



**Figure 5.28 outcome of Decision Tree**

# Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
NumOfProducts		1	1.0000	1.0000	1.0000
Age		1	0.8548	0.3882	0.4542
IsActiveMember		1	0.7149	0.6747	0.9438

# Tree Leaf Report

Figure 5.29 Variable Importance

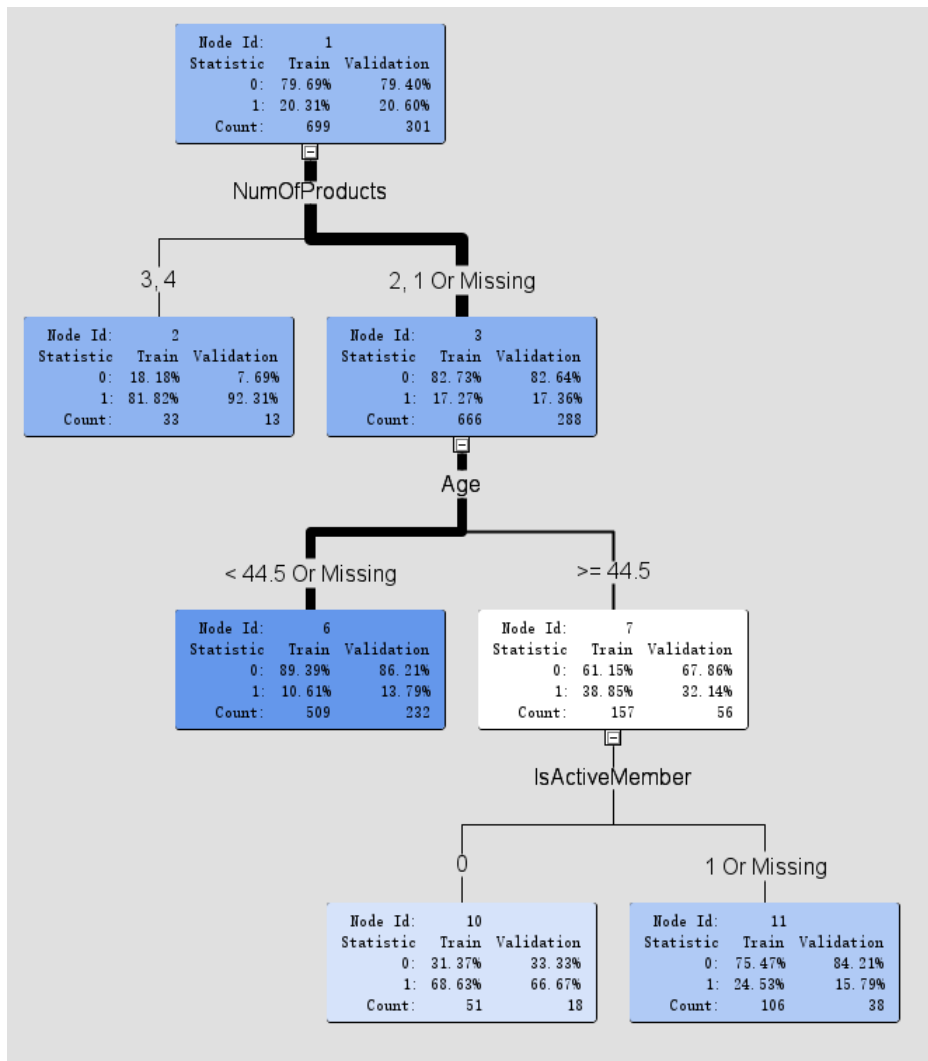


Figure 5.30 Tree of Predict in Decision Tree

#### Tree Leaf Report

Node Id	Depth	Training Observations	Training Percent	Validation Observations	Validation
			1		Percent 1
4	2	509	0.11	232	0.14
11	3	106	0.25	38	0.16
10	3	51	0.69	18	0.67
3	1	33	0.82	13	0.92

**Figure 5.31 Tree Leaf Report**

#### Fit Statistics

Target=Exited Target Label= ' '

Fit		Train	Validation
Statistics	Statistics Label		
_NOBS_	Sum of Frequencies	699.00	301.000
_MISC_	Misclassification Rate	0.15	0.150
_MAX_	Maximum Absolute Error	0.89	0.894
_SSE_	Sum of Squared Errors	167.57	76.475
_ASE_	Average Squared Error	0.12	0.127
_RASE_	Root Average Squared Error	0.35	0.356
_DIV_	Divisor for ASE	1398.00	602.000
_DFT_	Total Degrees of Freedom	699.00	.

**Figure 5.32 Fit Statistics**

Based on the decision tree report, it can be concluded that the Age variable has the highest importance in predicting customer churn, followed by the NumOfProducts and IsActiveMember variables. The model predicts that a customer is likely to exit if they are older, have multiple products, and are not active members. The balance and geography variables were also found to be important but to a lesser extent. Referring to Figure 5.30.

The tree leaf report shows that the model performed well on the training data, with low average squared errors, but slightly worse on the validation data, which is to be expected. However, the overall root average squared error is still relatively low, indicating that the model is reasonably accurate. Referring to Figure 5.31.

The fit statistics section shows the sum of squared errors and root mean squared error for the training, validation, and test datasets. The model's performance is relatively consistent across the different datasets, with an average squared error (ASE) of training is around 0.12 and root average squared error (RASE) of training is around 0.35. The overall root average

squared error is still relatively low, indicating that the model is reasonably accurate. Referring to Figure 5.32.

Here is the explanation from fit statistics:

- "NOBS"- the total number of observations in the data.
- "MAX"-the maximum absolute error of the model.
- "SSE"- the sum of squared errors of the model.
- "ASE"- the average squared error of the model.
- "RASE"- the root average squared error of the model.
- "DIV"- the divisor used in the calculation of the ASE.
- "DFT"- the total degrees of freedom of the model.

Overall, the decision tree model is a useful tool for predicting customer churn, and the report provides valuable insights into the key variables that drive customer behavior. However, as with any model, it is important to continue monitoring its performance and adjusting it as necessary to ensure that it remains accurate and relevant over time. (The full result is in the appendix)

## 6. Conclusion

Initially, the dataset contains 9 interval variables and 5 nominal variables. After manually revising the metadata, the output is shown in Table 6.1.

Role	Type of Variable	Count
Input	Binary	3
Input	Interval	6
Input	Nominal	4
Target	Binary	1

**Table 6.1: Revised metadata**

Based on the visualizations displayed in session 5.2 *Explore*, both objective 1 and objective 2 are achieved. The objectives are proven with key findings as listed in Table 6.3 below.

Objective	Key Findings
1. Understand the basic situation of bank customers.	From Figures 5.10 to 5.12, as well as Tables 5.1 to 5.3, the data displays that possible reasons for customer churn include age, tenure, and credit card usability. These factors are all related to some extent, but they cannot be seen from the surface.



<p>2. Determine the reason for customer churn based on the relationship between all attributes of the bank and all factors of the customer.</p>	<p>From the univariate analysis, multivariate analysis, and bivariate analysis in 5.2, it is found in a more detailed way, that most critical factors leading to customer churn are age, tenor, gender, and credit. Older and younger individuals tend to have fewer cases of attrition, while those in the middle age group tend to have more. For the customer tenure/age in banking, the longer they stay, the less likely they are to lose, while for those with shorter periods, they are more likely to lose. Furthermore, it is also found that women are more likely to have a loss.</p>
<p>3. Based on the analysis of all factors in the bank, predict whether customers will be lost.</p>	<p>By using the Decision Tree model, it can be concluded that the model can predict the churning customer accurately. The model reported that age variables have the highest importance in predicting customer churn, followed by NumOfProducts and IsActiveMember variables. Furthermore, if customers are older, have multiple products, and are not active members, they are more likely to churn. Balancing and geographical variables are also important, but to a lesser extent in term of churning prediction.</p>

---

# Appendix

\*-----\*

\* Training Output

\*-----\*

## Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
INPUT	BINARY	2
INPUT	INTERVAL	7
INPUT	NOMINAL	4
TARGET	BINARY	1

## Model Events

Target	Event	Measurement Level	Number of Levels	Order	Label
Exited	1	BINARY	2	Descending	

## Predicted and decision variables

Type	Variable	Label
TARGET	Exited	
PREDICTED	P_Exited1	Predicted: Exited=1
RESIDUAL	R_Exited1	Residual: Exited=1
PREDICTED	P_Exited0	Predicted: Exited=0
RESIDUAL	R_Exited0	Residual: Exited=0
FROM	F_Exited	From: Exited
INTO	I_Exited	Into: Exited

## Variable Importance

Ratio of		Number		of	
Validation		Splitting		Validation	
Training				to	
Variable Name	Label	Rules	Importance	Importance	Importance
NumOfProducts		1	1.0000	1.0000	1.0000
Age		1	0.8548	0.3882	0.4542
IsActiveMember		1	0.7149	0.6747	0.9438

## Tree Leaf Report

Node Id	Depth	Training Observations	Training Percent	Validation Observations	Validation Percent
			1		1
6	2	509	0.11	232	0.14
11	3	106	0.25	38	0.16
10	3	51	0.69	18	0.67
2	1	33	0.82	13	0.92

## Fit Statistics

Target=Exited Target Label=' '

Fit		Train	Validation
Statistics	Statistics Label		
_NOBS_	Sum of Frequencies	699.00	301.000
_MISC_	Misclassification Rate	0.15	0.150
_MAX_	Maximum Absolute Error	0.89	0.894

_SSE_	Sum of Squared Errors	167.57	76.475
_ASE_	Average Squared Error	0.12	0.127
_RASE_	Root Average Squared Error	0.35	0.356
_DIV_	Divisor for ASE	1398.00	602.000
_DFT_	Total Degrees of Freedom	699.00	.

### Classification Table

Data Role=TRAIN Target Variable=Exited Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	86.9919	96.0503	535	76.5379
1	0	13.0081	56.3380	80	11.4449
0	1	26.1905	3.9497	22	3.1474
1	1	73.8095	43.6620	62	8.8698

Data Role=VALIDATE Target Variable=Exited Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	85.9259	97.0711	232	77.0764
1	0	14.0741	61.2903	38	12.6246
0	1	22.5806	2.9289	7	2.3256
1	1	77.4194	38.7097	24	7.9734

### Event Classification Table

Data Role=TRAIN Target=Exited Target Label=' '

False	True	False	True	
Negative	Negative	Positive	Positive	
80	535		22	62

Data Role=VALIDATE Target=Exited Target Label=' '

False	True	False	True	
Negative	Negative	Positive	Positive	
38	232	7	24	

## Assessment Score Rankings

Data Role=TRAIN Target Variable=Exited Target Label=' '

Mean			Cumulative	%	Cumulative	Number of
Posterior	Gain	Lift	Lift	Response	% Response	Observations
Depth						
Probability						
5	299.042	3.99042	3.99042	81.0644	81.0644	35
0.81064						
10	268.432	3.37821	3.68432	68.6275	74.8459	35
0.68627						
15	214.812	2.07573	3.14812	42.1680	63.9533	35
0.42168						
20	166.295	1.20741	2.66295	24.5283	54.0970	35
0.24528						
25	137.184	1.20741	2.37184	24.5283	48.1833	35
0.24528						
30	111.251	0.81588	2.11251	16.5744	42.9151	35
0.16574						
35	88.533	0.52223	1.88533	10.6090	38.3000	35
0.10609						
40	71.494	0.52223	1.71494	10.6090	34.8386	35
0.10609						
45	58.242	0.52223	1.58242	10.6090	32.1464	35
0.10609						
50	47.640	0.52223	1.47640	10.6090	29.9927	35
0.10609						
55	38.966	0.52223	1.38966	10.6090	28.2306	35
0.10609						
60	31.737	0.52223	1.31737	10.6090	26.7621	35
0.10609						
65	25.621	0.52223	1.25621	10.6090	25.5195	35
0.10609						
70	20.378	0.52223	1.20378	10.6090	24.4545	35
0.10609						
75	15.835	0.52223	1.15835	10.6090	23.5315	35
0.10609						
80	11.859	0.52223	1.11859	10.6090	22.7238	35
0.10609						

85 0.10609	8.351	0.52223	1.08351	10.6090	22.0112	35
90 0.10609	5.233	0.52223	1.05233	10.6090	21.3777	35
95 0.10609	2.443	0.52223	1.02443	10.6090	20.8110	35
100 0.10609	0.000	0.52223	1.00000	10.6090	20.3147	34

Data Role=VALIDATE Target Variable=Exited Target Label=' '

Mean Posterior Depth Probability	Gain	Lift	Cumulative Lift	% Response	Cumulative % Response	Number of Observations
5 0.79345	324.798	4.24798	4.24798	87.5000	87.5000	16
10 0.68627	275.858	3.23656	3.75858	66.6667	77.4194	15
15 0.24528	178.292	0.76655	2.78292	15.7895	57.3227	15
20 0.24528	128.709	0.76655	2.28709	15.7895	47.1096	15
25 0.18033	97.806	0.72132	1.97806	14.8578	40.7441	15
30 0.10609	76.239	0.66963	1.76239	13.7931	36.3016	15
35 0.10609	60.775	0.66963	1.60775	13.7931	33.1165	15
40 0.10609	49.146	0.66963	1.49146	13.7931	30.7210	15
45 0.10609	40.081	0.66963	1.40081	13.7931	28.8540	15
50 0.10609	32.818	0.66963	1.32818	13.7931	27.3578	15
55 0.10609	26.867	0.66963	1.26867	13.7931	26.1321	15
60 0.10609	21.903	0.66963	1.21903	13.7931	25.1095	15
65	17.698	0.66963	1.17698	13.7931	24.2435	15

0.10609						
70	14.091	0.66963	1.14091	13.7931	23.5006	15
0.10609						
75	10.964	0.66963	1.10964	13.7931	22.8563	15
0.10609						
80	8.225	0.66963	1.08225	13.7931	22.2922	15
0.10609						
85	5.807	0.66963	1.05807	13.7931	21.7942	15
0.10609						
90	3.657	0.66963	1.03657	13.7931	21.3513	15
0.10609						
95	1.733	0.66963	1.01733	13.7931	20.9549	15
0.10609						
100	0.000	0.66963	1.00000	13.7931	20.5980	15
0.10609						

#### Assessment Score Distribution

Data Role=TRAIN Target Variable=Exited Target Label=' '

Posterior	Number		Mean	
Probability	of	Number of	Posterior	
Range	Events	Nonevents	Probability	Percentage
0.80-0.85	27	6	0.81818	4.7210
0.65-0.70	35	16	0.68627	7.2961
0.20-0.25	26	80	0.24528	15.1645
0.10-0.15	54	455	0.10609	72.8183

Data Role=VALIDATE Target Variable=Exited Target Label=' '

Posterior	Number		Mean	
Probability	of	Number of	Posterior	
Range	Events	Nonevents	Probability	Percentage
0.80-0.85	12	1	0.81818	4.3189
0.65-0.70	12	6	0.68627	5.9801
0.20-0.25	6	32	0.24528	12.6246
0.10-0.15	32	200	0.10609	77.0764