WQD7005 DATA MINING
2022/2023 SEMESTER 2

CREDIT SCORE CLASSIFICATION

GROUP ASSIGNMENT 2

YouTube link: https://youtu.be/YrV0HPwoJUg

DM G1(RL) – Group 1		
No.	**MATRIC ID**	**NAME**
1	22056764	Devayani A/P Balkrishnan
2	S2195613	Li Xin Qi
3	S2155659	Lim Yu Xuan
4	S2192763	Navaneeta A/P P Shanmugam

Table of Contents

1 Project Background

In this modern age, having a loan provides individual the access to cash for purchasing and gain ownership on assets instantly. In exchange, the individual will bear the financial commitment with a repayment of principal and interest to the lender over the years. However, giving out loan is a risky business to the lender as it imposes credit risk when borrowers fail to repay their loans, resulting in loan default. Thus, credit scoring plays a vital role in allowing lenders to assess the risk prior to approving any loan application, especially in recent years with the dramatic growth in consumer credit.

Credit scoring is a score that represents the likelihood of an individual repaying loans. It is a formal statistical method used for differentiating credit applicants into 'good' and 'bad' risk classes (Li et al., 2004).

From a borrower's standpoint, building a good credit scoring will not only allow their loan application to get approved easier as compared to a riskier borrower, but it also qualifies borrowers for lower interest rates. Having a good credit score reflects the individual as a trustworthy borrower and thus would enjoy more offers and rewards. Without a good credit score, it makes individuals vulnerable in getting funds for unexpected financial emergencies that would likely experience greater financial distress.

A few decades ago, financial institutions granted credit to individuals based on human judgement on loan default land experience in the industry (Hand & Henley, 1997). However, with the emergence of underwriting technologies and the economic pressure from the surging credit demand, financial institutions started to develop and leverage on credit scoring as a statistical model for making credit granting decisions. This shows a transition of assessment and pricing credit risk that are more data driven. With credit scoring classification, it optimizes financial institutions the ability to screen high-risk borrowers and target more generous facilities to lower-risk borrowers. These in turn enable financial institutions in minimizing default cost while increasing sales probability and overall expected profits (Einav et al., 2013).

2 Analysis Goal

In this project, we aim to predict the credit score of customers by leveraging their credit-related information, thus enabling finance companies to evaluate the customer's credit risk level. The analysis will be performed and visualized based on data mining steps developed by SAS institute named SEMMA, which follows with the process sequence of Sample, Explore, Modify, Model and Assess.

Therefore, the objectives of this project are:
1. To explore the key factors that have significant impact on an individual's credit score.
2. To build a credit score classification model.
3. To assess various models and determine the best model in classifying customers' credit score.

3 Accessing and Assaying Prepared Data

3.1 Data Acquisition

The datasets were acquired from https://www.kaggle.com/datasets/parisrohan/credit-score-classification?resource=download. It was uploaded by Rohan Paris in July 2022 under the title

‘Credit score classification’. The purpose of the datasets was to create a machine learning model that can categorize the credit score when given information about a person's credit history. Only the training dataset (train.csv) was chosen to be used in the model creation as it has the column denoting the credit score. The testing dataset (test.csv) will be used during the testing phase of our models.

3.2 Analysis of Data (Number of attributes/rows, anomalies, target features)

There are 28 attributes in the training dataset and 27 in the test dataset, with the credit score column being the difference. The training dataset has 100,000 rows while the testing dataset has 50,000 rows. The training dataset spans from January to August for each customer while the testing dataset covers the rest of the year, from September to December. Both datasets have a total of 12,500 unique customers and their data. The target feature of the data is the credit score which has three classifications, namely Poor, Standard, or Good.

3.3 Column Metadata

Table 3.3.1 below describes the columns in the datasets.

Column	Description
ID	Represents a unique identification of an entry
Customer_ID	Represents a unique identification of a person
Month	Represents the month of the year
Name	Represents the name of a person
Age	Represents the age of the person
SSN	Represents the social security number of a person
Occupation	Represents the occupation of the person
Annual_Income	Represents the annual income of the person
Monthly_Inhand_Salary	Represents the monthly base salary of a person
Num_Bank_Accounts	Represents the number of bank accounts a person holds
Num_Credit_Card	Represents the number of other credit cards held by a person
Interest_Rate	Represents the interest rate on credit card
Num_of_Loan	Represents the number of loans taken from the bank
Type_of_Loan	Represents the types of loan taken by a person
Delay_from_due_date	Represents the average number of days delayed from the payment date
Num_of_Delayed_Paymen	Represents the average number of payments delayed by a person
Changed_Credit_Limit	Represents the percentage change in credit card limit
Num_Credit_Inquiries	Represents the number of credit card inquiries
Credit_Mix	Represents the classification of the mix of credits
Outstanding_Debt	Represents the remaining debt to be paid (in USD)
Credit_Utilization_Ratio	Represents the utilization ratio of credit card
Credit_History_Age	Represents the age of credit history of the person
Payment_of_Min_Amount	Represents whether only the minimum amount was paid by the person

Total_EMI_per_month	Represents the monthly EMI payments (in USD)
Amount_invested_monthly	Represents the monthly amount invested by the customer (in USD)
Payment_Behaviour	Represents the payment behavior of the customer (in USD)
Monthly_Balance	Represents the monthly balance amount of the customer (in USD)
Credit_Score	Represents the bracket of credit score (Poor, Standard, Good)

Table 3.3.1 Column Metadata

4 Methodology: SEMMA

SAS Institute demonstrates data mining using the SEMMA methodology to uncover hidden information from massive amounts of data and improve business decision-making. The SEMMA process is well-known in handling business problems across various sectors as it aids businesses to gain a competitive edge, enhance performance, and provide customers with more beneficial services. Figure 4.1.1 below illustrates each phase in SEMMA in detail.

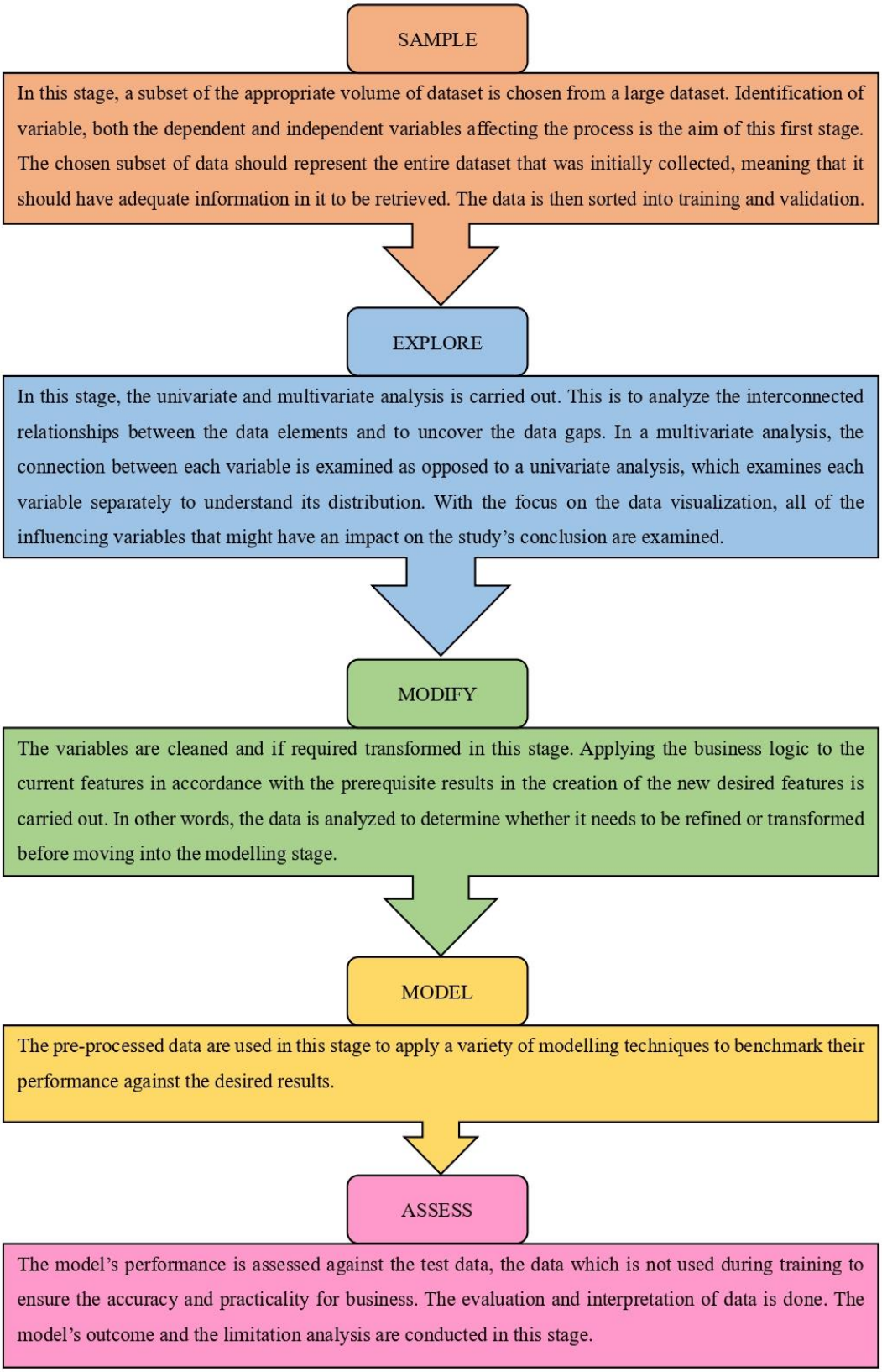


Figure 4.1.1 SEMMA

4.1 Project Setup

Setting up a new SAS Enterprise Miner project is required prior to applying the SEMMA process on the dataset. The step-by-step details include creating project, diagram, and library, and importing dataset are shown and described in the diagrams below.

1. A new SAS Enterprise Miner project is created by setting the project name "Credit Score Classification" and location of the project (Figures 12.1.1 - 12.1.4).
2. A SAS Enterprise Miner diagram workspace named "SE" is created to contain and display the steps involved in this project (Figures 12.1.5 - 12.1.6).
3. The training dataset, "train.csv" is uploaded to SAS server via SAS Studio located in SAS On Demand for Academics by selecting the "Upload" button to browse from local files. The uploaded dataset is ready for access in SAS Enterprise Miner to perform sampling and exploring techniques (Figures 12.1.7 - 12.1.11).
4. The File Import node located on the Sample tab of the toolbar is dragged to the SE diagram workspace to import the dataset from SAS server to SAS Enterprise Miner (Figures 12.1.12 - 12.1.18).
5. After importing the dataset, the Save Data node in the Utility tab is dragged to the SE diagram, then connected to the File Import node to save the dataset as a SAS table (Figures 12.1.19 - 12.1.23).
6. A library named "Dataset" is created with its path set as the location of the saved dataset to store the dataset (Figures 12.1.24 - 12.1.27).

5 SEMMA: SAMPLE

Once the project setup is completed, the dataset is ready for sampling process. The following describes the steps involved in the Sample stage, starting from creating and loading data sources to data sampling.

1. The data source is created via the Data Source Wizard by first setting the source as SAS Table (Figures 12.2.1 - 12.2.2).
2. The SAS table named "Em_save_train" located in the "Dataset" library is selected as our data source (Figures 12.2.3 - 12.2.6).
3. The information of the chosen data source like table name, number of variable and observations are displayed as shown in (Figure 12.2.7).
4. The advanced setting is used to modify the initial measurement levels and roles of the variables in the data source (Figure 12.2.8).
5. The figures below show the default setting (Figure 12.2.9) and modified setting (Figure 12.2.10) of the measurement levels and roles, respectively.

   Figure 5.1.1 below demonstrates and highlights the modifications applied to the column metadata, especially on the reclassification of variables role and level. The Customer ID and SSN variables are dropped as they are not required for exploring and modelling stage later.

Variables	Role	Level
Age	Input	Nominal
Amount_invested_monthly	Input	Nominal
Annual_Income	Input	Nominal
Changed_Credit_Limit	Input	Nominal
Credit_History_Age	Input	Nominal
Credit_Mix	Input	Nominal
Credit_Score	Input	Nominal
Credit_Utilization_Ratio	Input	Interval
Customer_ID	ID	Nominal
Delay_from_due_date	Input	Interval
ID	ID	Nominal
Interest_Rate	Input	Interval
Month	Input	Nominal
Monthly_Balance	Input	Interval
Monthly_Inhand_Salary	Input	Interval
Name	Rejected	Nominal
Num_Bank_Accounts	Input	Interval
Num_Credit_Card	Input	Interval
Num_Credit_Inquiries	Input	Interval
Num_of_Delayed_Payment	Rejected	Nominal
Num_of_Loan	Rejected	Nominal
Occupation	Input	Nominal
Outstanding_Debt	Rejected	Nominal
Payment_Behaviour	Input	Nominal
Payment_of_Min_Amount	Input	Nominal
SSN	Rejected	Nominal
Total_EMI_per_month	Input	Interval
Type_of_Loan	Text	Nominal

Variables	Role	Level
Age	Input	Nominal
Amount_invested_monthly	Input	Nominal
Annual_Income	Input	Ordinal
Changed_Credit_Limit	Input	Nominal
Credit_History_Age	Input	Nominal
Credit_Mix	Input	Ordinal
Credit_Score	Target	Ordinal
Credit_Utilization_Ratio	Input	Interval
Customer_ID	Input	Nominal
Delay_from_due_date	Input	Interval
ID	ID	Nominal
Interest_Rate	Input	Interval
Month	Input	Nominal
Monthly_Balance	Input	Interval
Monthly_Inhand_Salary	Input	Interval
Name	Rejected	Nominal
Num_Bank_Accounts	Input	Interval
Num_Credit_Card	Input	Interval
Num_Credit_Inquiries	Input	Interval
Num_of_Delayed_Payment	Input	Nominal
Num_of_Loan	Input	Nominal
Occupation	Input	Nominal
Outstanding_Debt	Input	Nominal
Payment_Behaviour	Input	Nominal
Payment_of_Min_Amount	Input	Nominal
SSN	Rejected	Nominal
Total_EMI_per_month	Input	Interval
Type_of_Loan	Input	Nominal

Figure 5.1.1 Reclassification of Variables Roles and Level

However, the level of the variables after reclassification still does not reflect the actual situation due to dirty and noisy data in the dataset. The variables level can only be classified accurately after the data cleaning process which will be applied at Modify stage in the next project phase. Figure 5.1.2 below demonstrates the correct level of variables of dataset.

Variables	Level
Age	Nominal
Amount_invested_monthly	Interval
Annual_Income	Interval
Changed_Credit_Limit	Interval
Credit_History_Age	Ordinal
Credit_Mix	Ordinal
Credit_Score	Ordinal
Credit_Utilization_Ratio	Interval
Customer_ID	Nominal
Delay_from_due_date	Interval
ID	Nominal
Interest_Rate	Interval
Month	Nominal
Monthly_Balance	Interval
Monthly_Inhand_Salary	Interval
Name	Nominal
Num_Bank_Accounts	Interval
Num_Credit_Card	Interval
Num_Credit_Inquiries	Interval
Num_of_Delayed_Payment	Interval
Num_of_Loan	Interval
Occupation	Nominal
Outstanding_Debt	Interval
Payment_Behaviour	Nominal
Payment_of_Min_Amount	Nominal
SSN	Nominal
Total_EMI_per_month	Interval
Type_of_Loan	Nominal

Figure 5.1.2 Correct Variables Level

6. The decision processing is not specified as illustrated in (Figure 12.2.11).
7. The sample dataset is not created via the Data Source Wizard (Figure 12.2.12), but instead the data sampling process is applied after defining the data source (where detailed discussion is provided at point 12).
8. No changes have been made to the name and role of the data source (Figure 12.2.13).
9. The summary of the data source is shown in (Figure 12.2.14). As observed, there are 16 nominal variables (1 ID, 13 Input, 2 Rejected), 9 interval variables with Input role and 3 ordinal variables (2 Input, 1 Target) in the dataset.
10. The data source named “EM_SAVE_TRAIN” is created successfully and dragged to the SE diagram (Figures 12.2.15 - 12.2.17).
11. The node is then renamed to “Training Data” and ready for data sampling (Figures 12.2.18 - 12.2.22).
12. Data sampling is performed by dragging the Sample node in the Sample tab (Figures 12.2.23 - 12.2.24) to the SE diagram and connecting it to the “Training Data” node. The stratified equal-sized sampling method is applied with a random seed 12345, and the size of the training dataset is set to 30% (Figures 12.2.25 - 12.2.29).

6 SEMMA: EXPLORE

In the Explore phase, data exploration is performed to identify patterns, trends, and relationship between variables, as well as discover anomalies and outliers in the dataset that require further investigation in later phase. Common Exploratory Data Analysis (EDA) techniques include summary statistics, univariate, bivariate and multivariate analysis are applied using the nodes under the Explore tab in SAS Enterprise Miner. The findings of the analyses are discussed in detail later in this section, alongside some screenshots of SAS Enterprise Miner.

6.1 Statistical Summary

To generate the summary statistics of the dataset, the StatExplore node is used by dragging the node to the SE diagram and connecting to the Sample node (Figures 12.3.1 - 12.3.4). Figure 6.1.1 below shows the summary statistics of the class variables while Figure 6.1.2 below shows the summary statistics of the interval variables.

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	Age	INPUT	513	0	25	2.95	28	2.94
TRAIN	Amount_invested_monthly	INPUT	513	19	__10000__	4.50		3.42
TRAIN	Annual_Income	INPUT	513	0	12613.56	0.61	75881.16	0.61
TRAIN	Changed_Credit_Limit	INPUT	513	0	_	3.11	5.71	0.86
TRAIN	Credit_History_Age	INPUT	404	0	NA	9.10	15 Years and 11 Months	0.53
TRAIN	Credit_Mix	INPUT	4	0	Good	32.16	Standard	29.42
TRAIN	Month	INPUT	8	0	June	12.72	July	12.65
TRAIN	Num_of_Delayed_Payment	INPUT	259	2103		7.01	10	5.18
TRAIN	Num_of_Loan	INPUT	159	0	3	14.66	2	14.36
TRAIN	Occupation	INPUT	16	0	_______	6.98	Scientist	6.53
TRAIN	Payment_Behaviour	INPUT	7	0	Low_spent_Small_value_payments	24.41	High_spent_Medium_value_payments	17.85
TRAIN	Payment_of_Min_Amount	INPUT	3	0	Yes	46.16	No	41.79
TRAIN	Type_of_Loan	INPUT	513	287		14.65	Mortgage Loan	2.40
TRAIN	Credit_Score	TARGET	3	0	Good	33.33	Poor	33.33

Figure 6.1.1 Class Variables Summary Statistics

Based on Figure 6.1.1, there are 14 class variables and out of these variables, 11 of them do not contain missing values, whereas the Amount invested monthly, Num of Delayed Payment, and Type of Loan variables have missing values. It is noticeable that the Age, Amount invested monthly, Annual Income, Changed Credit Limit, Credit History Age, Num of Delayed Payment, Num of Loan, and Type of Loan variables have extremely high number of levels due to dirty and noisy data, and large numbers of distinct values. The Age, Amount invested monthly, Annual Income, Changed Credit Limit, Credit History Age, Num of Delayed Payment, and Num of Loan variables are classified as class variables, which are supposed to be classified as interval variables. The Type of Loan variable possess high cardinality, and high dimensional data can result in a range of issues such as overfitting, low model performance and interpretability. These issues will be properly addressed and discussed further in the Modify phase.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Credit_Utilization_Ratio	INPUT	32.28744	5.135271	30000	0	20.25707	32.3135	50	0.041792	-0.93116
Delay_from_due_date	INPUT	20.0766	15.03858	30000	0	-5	16	67	1.024543	0.430127
Interest_Rate	INPUT	77.20173	494.2043	30000	0	1	12	5775	8.630536	77.49108
Monthly_Balance	INPUT	409.4905	221.5765	29624	376	0.503582	341.2423	1566.613	1.585072	2.823866
Monthly_Inhand_Salary	INPUT	4328.088	3296.71	25470	4530	303.6454	3173.971	15204.63	1.114643	0.521351
Num_Bank_Accounts	INPUT	16.77553	117.1833	30000	0	-1	5	1798	11.2261	133.0324
Num_Credit_Card	INPUT	22.56953	131.3644	30000	0	0	5	1499	8.394718	72.82476
Num_Credit_Inquiries	INPUT	28.20697	197.0245	29439	561	0	5	2594	9.652007	97.51068
Total_EMI_per_month	INPUT	1432.826	8401.09	30000	0	0	70.58768	82236	7.04676	51.34881

Figure 6.1.2 Interval Variable Summary Statistics

Based on Figure 6.1.2, the Monthly Balance, Monthly Inhand Salary, and Num Credit Inquiries variables have missing values, while the remaining 6 variables have no missing values. It is not reasonable that the Delay from due date and the Num Bank Accounts variables contain negative values (with minimum values -5 and -1 respectively), and the Num Bank Accounts, Num Credit Card and Interest rate variables have extremely high values (with maximum values 1798, 1499 and 5775 respectively).

Ob...	ID	Cu...	M...	Na...	Age	SSN	O...	Ann...	Mont...	Nu...	Num_...	Intere...	Nu...	Type...	Dela...	Nu...	Ch...	Nu...	Cred...	Outst...	Credi...	Cre...	Pay...	Total...	Amo...	Paym...	Monthl...	Credit_S...
1	0x1602	CU...	Jan...	Aaro...	23	821-0...	Scie...	19114...	1824...	3	4	34		Auto L...	37		11.27	4_		809.98	26.822...	22 Ye...	No	49.57...	80.415...	High_sp...	312.494...	Good
2	0x1603	CU...	Feb...	Aaro...	23	821-0...	Scie...	19114...		3	4	34		Auto L...	-1		11.27	4	Good	809.98	31.944...	NA	No	49.57...	118.28...	Low_sp...	284.629...	Good
3	0x1604	CU...	Mar...	Aaro...	-500	821-0...	Scie...	19114...		3	4	34		Auto L...	37		_	4	Good	809.98	28.609...	22 Ye...	No	49.57...	81.699...	Low_sp...	331.209...	Good
4	0x1605	CU...	Apri...	Aaro...	23	821-0...	Scie...	19114...		3	4	34		Auto L...	54		6.27	4	Good	809.98	31.377...	22 Ye...	No	49.57...	199.45...	Low_sp...	223.451...	Good
5	0x1606	CU...	May...	Aaro...	23	821-0...	Scie...	19114...	1824...	3	4	34		Auto L...	6		11.27	4	Good	809.98	24.797...	22 Ye...	No	49.57...	41.420...	High_sp...	341.48...	Good
6	0x1607	CU...	Jun...	Aaro...	23	821-0...	Scie...	19114...		3	4	34		Auto L...	84		9.27	4	Good	809.98	27.262...	22 Ye...	No	49.57...	62.430...	!@9#%8...	340.479...	Good
7	0x1608	CU...	July...	Aaro...	23	821-0...	Scie...	19114...	1824...	3	4	34		Auto L...	38_		11.27	4	Good	809.98	22.537...	22 Ye...	No	49.57...	178.34...	Low_sp...	244.565...	Good
8	0x1609	CU...	Aug...		23	#F%$...	Scie...	19114...	1824...	3	4	34		Auto L...	36		11.27	4	Good	809.98	23.93...	NA	No	49.57...	24.785...	High_sp...	358.124...	Standard
9	0x160e	CU...	Jan...	Rick ...	28_	004-0...	___	34847...	3037...	2	4	61		Credit-...	34		5.42	2	Good	605.03	24.464...	26 Ye...	No	18.81...	104.29...	Low_sp...	470.690...	Standard
10	0x160f	CU...	Feb...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	4	61		Credit-...	71		7.42	2	Good	605.03	38.550...	26 Ye...	No	18.81...	40.391...	High_sp...	484.591...	Good
11	0x1610	CU...	Mar...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	1385	61		Credit-...	3-1		5.42	2_		605.03	33.224...	26 Ye...	No	18.81...	58.515...	High_sp...	466.466...	Standard
12	0x1611	CU...	Apri...	Rick ...	28	004-0...	Tea...	34847...		2	4	61		Credit-...	33_		5.42	2	Good	605.03	39.182...	26 Ye...	No	18.81...	99.306...	Low_sp...	465.676...	Good
13	0x1612	CU...	May...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	4	61		Credit-...	31		6.42	2	Good	605.03	34.977...	26 Ye...	No	18.81...	130.11...	Low_sp...	444.867...	Good
14	0x1613	CU...	Jun...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	4	61		Credit-...	30		5.42	2	Good	605.03	33.38...	27 Ye...	No	18.81...	43.477...	High_sp...	481.505...	Good
15	0x1614	CU...	July...	Rick ...	28	004-0...	Tea...	34847...		2	4	61		Credit-...	34		5.42	2	Good	605.03	31.131...	27 Ye...	NM	18.81...	70.101...	High_sp...	464.880...	Good
16	0x1615	CU...	Aug...	Rick ...	28	004-0...	Tea...	34847...	3037...	2	4	61		Credit-...	34		5.42	2	Good	605.03	32.933...	27 Ye...	No	18.81...	218.90...	Low_sp...	356.078...	Good
17	0x161a	CU...	Jan...	Lang...	34	486-8...	____	14316...	121...	1	5	83		Auto L...	58		7.1	3	Good	1303.01	28.616...	17 Ye...	No	246.9...	168.41...	!@9#%8...	1043.31...	Good
18	0x161b	CU...	Feb...		34	486-8...	Eng...	14316...	121...	1	5	83		Auto L...	136		7.1	3	Good	1303.01	41.702...	17 Ye...	No	246.9...	232.86...	High_sp...	998.869...	Good
19	0x161c	CU...	Mar...	Lang...	34	486-8...	____	14316...		1	5	83		Auto L...	87		11.1		Good	1303.01	26.519...	17 Ye...	No	246.9...	__100...	High_sp...	715.741...	Good
20	0x161d	CU...	Apri...	Lang...	34	486-8...	Eng...	14316...	121...	1	5	83		Auto L...	85		9.1	3_		1303.01	39.501...	NA	No	246.9...	825.21...	Low_sp...	426.513...	Good
21	0x161e	CU...	May...	Lang...	34	486-8...	____	14316...	121...	1	5	83		Auto L...	105		7.1	3	Good	1303.01	31.376...	18 Ye...	No	246.9...	430.94...	Low_sp...	810.782...	Good
22	0x161f	CU...	Jun...	Lang...	34	486-8...	Eng...	14316...	121...	1	5	8967		Auto L...	86		7.1	3	Good	1303.01	39.783...	18 Ye...	No	246.9...	257.80...	High_sp...	963.921...	Good
23	0x1620	CU...	July...		34	486-8...	Eng...	14316...	121...	1	5	83		Auto L...	86		7.1	3	Good	1303.01	38.068...	18 Ye...	No	246.9...	263.17...	High_sp...	968.555...	Standard
24	0x1621	CU...	Aug...	Lang...	34	486-8...	Eng...	14316...	121...	1	5	83		Auto L...	86		7.1	3	Good	1303.01	38.374...	18 Ye...	No	246.9...	__100...	High_sp...	895.494...	Standard
25	0x1626	CU...	Jan...	Jaso...	54	072-3...	Entr...	30689...	2612...	2	5	41		Not Sp...	06		1.99	4	Good	632.46	26.544...	17 Ye...	No	16.41...	81.228...	Low_sp...	433.60...	Standard
26	0x1627	CU...	Feb...	Jaso...	54	072-3...	Entr...	30689...	2612...	2	5	41		Not Sp...	53		1.99	4	Good	632.46	35.279...	17 Ye...	No	16.41...	124.88...	Low_sp...	409.951...	Standard
27	0x1628	CU...	Mar...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	41		Not Sp...	39		1.99	4	Good	632.46	32.301...	17 Ye...	NM	16.41...	83.406...	High_sp...	411.427...	Standard
28	0x1629	CU...	Apri...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	41		Not Sp...	76		-2.01	4	Good	632.46	38.132...	17 Ye...	No	16.41...	272.33...	Low_sp...	262.499...	Standard
29	0x162a	CU...	May...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	41		Not Sp...	56		-1.01	4	Good	632.46	41.154...	17 Ye...	No	16.41...	__100...	Low_sp...	359.374...	Standard
30	0x162b	CU...	Jun...	Jaso...	55	#F%$...	___	30689...	2612...	2	5	41		Not Sp...	56		-3.01	4_		632.46	27.445...	17 Ye...	No	16.41...	84.952...	High_sp...	419.880...	Standard
31	0x162c	CU...	July...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	41		Not Sp...	5		1.99	4	Good	632.46	26.056...	17 Ye...	No	16.41...	71.283...	Low_sp...	443.549...	Standard
32	0x162d	CU...	Aug...	Jaso...	55	072-3...	Entr...	30689...	2612...	2	5	4-100		Not Sp...	49		1.99	4	Good	632.46	27.332...	17 Ye...	No	16.41...	125.61...	High_sp...	379.216...	Standard
33	0x1632	CU...	Jan...	Dee...	21	615-0...	Dev...	35547...	2853...	7	5	50		...	5		2.58	4	Stand...	943.86	39.797...	30 Ye...	Yes		0276.72...	!@9#%8...	288.605...	Standard
34	0x1633	CU...	Feb...	Dee...	21	615-0...	Dev...	35547...		7	5	50		...	9		2.58	4	Stand...	943.86	27.020...	30 Ye...	NM		074.443...	High_sp...	460.887...	Standard
35	0x1634	CU...	Mar...	Dee...	21	615-0...	Dev...	35547...	2853...	7	5	5-100		...	512		2.58	4	Stand...	943.86	23.462...	30 Ye...	Yes		0173.13...	Low_sp...	392.192...	Standard
36	0x1635	CU...	Apri...	Dee...	21	615-0...	Dev...	35547...	2853...	7	5	50		...	115		2.58	4_		943.86	28.924...	30 Ye...	Yes		096.785...	High_sp...	438.545...	Standard
37	0x1636	CU...	May...	Dee...	21	615-0...	Dev...	35547...	2853...	7	5	50		...	917		2.58	4_		943.86	41.776...	31 Ye...	Yes		062.723...	High_sp...	482.607...	Standard
38	0x1637	CU...	Jun...	Dee...	21	615-0...	Dev...	35547...		7	5	50_		...	515		2.58	4	Stand...	943.86	29.217...	31 Ye...	Yes		037.643...	High_sp...	497.687...	Standard

Figure 6.1.3 Overview of Data

Based on Figure 6.1.3, some of the numerical variables such as the Age variable consists of '_' character, which makes it unable to classify as interval variable, but instead character variable. In fact, the Credit History Age variable consists of missing values, but the missing values are labelled as 'NA' that is not recognized as missing values in SAS Enterprise Miner. Furthermore, the Payment of Min Amount column has a new label besides the 'Yes' and 'No' label, and the Credit Mix column also contains '_' character in the data. All missing values, noisy and inconsistent data will be cleaned and transformed in the next phase.

6.2 Graph Visualization

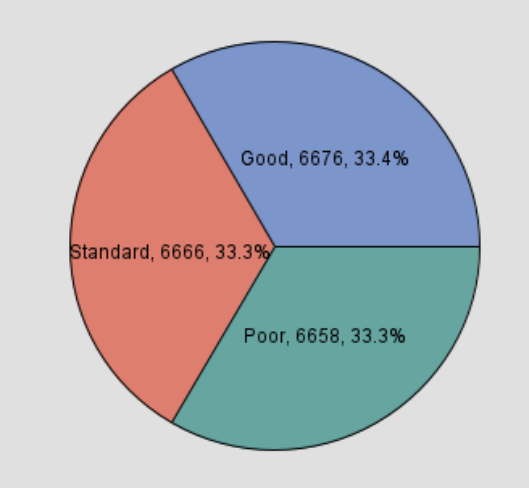
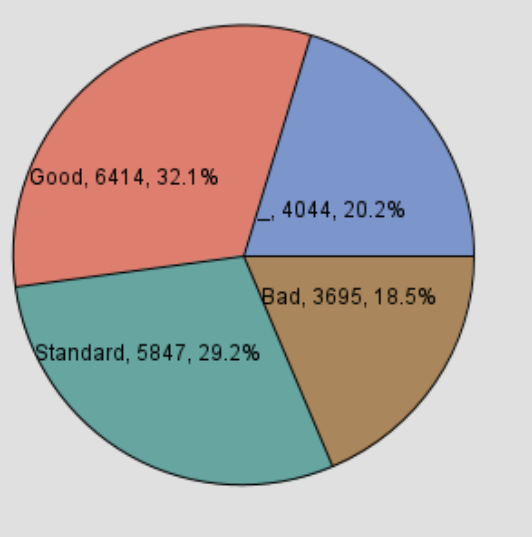
In the graph visualization, univariate, bivariate and multivariate analysis are performed. The data exploration will not reflect with the original dataset as the proportion is adjusted based on sampling

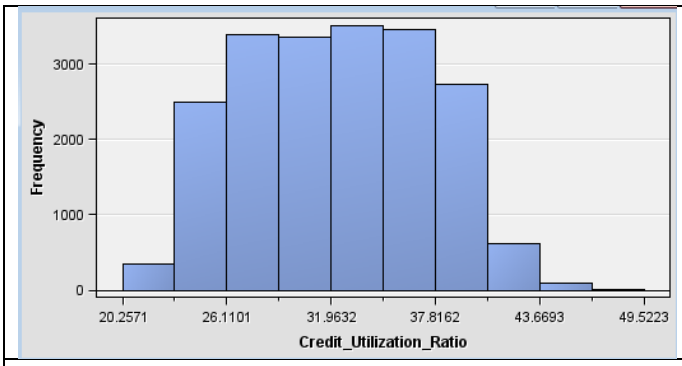
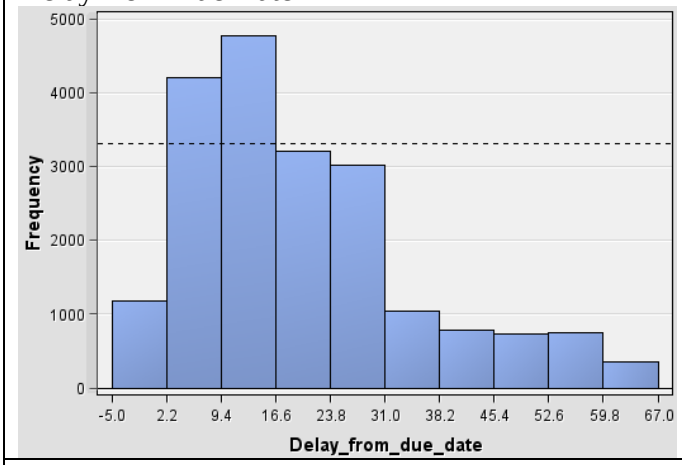
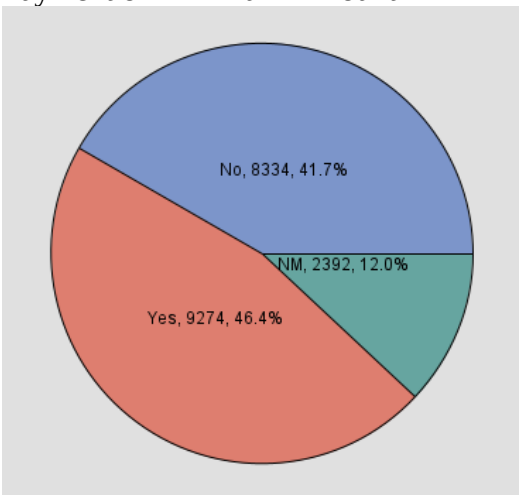
set. The GraphExplore, StatExplore, Multiplot, and Variable Clustering nodes are used by dragging the nodes to the SE diagram and connecting to the Sample node (shown in Appendix).

6.2.1 Univariate Analysis

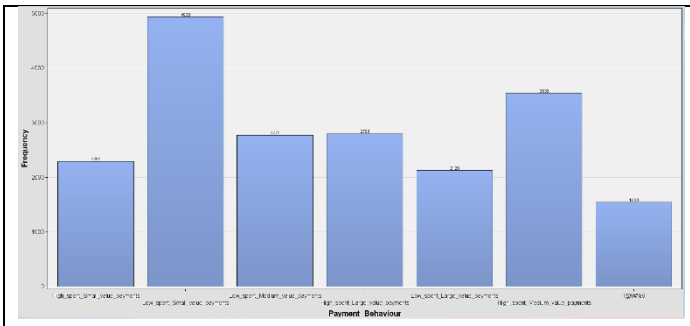
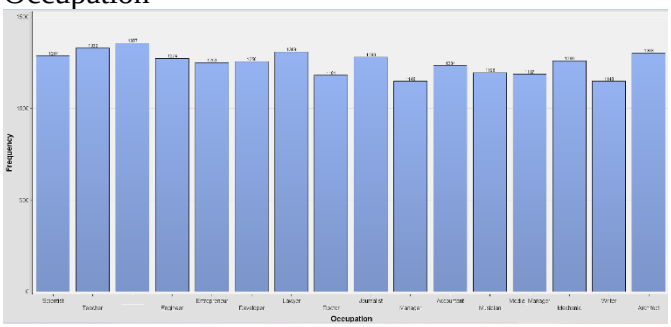
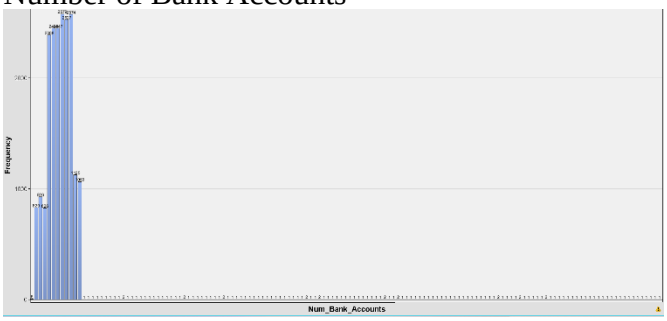
Univariate analysis is conducted by analyzing a single variable in the data using common techniques such as histogram, box plot etc. This analysis aims to describe the data, compare the spread of the variable as well as to find patterns and anomalies in the data.

In Figure 6.2.1 below, pie charts, bar graphs and histograms are presented to describe the patterns of each variable of the dataset.

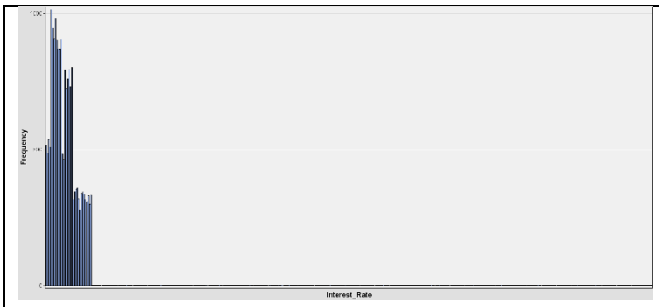
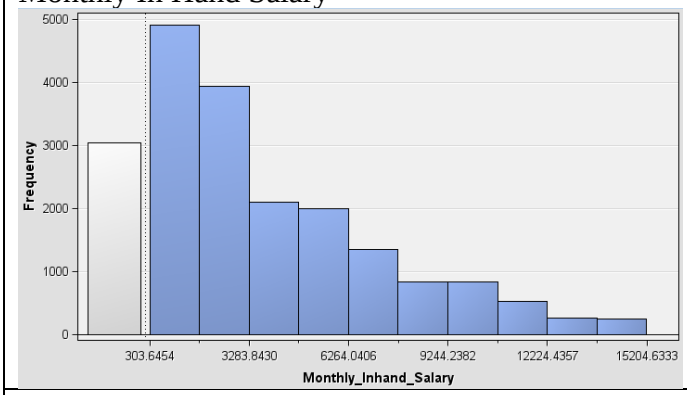
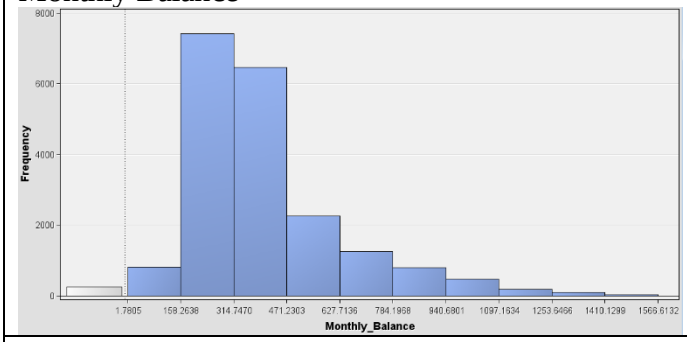
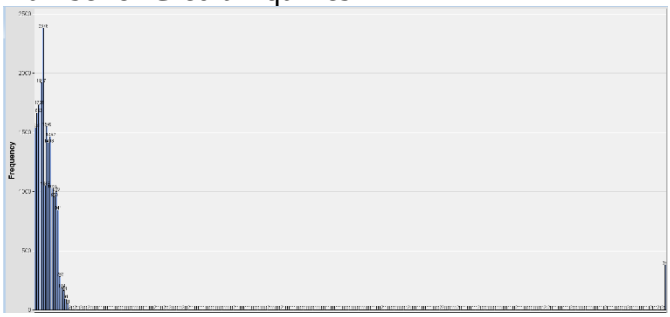
Variables	**Findings**
Credit Score 	Class target variable which uses stratified sampling and break down into 3 categories with different proportions and percentage: • Good – 6,676 (33%) • Standard – 6,666 (34%) • Poor – 6,658 (33%)
Credit Mix 	In credit mix variable, there are 4 categories with different proportions and percentage as below: • Standard – 5,847 (29%) • Good – 6,414 (32%) • Missing value – 4,044 (20%) • Bad – 3,695 (19%)
Credit Utilization Ratio	Based on credit utilization ratio variable, most borrowers' credit utilization ratio falls between 26.11 - 37.82, followed by

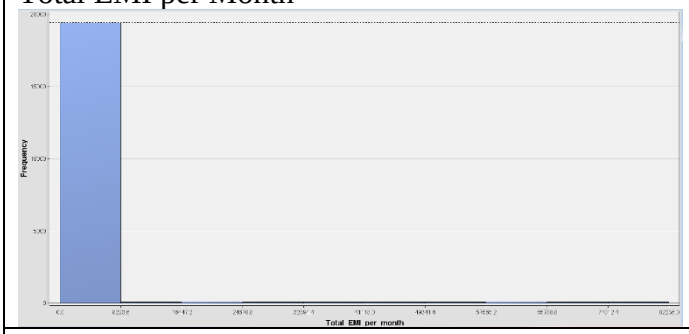
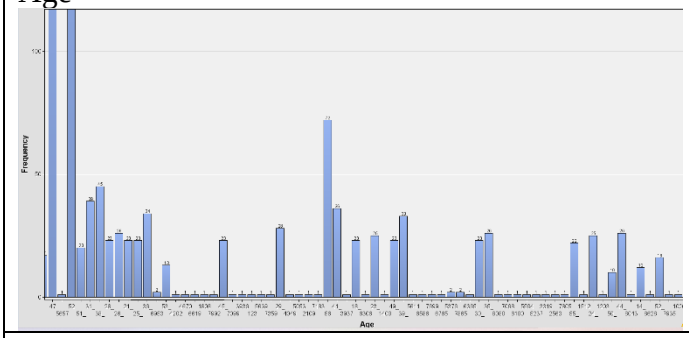
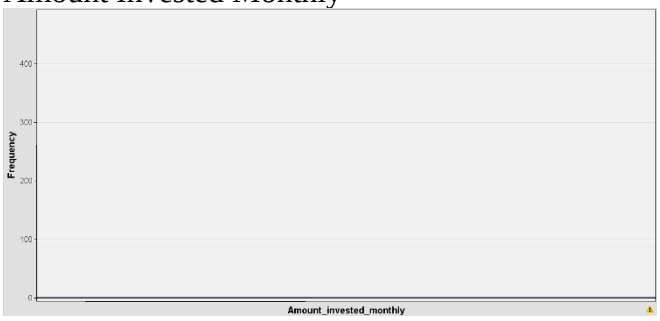
	the range between 23.18 - 26.11, and range between 37.82 - 40.74, with 20.23 as the lowest and 49.5 as the highest ratio.
Delay from Due Date 	Most borrowers delay their monthly instalment for 9 - 17 days (about 2 and a half weeks), followed by 2- 9 days and 17 - 24 days (about 3 and a half weeks) from their due date. The maximum day from delaying their payment is 67 days (about 2 months), which is more than 2 months. The minimum days is -5 days which is unusual, as this may indicate that borrowers make advanced payment.
Payment of Minimum Amount 	Categorized into 3 categories: • Made minimum amount payment 9,274 (46%) • Did not make minimum amount payment 8,334 (42%) • Not mentioned 2,392 (12%)

Payment Behavior	

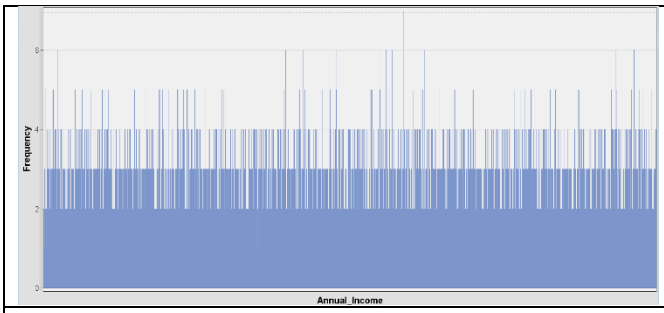
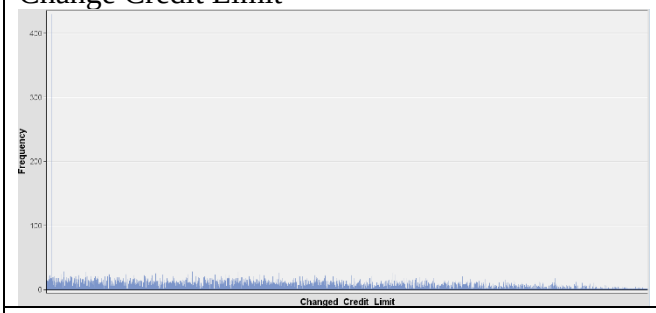
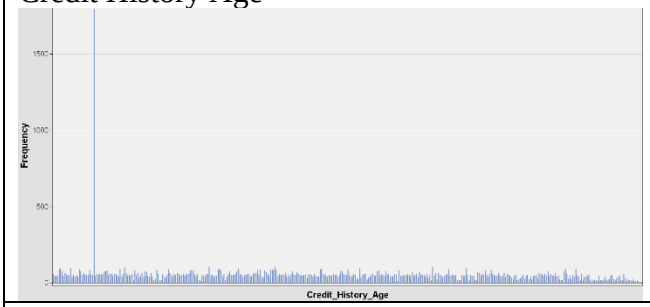
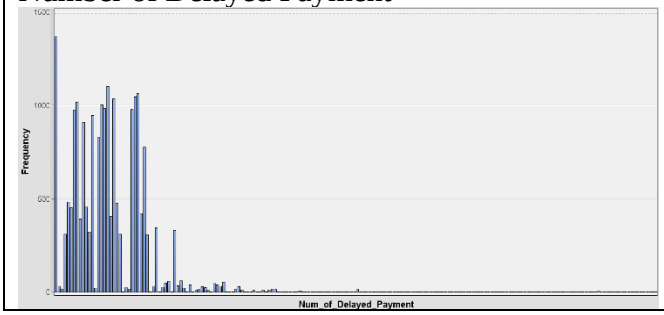
	Payment behavior is a breakdown into spending power and value of payments categories. The categories are ranked from highest to lowest frequency as below: • Low spending with small value payment 4,938 (25%) • High spending with medium value payment 3,538 (18%) • High spending with large value payment 2,796 (14%) There are 1,540 invalid values found in this variable.
Occupation 	The occupation across the borrowers is fairly distributed, with the occupation types ranked from highest to lowest frequency: • Teacher 1,330 (7%) • Lawyer 1,309 (7%) • Architect 1,303 (7%) There are unknown values amounting to 1,357 (7%) of the dataset.
Number of Bank Accounts 	The variable shows that mostly the number of bank accounts a borrower has lies between 1-10 accounts. Unusual pattern detected in the variable. Reason being are due to: • Dirty data is found, which the data type is detected as character, rather than numeric. • Noisy data is detected with 835 invalid values showing 0 number of accounts, which is unusual as each borrower must have at least 1 bank account to obtain a loan.

Interest Rate	The variable shows that the interest rate mostly falls between 0-34%, with 8% as

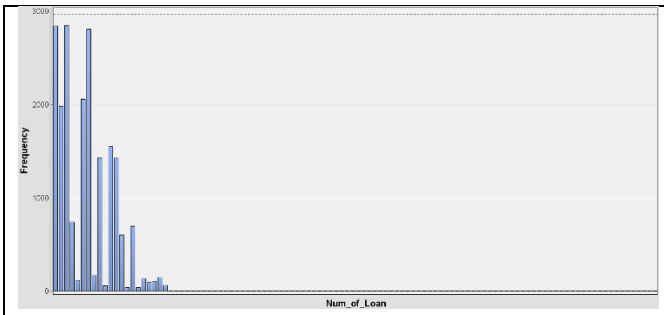
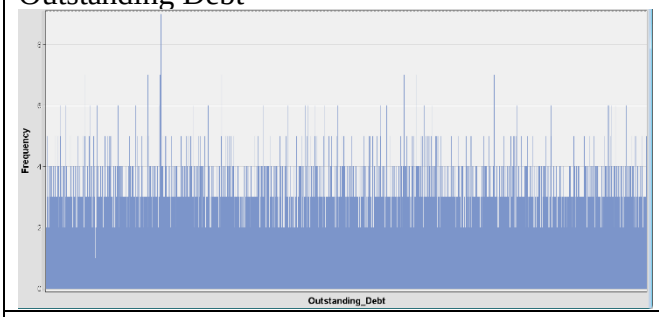
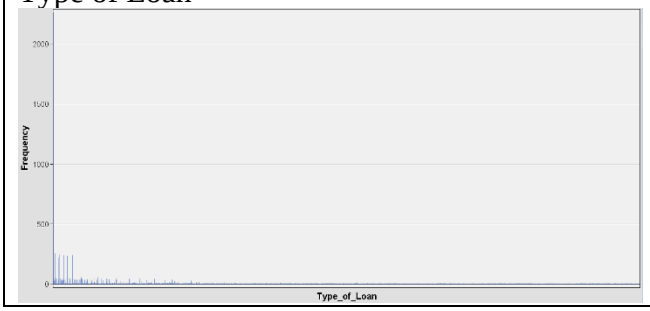
	the most common interest rate for 1,081 borrowers. Unusual pattern detected in the variable. Reason being are due to: • Noisy data detected with maximum interest rate of 5,775% shown which is abnormal as the maximum value for interest rate should be 100%.
Monthly In Hand Salary 	• Right-skewed distribution with 4,898 borrowers (24%) having a monthly in hand salary between $303 and $1,794, and 3,936 (20%) having salary in between $1,793 and $3,286. • 3,042 missing values detected in this variable.
Monthly Balance 	• 7,420 borrowers (37%) fall in the range of $158 and $315 monthly balance, followed by 6,466 borrowers (32%) with monthly balance between $315 and $471. • 255 missing values are found in the variable.
Number of Credit Inquiries 	• Highly skewed distribution. • 2,378 values fall under 4 credit enquiries. • There are 1,536 and 379 with no enquiries and missing values. • Outliers detected as with maximum value of 2,592 credit inquiries, which is unusual.
Number of Credit Cards	• Highly skewed distribution.

	• 3,726 borrowers (19%) own 5 credit cards. • Most borrowers own between 0 and 10 credit cards. • Outliers detected with 1,498 as the maximum number of credit cards owned by a single borrower, which is very unusual.
Total EMI per Month 	• Highly skewed distribution. • 19,397 of the Estimated monthly instalment (EMI) records (97%) lie between $0 and $8226. • Outliers detected with a maximum value of $82,236 EMI.
Age 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the age distribution.
Amount Invested Monthly 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the amount invested monthly distribution.

Annual Income	

	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the annual income distribution.
Change Credit Limit 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the change credit limit distribution.
Credit History Age 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the credit history age distribution.
Number of Delayed Payment 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. 1,416 records shown as missing values. Data cleaning is required to reflect the accuracy of the number of delayed payment distribution.

Number of Loan	

	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. 978 is shown as the maximum value of loan, which is unusual and inaccurate as the number of loans is not arranged according to numerical order. Data cleaning is required to reflect the accuracy of the number of loan distribution.
Outstanding Debt 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the outstanding debt distribution.
Type of Loan 	Unusual pattern in the data due to dirty data as the data type is detected as character, rather than numeric. Data cleaning is required to reflect the accuracy of the loan type distribution.

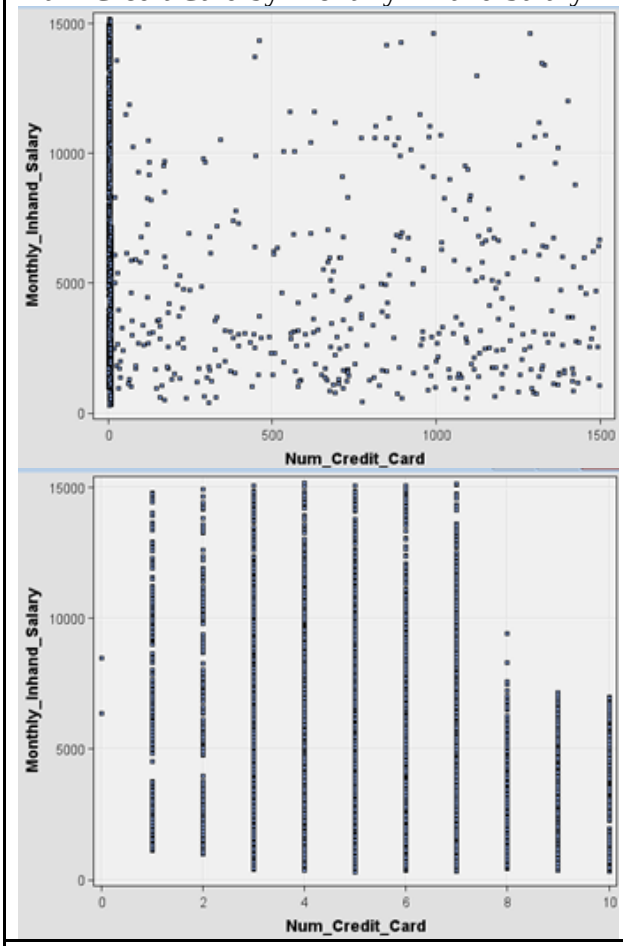
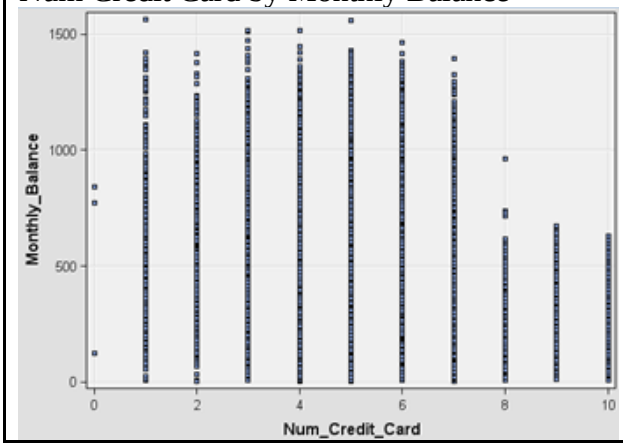
6.2.2 Bivariate Analysis

Unlike univariate analysis, bivariate analysis is a statistical method used to examine the relationship between two variables by analyzing the degree of association or correlation between them and identifying patterns or trends in the data. In this project, we focus on exploring the relationship between two input variables and examining how each continuous input variable is connected to the target variable with scatter plots and bar plots to illustrate these relationships. By conducting this bivariate analysis, valuable insights on how the variables are related can be obtained, thus informed decisions can be made based on the findings.

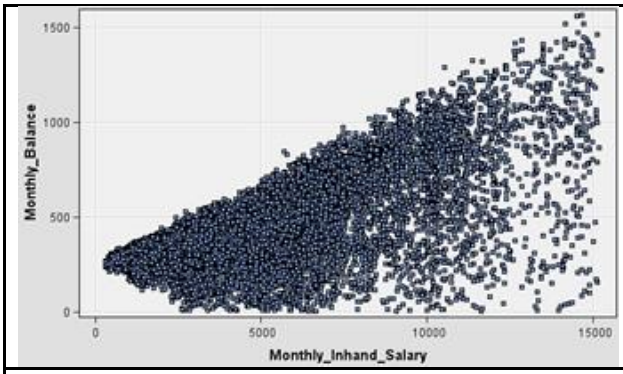
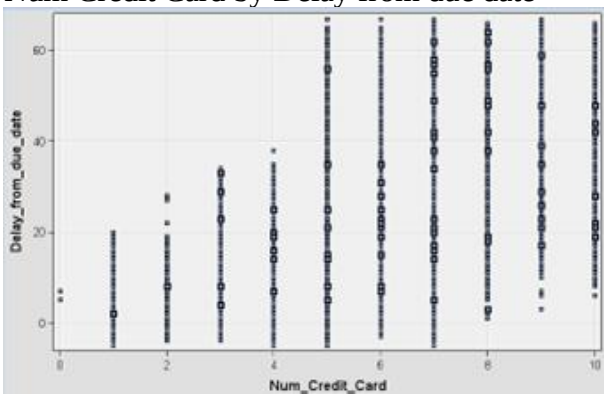
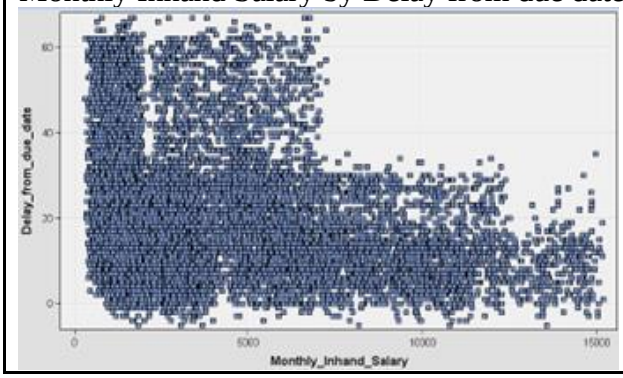
6.2.2.1 Bivariate Analysis on Input Variables

Identifying relationships between features is crucial in selecting pertinent input variables for a machine learning model, guiding feature engineering efforts, improving model performance, as well as interpreting the analysis results for better decision making.

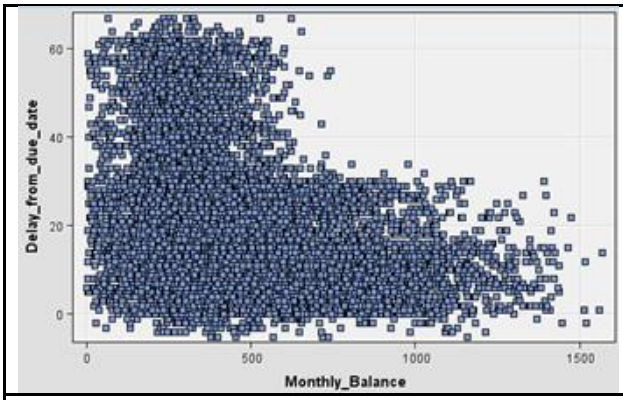
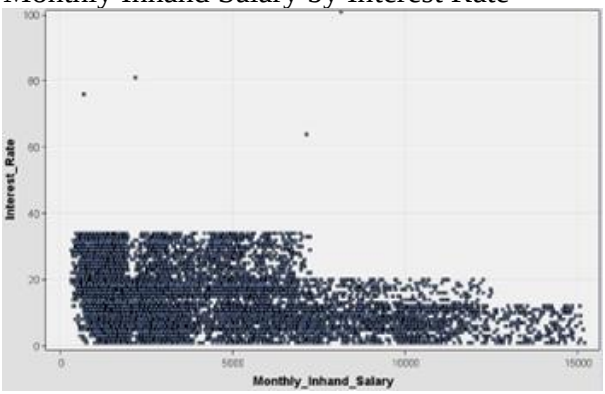
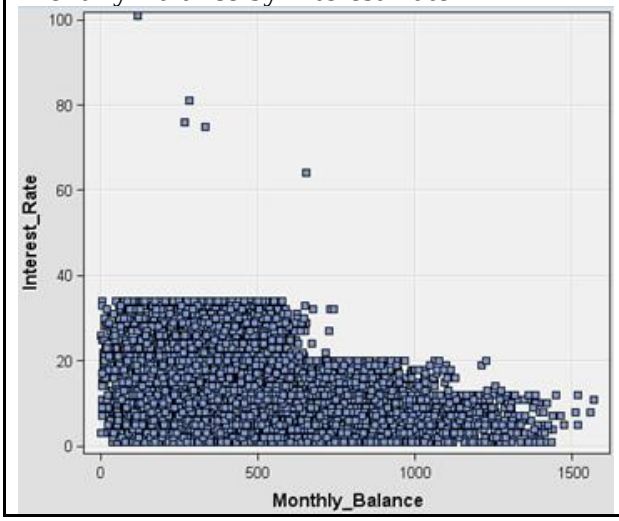
Figure 6.2.2.1 below displays and describes the connection between two input variables in the dataset.

Bivariate Analysis Graphs	Findings
Num Credit Card by Monthly Inhand Salary 	Based on the first plot, it is observed that the Num Credit Card variable contains a lot of outliers, and most customers fall in the range of 0 to 10 credit cards. Focusing on customers with credit cards less than or equal to 10, there is a huge drop in the monthly in-hand salary for customers owned more than 7 credit cards.
Num Credit Card by Monthly Balance 	It is clearly seen that the monthly balance experienced a huge fall for customers owned more than 7 credit cards, showing a similar trend described in Num Credit Card and Monthly Inhand Salary. This indicates there may be a strong connection between monthly in-hand salary and monthly balance, but more study is needed to assure their relationship.

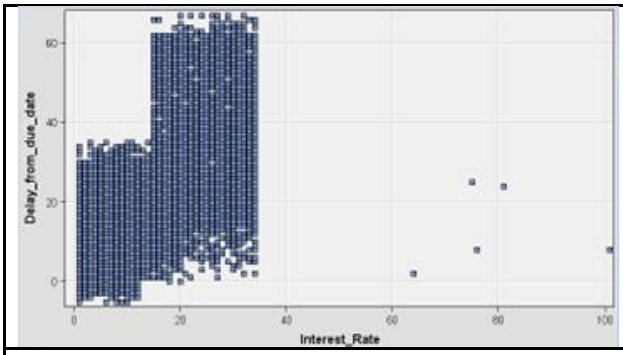
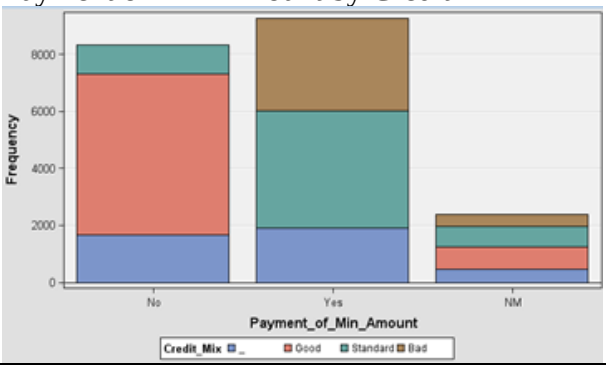
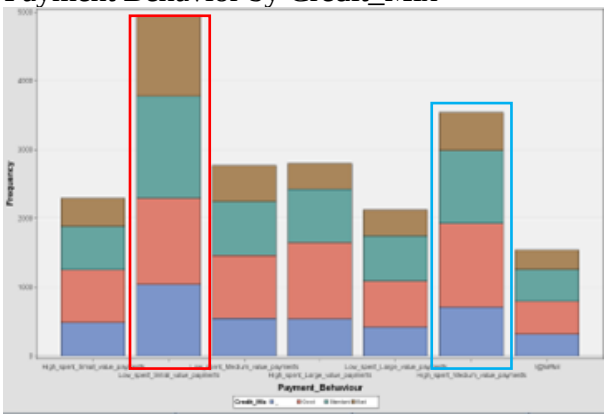
Monthly Inhand Salary by Monthly Balance	

	It is proven that the higher the monthly in-hand salary, the higher the monthly balance of the customer.
Num Credit Card by Delay from due date 	Customers tend to pay their credit card debt later if they have more credit cards. The number of days delayed from the payment due date increases sharply if the customer holds more than 4 credit cards. This is reasonable as taking multiple credit cards can make debt repayments unsustainable. Since there is some connection between the number of credit cards and delay from due date, and number of credit card and monthly in-hand salary, then it is worth to investigate the relationship between monthly in-hand salary and delay from due date.
Monthly Inhand Salary by Delay from due date 	It is observed that the number of customers with monthly in-hand salary more than 2000 and pay loan later than 40 days from due date is reduced, and all customers with monthly in-hand salary greater than 7000 settled their debt in less than 40 days from due date. Customers with higher monthly in-hand salary pay their debt earlier, as compared to those with lower monthly in-hand salary.

Monthly Balance by Delay from due date	

	Similar pattern is observed in this plot as compared to the monthly in-hand salary and delay from due date scatter plot. There is also a huge drop in the number of customers who make payment 40 days after the due date if their monthly balance is more than 500.
Monthly Inhand Salary by Interest Rate  Monthly Balance by Interest Rate 	Similar patterns are observed in these plots as compared to the monthly in-hand salary and delay from due date scatter plot. The interest rate falls by half for customers with monthly in-hand salary more than 7000 and monthly balance more than 500.

Interest Rate by Delay from due date	

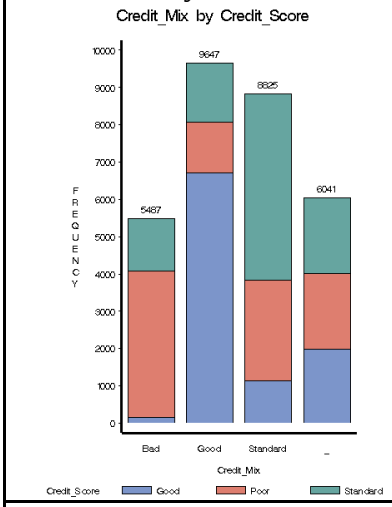
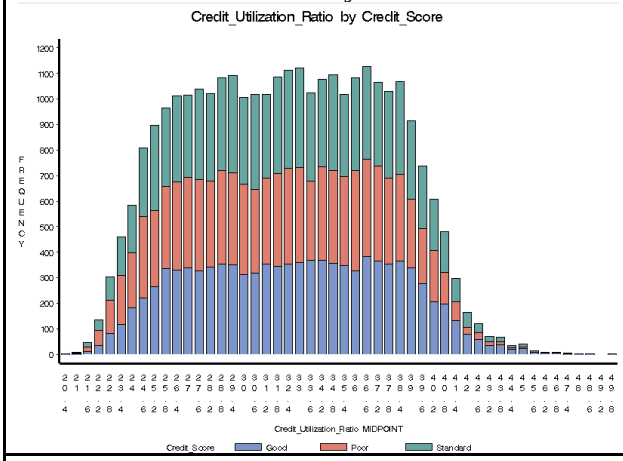
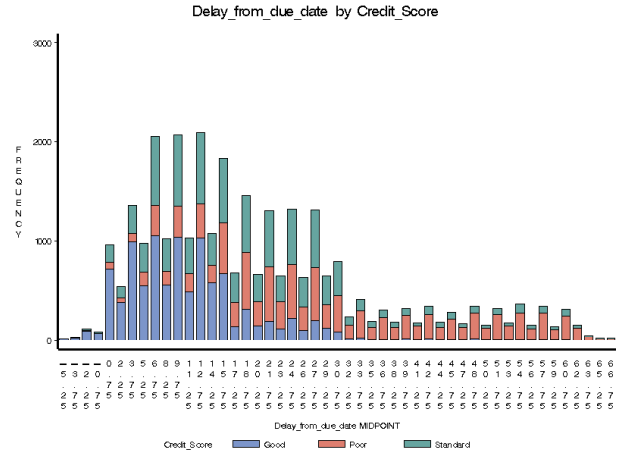
	Customers with higher interest rates on their credit cards are suffering from paying their credit card debt as those with credit card interest higher than 20% delayed their payment twice times more than those with credit card interest less than 20%.
Payment of Min Amount by Credit Mix 	Most customers who paid only minimum monthly payments on their credit card have standard to low credit mix score while customers who made payment more than required minimum amount or full amount have standard to good score.
Payment Behavior by Credit_Mix 	The proportion of customers with poor credit mix score is significantly higher if their payment behavior is low spent and small value payment (red box), compared to customers with low spent and large value payment (blue box). Therefore, it is said that credit card payment is one of the essential elements affecting credit mix score.

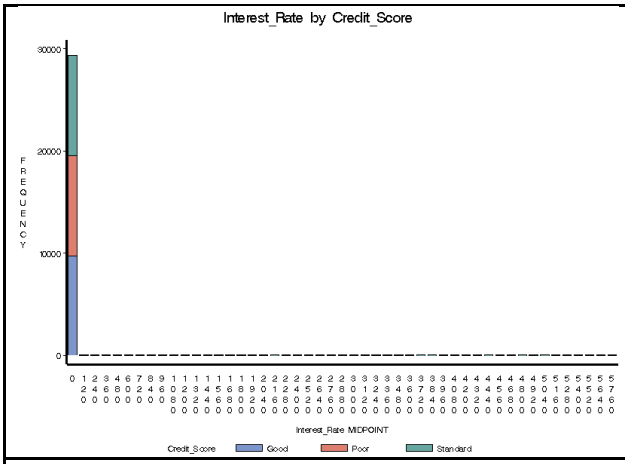
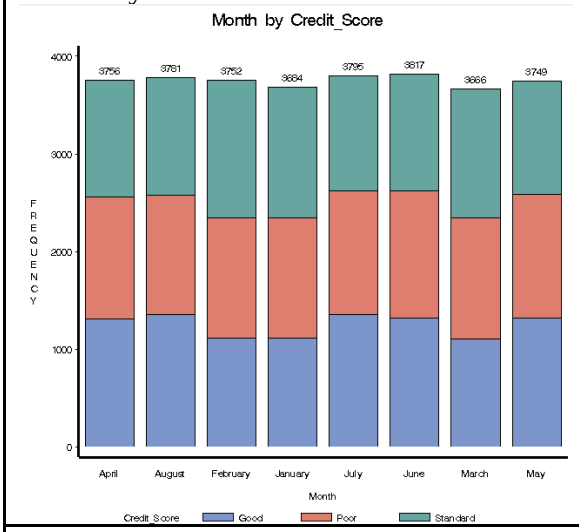
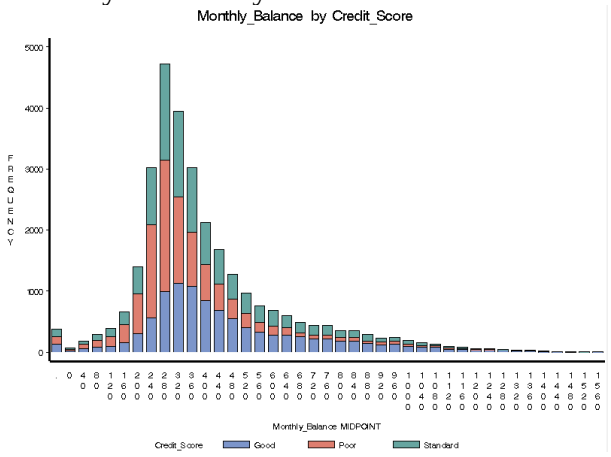
6.2.2.2 Bivariate Analysis on Input and Target Variables

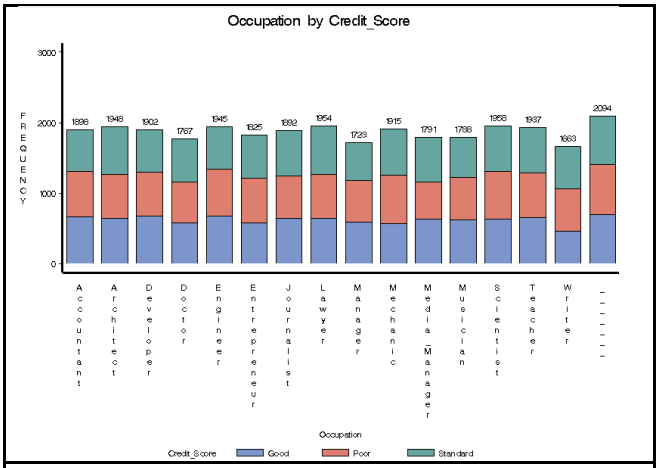
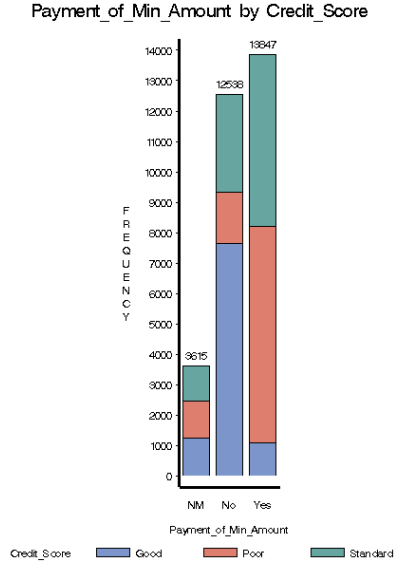
Bivariate analysis serves as one of the tools for determining whether there is a correlation between the input features and the target variable. This is vital in a predictive modelling context as the strength of the relationship will significantly impact the accuracy and interpretation of the model. A feature with strong positive or negative correlation to the target variable is treated as a dependable predictor, and thus it is worth to incorporate into the model training stage to boost the accuracy. Conversely, a feature with weak to no correlation is not useful to the predictive model as it will not enhance the model accuracy.

Figure 6.2.2.2 below illustrates and explains the relationship between input variables and target variable.

Bivariate Analysis Graphs	**Findings**

Credit Mix by Credit Score 	The credit mix is related to credit score as shown in the graph where most customers with good credit mix also obtained good credit score.
Credit Utilization Ratio by Credit Score 	The credit utilization ratio shows a normal distribution with almost equal portion of good, standard, and poor credit score at most of the bins.
Delay from due date by Credit Score 	The proportion of customers with standard and good credit score reduced, and the number of customers with poor credit score rose when the number of days delayed from due date increased. This indicates that customers with good credit scores pay their credit card debt much earlier than those with poor credit score.
Interest Rate by Credit Score	

	The bar plot did not show any meaningful results due to outliers in interest rate.
Month by Credit Score 	There is no significant difference in the amount of customers in various credit scores across months.
Monthly Balance by Credit Score 	The plot illustrates a right-skewed normal distribution with a small spike at monthly balance less than 0. Customers with poor credit scores are relatively high when the monthly balance is closer to mean, but the number of customers drops drastically when the monthly balance gets higher. This shows the connection between monthly balance and credit score although it is not that obvious as compared to other plots.
Occupation by Credit Score	

	There is no obvious difference in the distribution of credit score in each segment of the customers' occupation.
Payment of Min Amount by Credit Score 	The number of customers with standard to poor credit scores and paying only minimum payment is significantly higher. Hence, payment of minimum amount is also directly affecting the credit score.

6.2.3 Multivariate Analysis

Multivariate analysis is a statistical method that studies simultaneous effect of more than two independent variables and the relationships among the datasets.

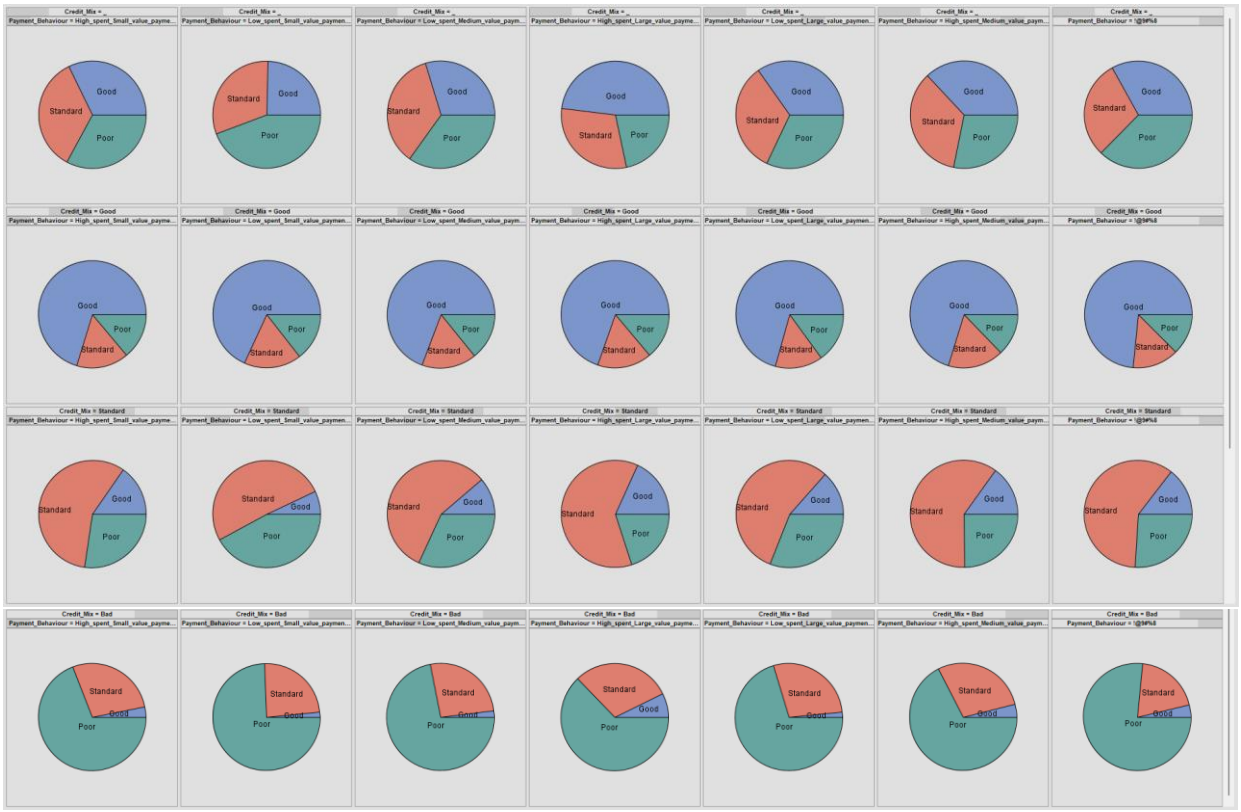


Figure 6.2.3.1 Lattice Pie Charts

The Lattice Bar Charts show how the categories Payment_Behaviour, and Credit_Mix, correlate with Credit_Score. Generally, having a good Credit_Mix gives a good Credit_Score, standard Credit_Mix gives a standard Credit_Score, and poor Credit_Mix gives a poor Credit_Score regardless of the Payment_Behaviour. When the Credit_Mix is undefined, High_spent Payment_Behaviour does not really affect the Credit_Score unless Large_value_payments are made, which increases the likelihood of a good Credit_Score. Low_spent Payment Behaviour also does not affect Credit_Score unless Small_value_payments are made, which results in poor Credit_Score.

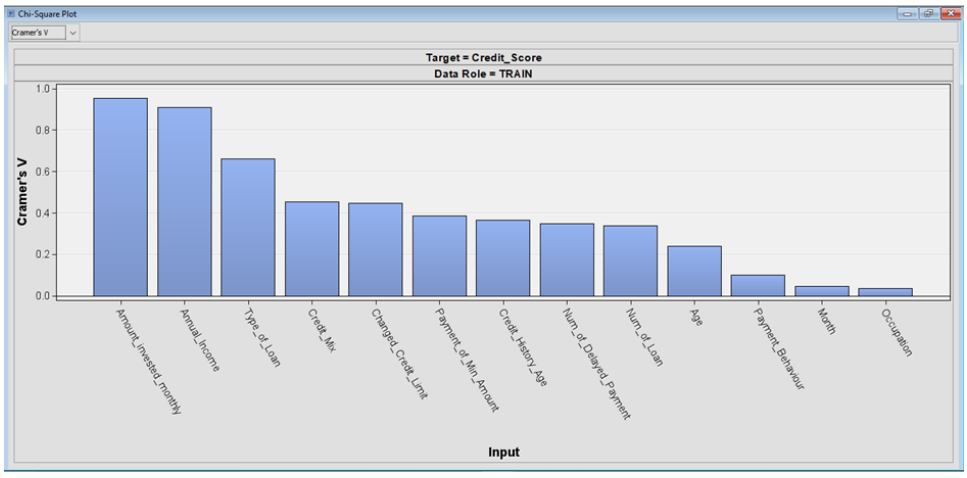


Figure 6.2.3.2 Cramer's V Statistics with respect to the Target Variable

Figure 6.2.3.2 shows Cramer's V measure in which the 0 value indicates no association whereas the 1 indicates perfect association between the two variables. The threshold value is set to 0.1. Thus, Amount invested monthly, Annual Income, Type of Loan, Credit Mix, Changed Credit Limit, Payment of Min Amount, Credit History Age, Num of Delayed Payment, Num of Loan, Age, and Payment Behaviour show association with the target variable which is the Credit Score. Meanwhile, Month and Occupation show nearly no association with Credit Score.

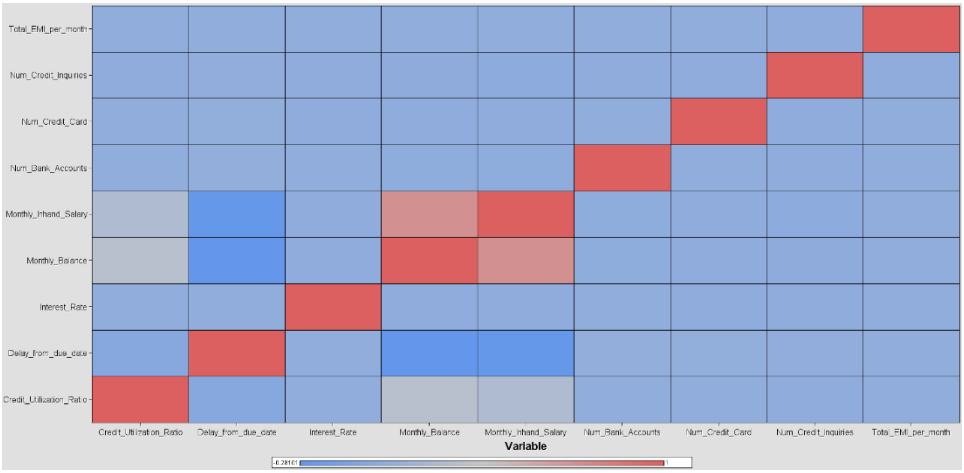


Figure 6.2.3.3 Correlation Matrix

Table 6.3.2.1 Correlation Table

Variable 1	Variable 2	Correlation
Credit_Utilization_Ratio	Credit_Utilization_Ratio	1.000
Credit_Utilization_Ratio	Delay_from_due_date	-0.055
Credit_Utilization_Ratio	Interest_Rate	0.001
Credit_Utilization_Ratio	Monthly_Balance	0.258
Credit_Utilization_Ratio	Monthly_Inhand_Salary	0.196
Credit_Utilization_Ratio	Num_Bank_Accounts	0.001
Credit_Utilization_Ratio	Num_Credit_Card	0.002
Credit_Utilization_Ratio	Num_Credit_Inquiries	-0.002
Credit_Utilization_Ratio	Total_EMI_per_month	0.004
Delay_from_due_date	Credit_Utilization_Ratio	-0.055
Delay_from_due_date	Delay_from_due_date	1.000
Delay_from_due_date	Interest_Rate	0.009
Delay_from_due_date	Monthly_Balance	-0.281
Delay_from_due_date	Monthly_Inhand_Salary	-0.265
Delay_from_due_date	Num_Bank_Accounts	0.023
Delay_from_due_date	Num_Credit_Card	0.016
Delay_from_due_date	Num_Credit_Inquiries	0.011

Delay_from_due_date	Total_EMI_per_month	0.002
Interest_Rate	Credit_Utilization_Ratio	0.001
Interest_Rate	Delay_from_due_date	0.009
Interest_Rate	Interest_Rate	1.000
Interest_Rate	Monthly_Balance	-0.002
Interest_Rate	Monthly_Inhand_Salary	-0.005
Interest_Rate	Num_Bank_Accounts	0.002
Interest_Rate	Num_Credit_Card	-0.007
Interest_Rate	Num_Credit_Inquiries	-0.002
Interest_Rate	Total_EMI_per_month	0.001
Monthly_Balance	Credit_Utilization_Ratio	0.258
Monthly_Balance	Delay_from_due_date	-0.281
Monthly_Balance	Interest_Rate	-0.002
Monthly_Balance	Monthly_Balance	1.000
Monthly_Balance	Monthly_Inhand_Salary	0.705
Monthly_Balance	Num_Bank_Accounts	-0.012
Monthly_Balance	Num_Credit_Card	-0.005
Monthly_Balance	Num_Credit_Inquiries	-0.015
Monthly_Balance	Total_EMI_per_month	0.007
Monthly_Inhand_Salary	Credit_Utilization_Ratio	0.196
Monthly_Inhand_Salary	Delay_from_due_date	-0.265
Monthly_Inhand_Salary	Interest_Rate	-0.005
Monthly_Inhand_Salary	Monthly_Balance	0.705
Monthly_Inhand_Salary	Monthly_Inhand_Salary	1.000
Monthly_Inhand_Salary	Num_Bank_Accounts	-0.011
Monthly_Inhand_Salary	Num_Credit_Card	-0.002
Monthly_Inhand_Salary	Num_Credit_Inquiries	-0.015
Monthly_Inhand_Salary	Total_EMI_per_month	0.008
Num_Bank_Accounts	Credit_Utilization_Ratio	0.001
Num_Bank_Accounts	Delay_from_due_date	0.023
Num_Bank_Accounts	Interest_Rate	0.002
Num_Bank_Accounts	Monthly_Balance	-0.012
Num_Bank_Accounts	Monthly_Inhand_Salary	-0.011
Num_Bank_Accounts	Num_Bank_Accounts	1.000
Num_Bank_Accounts	Num_Credit_Card	0.003
Num_Bank_Accounts	Num_Credit_Inquiries	-0.006
Num_Bank_Accounts	Total_EMI_per_month	0.003
Num_Credit_Card	Credit_Utilization_Ratio	0.002
Num_Credit_Card	Delay_from_due_date	0.016
Num_Credit_Card	Interest_Rate	-0.007
Num_Credit_Card	Monthly_Balance	-0.005

Num_Credit_Card	Monthly_Inhand_Salary	-0.002
Num_Credit_Card	Num_Bank_Accounts	0.003
Num_Credit_Card	Num_Credit_Card	1.000
Num_Credit_Card	Num_Credit_Inquiries	-0.005
Num_Credit_Card	Total_EMI_per_month	0.000
Num_Credit_Inquiries	Credit_Utilization_Ratio	-0.002
Num_Credit_Inquiries	Delay_from_due_date	0.011
Num_Credit_Inquiries	Interest_Rate	-0.002
Num_Credit_Inquiries	Monthly_Balance	-0.015
Num_Credit_Inquiries	Monthly_Inhand_Salary	-0.015
Num_Credit_Inquiries	Num_Bank_Accounts	-0.006
Num_Credit_Inquiries	Num_Credit_Card	-0.005
Num_Credit_Inquiries	Num_Credit_Inquiries	1.000
Num_Credit_Inquiries	Total_EMI_per_month	-0.005
Total_EMI_per_month	Credit_Utilization_Ratio	0.004
Total_EMI_per_month	Delay_from_due_date	0.002
Total_EMI_per_month	Interest_Rate	0.001
Total_EMI_per_month	Monthly_Balance	0.007
Total_EMI_per_month	Monthly_Inhand_Salary	0.008
Total_EMI_per_month	Num_Bank_Accounts	0.003
Total_EMI_per_month	Num_Credit_Card	0.000
Total_EMI_per_month	Num_Credit_Inquiries	-0.005
Total_EMI_per_month	Total_EMI_per_month	1.000

The correlation matrix and table can be used to determine the variables to be dropped based on the correlation coefficients. The correlation matrix takes the outliers into consideration when calculating the correlation. The closer the number is to 1 or -1, the stronger the correlation. The closest positive correlation value obtained was 0.7 from Monthly In-hand Salary and Monthly Balance. However, since none of the values were in the range of |0.9| to |1|, none of the variables were dropped.

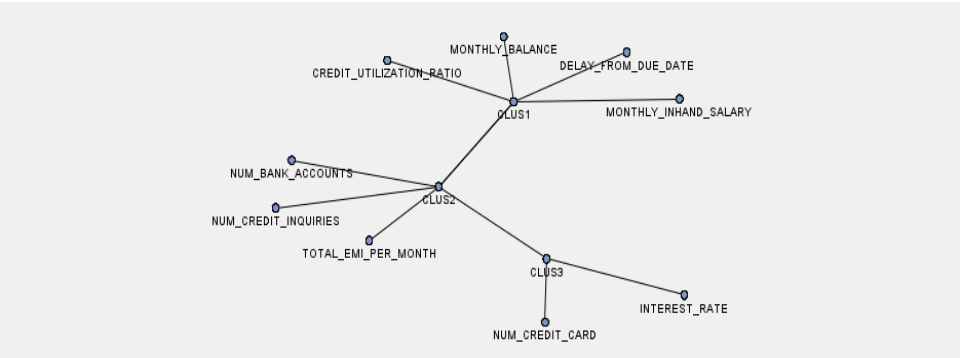


Figure 6.2.3.4 Cluster Plot

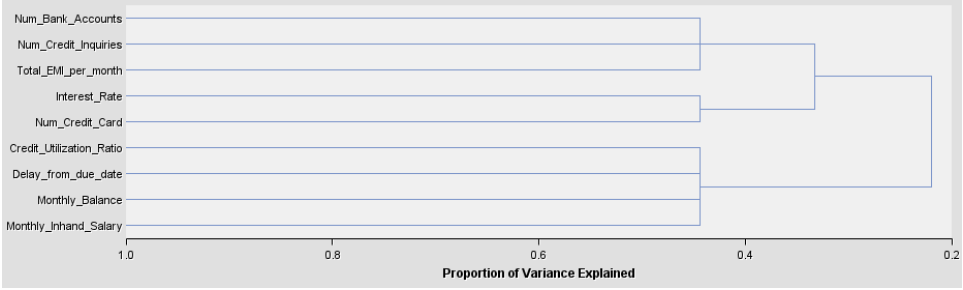


Figure 6.2.3.5 Dendrogram of Variance Proportion

Variable Selection Table

Cluster	Variable	R-Square With Own Cluster Component	Next Closest Cluster	R-Square with Next Cluster Component	Type	Label	1-R2 Ratio	Variable Selected
CLUS1	MONTHLY_BALANCE	0.781366	CLUS2	3.433E-5	Variable		0.218641	YES
CLUS2	NUM_CREDIT_INQUIRIES	0.416614	CLUS1	.0002716	Variable		0.583544	YES
CLUS3	INTEREST_RATE	0.503625	CLUS1	2.494E-5	Variable		0.496388	YES
CLUS1	CLUS1	1	CLUS2	2.474E-5	ClusterComp	Cluster 1	0	NO
CLUS1	MONTHLY_INHAND_SALA...	0.744971	CLUS2	5.799E-5	Variable		0.255044	NO
CLUS1	DELAY_FROM_DUE_DATE	0.262602	CLUS2	5.051E-5	Variable		0.737435	NO
CLUS1	CREDIT_UTILIZATION_RA...	0.189093	CLUS2	1.582E-5	Variable		0.81092	NO
CLUS2	CLUS2	1	CLUS1	2.474E-5	ClusterComp	Cluster 2	0	NO
CLUS2	NUM_BANK_ACCOUNTS	0.338754	CLUS1	.0002425	Variable		0.661406	NO
CLUS2	TOTAL_EMI_PER_MONTH	0.254022	CLUS1	4.678E-5	Variable		0.746013	NO
CLUS3	CLUS3	1	CLUS2	3.477E-6	ClusterComp	Cluster 3	0	NO
CLUS3	NUM_CREDIT_CARD	0.503625	CLUS1	4.742E-5	Variable		0.496399	NO

Figure 6.2.3.6 Variable Selection Table with Best Variable

Figure 6.2.3.4 and Figure 6.2.3.5 depict the variable clustering of the dataset prior to any data cleaning and transformation. There were three clusters made due to each cluster satisfies the stopping criteria specified in the Variation Proportion property. The first cluster is made of Credit_Utilization_Ratio, Monthly_Balance, Delay_from_due_date, Monthly_Inhand_Salary. The second cluster consists of Num_Bank_Accounts, Num_Credit_Inquiries, and Total_EMI_per_month. The last cluster has Num_Credit_Card and Interest_Rate. Monthly_Balance, Num_Credit_Inquiries, and Interest_Rate were the best variable in each cluster with the lowest 1–R^2 Ratio.

6.2.4 Interesting Visualization

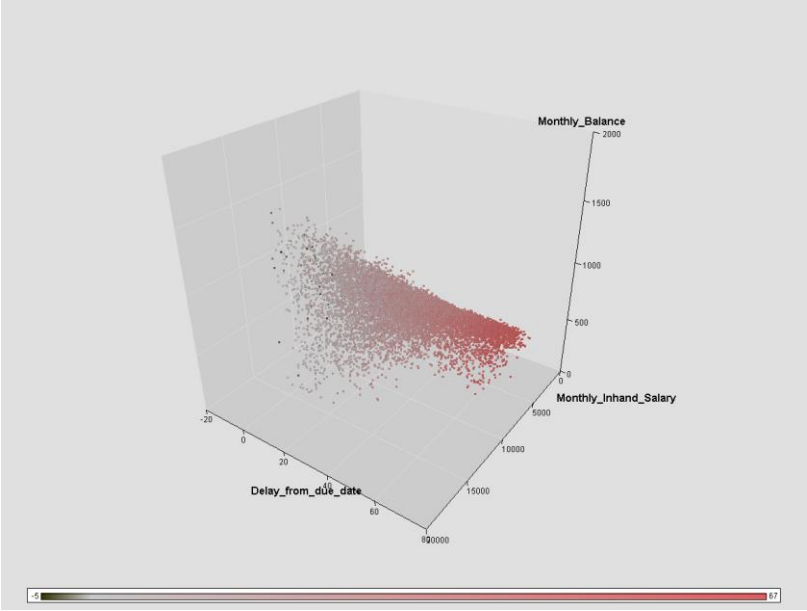


Figure 6.2.4.1 3-dimensional Scatter Plot

The Scatter plot depicts the relationship between Monthly_Inhand_Salary (x-axis), Delay_from_due_date (y-axis, colour), and Monthly_Balance (z-axis). From the plot it can be seen that Monthly_Inhand_Salary and Monthly_Balance increase proportionally. As for Delay_from_due_date, it decreases as the Monthly_Inhand_Salary and Monthly_Balance increase.

7 SEMMA: Modify

In the Modify phase, the dataset is first pre-modified via base SAS using SAS Studio, which is a programming language for data access, data transformation, analysis, and reporting. The dataset is then further modified in SAS Enterprise Miner using Modify nodes, such as Impute, Drop, Transform, Replacement nodes etc. After modifying the dataset, each of the variables is explored and the results before and after transformation are compared.

7.1 Pre-modifying Dataset using Base SAS

Base SAS code is used to address the variables level misclassification issues and perform some data cleaning. The following demonstrates the code and describes all the data cleaning steps involved in detail.

1. A library named "cleaning" is created and the dataset is copied to the library for data cleaning purposes. (Figure 12.4.1)
2. A SAS Macro is defined to remove special characters except dot and comma, in each of the character variables in the dataset. (Figures 12.4.2 - 12.4.3)
3. A new unique id column named "id_new" is constructed to overcome the duplication of ID in the dataset. (Figure 12.4.4)
4. The levels of the character variables such as Age, Annual Income, Num of Loan, Num of Delayed Payment, Changed Credit Limit, Outstanding Debt and Amount invested monthly

are finally ready for reclassification. These variables are also rounded to 2 decimal places after converting from character to numeric. (Figure 12.4.5)

5. The Credit History Age variable is transformed into Credit History Year and Credit History Month, then into new Credit History Age variable in year format. For instance, 22 years 10 months is transformed into 22 Credit History Year and 10 Credit History Month, then into 22.83 Credit History Age (year). (Figure 12.4.6)
6. The Type of Loan variable, which contains a bunch of loans with commas in between, is extracted and transformed by splitting the loan by comma, then transposing from column into rows. (Figure 12.4.7)
7. After transposed, the distinct type of loan is obtained, and the dirty data in the type of loan column is cleaned to ensure that the data is consistent, accurate and reliable. (Figure 12.4.8)
8. New variables (auto_loan_f, creditbuilder_loan_f, debt_con_loan_f, home_equi_loan_f, mg_loan_f, not_spec_loan_f, payday_loan_f, personal_loan_f, student_loan_f) are constructed to indicate whether each customer holds a specific loan product. For example, if the customer holds an auto loan, then the auto_loan_f will be tagged as Y, otherwise N. (Figure 12.4.9)
9. These new variables are then merged into the main dataset (the dataset after transformed the Credit History Age at point 5) using the id_new column. (Figure 12.4.10)
10. The Month variables are transformed from character to numeric to ease the exploratory data analysis process later. (Figure 12.4.11)
11. The newly created and transformed variables are renamed accordingly. (Figure 12.4.12)
12. The modified dataset is saved as "em_save_train_modified" in the library which was created in the previous phase to store the data source for Credit Classification Score project. (Figure 12.4.13)

7.2 Modifying Dataset using SAS Enterprise Miner

A new data source is created for the modified dataset named "em_save_train_modified" to further perform data pre-processing process in SAS Enterprise Miner. The roles and levels of the variables in the dataset are reclassified appropriately as shown in Figure 7.2.1, whereas the metadata of the new variables created are described in Table 7.2.1. 1.

As observed in Figure 7.2.1, there are 34 variables in total, which are 1 nominal variable with ID role, 5 nominal variables with Input role, 17 interval variables with Input role, 1 ordinal variable with Input role, 9 binary variables with input role, and 1 binary variable as target role in the dataset.

Name	Role	Level
Age	Input	Interval
Amount_invested_monthly	Input	Interval
Annual_Income	Input	Interval
Changed_Credit_Limit	Input	Interval
Credit_History_Age	Input	Interval
Credit_Mix	Input	Ordinal
Credit_Score	Target	Binary
Credit_Utilization_Ratio	Input	Interval
Customer_ID	Input	Nominal
Delay_from_due_date	Input	Interval
Interest_Rate	Input	Interval
Month	Input	Nominal
Monthly_Balance	Input	Interval
Monthly_Inhand_Salary	Input	Interval
Num_Bank_Accounts	Input	Interval
Num_Credit_Card	Input	Interval
Num_Credit_Inquiries	Input	Interval
Num_of_Delayed_Payment	Input	Interval
Num_of_Loan	Input	Interval
Occupation	Input	Nominal
Outstanding_Debt	Input	Interval
Payment_Behaviour	Input	Nominal
Payment_of_Min_Amount	Input	Nominal
Total_EMI_per_month	Input	Interval
auto_loan_f	Input	Binary
creditbuilder_loan_f	Input	Binary
debt_con_loan_f	Input	Binary
home_equi_loan_f	Input	Binary
id	ID	Nominal
mg_loan_f	Input	Binary
not_spec_loan_f	Input	Binary
payday_loan_f	Input	Binary
personal_loan_f	Input	Binary
student_loan_f	Input	Binary

Figure 7.2.1: The roles and levels of the variables

Column	Description
auto_loan_f	Represents whether the customer holding auto loan
creditbuilder_loan_f	Represents whether the customer holding credit-builder loan
debt_con_loan_f	Represents whether the customer holding debt consolidation loan
home_equi_loan_f	Represents whether the customer holding home equity loan
mg_loan_f	Represents whether the customer holding mortgage loan
not_spec_loan_f	Represents whether the customer holding non-specified loan
payday_loan_f	Represents whether the customer holding payday loan
personal_loan_f	Represents whether the customer holding personal loan
student_loan_f	Represents whether the customer holding student loan

Table 7.2.1: New columns metadata

After creating the data source, the dataset is dragged to the diagram and ready for data pre-processing. The steps involved in Modify stage are discussed in detail as follows, along with the screenshots of how the steps performed in SAS Enterprise Miner in brackets after each step.

1. The values in class variables are replaced with their correct and default values using the Replacement node under the Modify tab as shown in Figures 7.2.2 and 7.2.3 (Figures 12.5.1 - 12.5.8).

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
Credit_Mix	Good		9647	C	Good	.
Credit_Mix	Standard		8825	C	Standard	.
Credit_Mix		_MISSING_	6041	C		.
Credit_Mix	Bad		5487	C	Bad	.
Credit_Mix	_UNKNOWN_	_DEFAULT_	.	C		.
Credit_Score	Good		10000	C	Good	.
Credit_Score	Poor		10000	C	Poor	.
Credit_Score	Standard		10000	C	Standard	.
Credit_Score	_UNKNOWN_	_DEFAULT_	.	C		.
Customer_ID	_UNKNOWN_	_DEFAULT_	.	C		.
Month	6		3817	N		6
Month	7		3795	N		7
Month	8		3781	N		8
Month	4		3756	N		4
Month	2		3752	N		2
Month	5		3749	N		5
Month	1		3684	N		1
Month	3		3666	N		3
Month	_UNKNOWN_	_DEFAULT_	.	N		.
Occupation		_MISSING_	2094	C		.
Occupation	Scientist		1958	C	Scientist	.
Occupation	Lawyer		1954	C	Lawyer	.
Occupation	Architect		1948	C	Architect	.
Occupation	Engineer		1945	C	Engineer	.
Occupation	Teacher		1937	C	Teacher	.
Occupation	Mechanic		1915	C	Mechanic	.
Occupation	Developer		1902	C	Developer	.
Occupation	Accountant		1898	C	Accountant	.
Occupation	Journalist		1892	C	Journalist	.
Occupation	Entrepreneur		1825	C	Entrepreneur	.
Occupation	MediaManager		1791	C	MediaManager	.
Occupation	Musician		1788	C	Musician	.
Occupation	Doctor		1767	C	Doctor	.
Occupation	Manager		1723	C	Manager	.
Occupation	Writer		1663	C	Writer	.
Occupation	_UNKNOWN_	_MISSING_	.	C		.
Payment_Behaviour	LowspentSmallvaluepayments	**Low Spent Small Value Payment**	7323	C	LowspentSmallvaluepayments	.
Payment_Behaviour	HighspentMediumvaluepayments	**High Spent Medium Value Payment**	5354	C	HighspentMediumvaluepayments	.
Payment_Behaviour	HighspentLargevaluepayments	**High Spent Large Value Payment**	4253	C	HighspentLargevaluepayments	.
Payment_Behaviour	LowspentMediumvaluepayments	**Low Spent Medium Value Payment**	4171	C	LowspentMediumvaluepayments	.
Payment_Behaviour	HighspentSmallvaluepayments	**High Spent Small Value Payment**	3375	C	HighspentSmallvaluepayments	.
Payment_Behaviour	LowspentLargevaluepayments	**Low Spent Large Value Payment**	3176	C	LowspentLargevaluepayments	.
Payment_Behaviour	98	_MISSING_	2348	C	98	.
Payment_Behaviour	_UNKNOWN_	_DEFAULT_	.	C		.
Payment_of_Min_Amount	Yes		13847	C	Yes	.
Payment_of_Min_Amount	No		12538	C	No	.
Payment_of_Min_Amount	NM	_MISSING_	3615	C	NM	.
Payment_of_Min_Amount	_UNKNOWN_	_DEFAULT_	.	C		.

Figures 7.2.2 - 7.2.3: Replacement values for each class variables

2. The values in interval variables that fall outside the specified range are replaced with missing values using the Replacement node as shown in Figure 7.2.4 (Figures 12.5.9 - 12.5.11).

Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace Method	Lower Replacement Value	Upper Replacement Value	Role	Level
Age	Default	User Specified	18	75	Missing	.	.	Input	Interval
Amount_invested_monthly	Default	User Specified	0	.	Missing	.	.	Input	Interval
Annual_Income	Default	User Specified	0	.	Missing	.	.	Input	Interval
Changed_Credit_Limit	Default	User Specified	.	.	Default	.	.	Input	Interval
Credit_History_Age	Default	User Specified	0	.	Missing	.	.	Input	Interval
Credit_Utilization_Ratio	Default	User Specified	0	100	Missing	.	.	Input	Interval
Delay_from_due_date	Default	User Specified	0	.	Missing	.	.	Input	Interval
Interest_Rate	Default	User Specified	0	100	Missing	.	.	Input	Interval
Monthly_Balance	Default	User Specified	0	.	Missing	.	.	Input	Interval
Monthly_Inhand_Salary	Default	User Specified	0	.	Missing	.	.	Input	Interval
Num_Bank_Accounts	Default	User Specified	1	10	Missing	.	.	Input	Interval
Num_Credit_Card	Default	User Specified	0	15	Missing	.	.	Input	Interval
Num_Credit_Inquiries	Default	User Specified	0	.	Missing	.	.	Input	Interval
Num_of_Delayed_Payment	Default	User Specified	0	700	Missing	.	.	Input	Interval
Num_of_Loan	Default	User Specified	0	15	Missing	.	.	Input	Interval
Outstanding_Debt	Default	User Specified	0	.	Missing	.	.	Input	Interval
Total_EMI_per_month	Default	User Specified	0	.	Missing	.	.	Input	Interval
auto_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval
creditbuilder_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval
debt_con_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval
home_equi_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval
mg_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval
not_spec_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval
payday_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval
personal_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval
student_loan_f	Default	User Specified	.	.	Default	.	.	Input	Interval

Figure 7.2.4: Replacement of interval values

3. After replacing the values in these variables, new columns with column names started with "REP_" are created. Hence, these variables remain except "REP_Customer_ID" while the old variables are dropped as shown in Figure 7.2.5 (Figures 12.5.12 - 12.5.15).

Name	Drop	Role	Level
Age	Yes	Rejected	Interval
Amount_invested_monthly	Yes	Rejected	Interval
Annual_Income	Yes	Rejected	Interval
Changed_Credit_Limit	Default	Input	Interval
Credit_History_Age	Yes	Rejected	Interval
Credit_Mix	Yes	Rejected	Ordinal
Credit_Score	Yes	Rejected	Ordinal
Credit_Utilization_Ratio	Yes	Rejected	Interval
Customer_ID	Yes	Rejected	Nominal
Delay_from_due_date	Yes	Rejected	Interval
Interest_Rate	Yes	Rejected	Interval
Month	Yes	Rejected	Nominal
Monthly_Balance	Yes	Rejected	Interval
Monthly_Inhand_Salary	Yes	Rejected	Interval
Num_Bank_Accounts	Yes	Rejected	Interval
Num_Credit_Card	Yes	Rejected	Interval
Num_Credit_Inquiries	Yes	Rejected	Interval
Num_of_Delayed_Payment	Yes	Rejected	Interval
Num_of_Loan	Yes	Rejected	Interval
Occupation	Yes	Rejected	Nominal
Outstanding_Debt	Yes	Rejected	Interval
Payment_Behaviour	Yes	Rejected	Nominal
Payment_of_Min_Amount	Yes	Rejected	Nominal
REP_Age	Default	Input	Interval
REP_Amount_invested_monthly	Default	Input	Interval
REP_Annual_Income	Default	Input	Interval
REP_Credit_History_Age	Default	Input	Interval
REP_Credit_Mix	Default	Input	Ordinal
REP_Credit_Score	Default	Target	Ordinal
REP_Credit_Utilization_Ratio	Default	Input	Interval
REP_Customer_ID	Yes	Input	Nominal
REP_Delay_from_due_date	Default	Input	Interval
REP_Interest_Rate	Default	Input	Interval
REP_Month	Default	Input	Nominal
REP_Monthly_Balance	Default	Input	Interval
REP_Monthly_Inhand_Salary	Default	Input	Interval
REP_Num_Bank_Accounts	Default	Input	Interval
REP_Num_Credit_Card	Default	Input	Interval
REP_Num_Credit_Inquiries	Default	Input	Interval
REP_Num_of_Delayed_Payment	Default	Input	Interval
REP_Num_of_Loan	Default	Input	Interval
REP_Occupation	Default	Input	Nominal
REP_Outstanding_Debt	Default	Input	Interval
REP_Payment_Behaviour	Default	Input	Nominal
REP_Payment_of_Min_Amount	Default	Input	Nominal
REP_Total_EMI_per_month	Default	Input	Interval
REP_auto_loan_f	Default	Input	Binary
REP_creditbuilder_loan_f	Default	Input	Binary
REP_debt_con_loan_f	Default	Input	Binary
REP_home_equi_loan_f	Default	Input	Binary
REP_mg_loan_f	Default	Input	Binary
REP_not_spec_loan_f	Default	Input	Binary
REP_payday_loan_f	Default	Input	Binary
REP_personal_loan_f	Default	Input	Binary
REP_student_loan_f	Default	Input	Binary
Total_EMI_per_month	Yes	Rejected	Interval
auto_loan_f	Default	Rejected	Binary
creditbuilder_loan_f	Default	Rejected	Binary
debt_con_loan_f	Default	Rejected	Binary
home_equi_loan_f	Default	Rejected	Binary
id	Yes	ID	Nominal
mg_loan_f	Default	Rejected	Binary
not_spec_loan_f	Default	Rejected	Binary
payday_loan_f	Default	Rejected	Binary
personal_loan_f	Default	Rejected	Binary
student_loan_f	Default	Rejected	Binary

Figure 7.2.5: Dropped variables

4. The missing values for class and interval variables are impute using the count and median respectively as shown in Figure 7.2.6 using the Impute node (Figures 12.5.16 - 12.5.19). The results after imputation are displayed in Figure 7.2.7 as below.

Name	Use	Method	Use Tree	Role	Level
Changed_Credit_Limit	**Default**	**Median**	**Default**	Input	Interval
REP_Age	**Default**	**Median**	**Default**	Input	Interval
REP_Amount_invested_monthly	**Default**	**Median**	**Default**	Input	Interval
REP_Annual_Income	**Default**	**Median**	**Default**	Input	Interval
REP_Credit_History_Age	**Default**	**Median**	**Default**	Input	Interval
REP_Credit_Mix	**Default**	**Count**	**Default**	Input	Ordinal
REP_Credit_Score	**Default**	**Default**	**Default**	Target	Ordinal
REP_Credit_Utilization_Ratio	**Default**	**Median**	**Default**	Input	Interval
REP_Delay_from_due_date	**Default**	**Median**	**Default**	Input	Interval
REP_Interest_Rate	**Default**	**Median**	**Default**	Input	Interval
REP_Month	**Default**	**Count**	**Default**	Input	Nominal
REP_Monthly_Balance	**Default**	**Median**	**Default**	Input	Interval
REP_Monthly_Inhand_Salary	**Default**	**Median**	**Default**	Input	Interval
REP_Num_Bank_Accounts	**Default**	**Median**	**Default**	Input	Interval
REP_Num_Credit_Card	**Default**	**Median**	**Default**	Input	Interval
REP_Num_Credit_Inquiries	**Default**	**Median**	**Default**	Input	Interval
REP_Num_of_Delayed_Payment	**Default**	**Median**	**Default**	Input	Interval
REP_Num_of_Loan	**Default**	**Median**	**Default**	Input	Interval
REP_Occupation	**Default**	**Count**	**Default**	Input	Nominal
REP_Outstanding_Debt	**Default**	**Median**	**Default**	Input	Interval
REP_Payment_Behaviour	**Default**	**Count**	**Default**	Input	Nominal
REP_Payment_of_Min_Amount	**Default**	**Count**	**Default**	Input	Nominal
REP_Total_EMI_per_month	**Default**	**Median**	**Default**	Input	Interval
REP_auto_loan_f	**Default**	**Default**	**Default**	Input	Binary
REP_creditbuilder_loan_f	**Default**	**Default**	**Default**	Input	Binary
REP_debt_con_loan_f	**Default**	**Default**	**Default**	Input	Binary
REP_home_equi_loan_f	**Default**	**Default**	**Default**	Input	Binary
REP_mg_loan_f	**Default**	**Default**	**Default**	Input	Binary
REP_not_spec_loan_f	**Default**	**Default**	**Default**	Input	Binary
REP_payday_loan_f	**Default**	**Default**	**Default**	Input	Binary
REP_personal_loan_f	**Default**	**Default**	**Default**	Input	Binary
REP_student_loan_f	**Default**	**Default**	**Default**	Input	Binary

Figure 7.2.6: Imputing class and interval variables

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label ▲	Number of Missing for TRAIN
Changed_Credit_Limit	MEDIAN	IMP_Changed_Credit_Limit	8.76	INPUT	INTERVAL		653
REP_Age	MEDIAN	IMP_REP_Age	34	INPUT	INTERVAL	Replacement: Age	2432
REP_Amount_invested_monthly	MEDIAN	IMP_REP_Amount_invested_monthly	138.76	INPUT	INTERVAL	Replacement: Amount invested monthly	1331
REP_Annual_Income	MEDIAN	IMP_REP_Annual_Income	38624.06	INPUT	INTERVAL	Replacement: Annual Income	0
REP_auto_loan_f	COUNT	IMP_REP_auto_loan_f	0	INPUT	BINARY	Replacement: auto loan f	3630
REP_Credit_History_Age	MEDIAN	IMP_REP_Credit_History_Age	18.8	INPUT	INTERVAL	Replacement: Credit History Age	2729
REP_Credit_Mix	COUNT	IMP_REP_Credit_Mix	Good	INPUT	ORDINAL	Replacement: Credit Mix	6041
REP_Credit_Utilization_Ratio	MEDIAN	IMP_REP_Credit_Utilization_Ratio	32.31	INPUT	INTERVAL	Replacement: Credit Utilization Ratio	0
REP_creditbuilder_loan_f	COUNT	IMP_REP_creditbuilder_loan_f	0	INPUT	BINARY	Replacement: creditbuilder loan f	3630
REP_debt_con_loan_f	COUNT	IMP_REP_debt_con_loan_f	0	INPUT	BINARY	Replacement: debt con loan f	3630
REP_Delay_from_due_date	MEDIAN	IMP_REP_Delay_from_due_date	16	INPUT	INTERVAL	Replacement: Delay from due date	236
REP_home_equi_loan_f	COUNT	IMP_REP_home_equi_loan_f	0	INPUT	BINARY	Replacement: home equi loan f	3630
REP_Interest_Rate	MEDIAN	IMP_REP_Interest_Rate	12	INPUT	INTERVAL	Replacement: Interest Rate	629
REP_mg_loan_f	COUNT	IMP_REP_mg_loan_f	0	INPUT	BINARY	Replacement: mg loan f	3630
REP_Month	COUNT	IMP_REP_Month	6	INPUT	NOMINAL	Replacement: Month	0
REP_Monthly_Balance	MEDIAN	IMP_REP_Monthly_Balance	341.245	INPUT	INTERVAL	Replacement: Monthly Balance	376
REP_Monthly_Inhand_Salary	MEDIAN	IMP_REP_Monthly_Inhand_Salary	3174.005	INPUT	INTERVAL	Replacement: Monthly Inhand Salary	4530
REP_not_spec_loan_f	COUNT	IMP_REP_not_spec_loan_f	0	INPUT	BINARY	Replacement: not spec loan f	3630
REP_Num_Bank_Accounts	MEDIAN	IMP_REP_Num_Bank_Accounts	5	INPUT	INTERVAL	Replacement: Num Bank Accounts	2077
REP_Num_Credit_Card	MEDIAN	IMP_REP_Num_Credit_Card	5	INPUT	INTERVAL	Replacement: Num Credit Card	677
REP_Num_Credit_Inquiries	MEDIAN	IMP_REP_Num_Credit_Inquiries	5	INPUT	INTERVAL	Replacement: Num Credit Inquiries	561
REP_Num_of_Delayed_Payment	MEDIAN	IMP_REP_Num_of_Delayed_Payment	13	INPUT	INTERVAL	Replacement: Num of Delayed Payment	2275
REP_Num_of_Loan	MEDIAN	IMP_REP_Num_of_Loan	3	INPUT	INTERVAL	Replacement: Num of Loan	1284
REP_Occupation	COUNT	IMP_REP_Occupation	Scientist	INPUT	NOMINAL	Replacement: Occupation	2094
REP_Outstanding_Debt	MEDIAN	IMP_REP_Outstanding_Debt	1151.22	INPUT	INTERVAL	Replacement: Outstanding Debt	0
REP_payday_loan_f	COUNT	IMP_REP_payday_loan_f	0	INPUT	BINARY	Replacement: payday loan f	3630
REP_Payment_Behaviour	COUNT	IMP_REP_Payment_Behaviour	Low_Spent_Small_Value_Payments	INPUT	NOMINAL	Replacement: Payment Behaviour	2348
REP_Payment_of_Min_Amount	COUNT	IMP_REP_Payment_of_Min_Amount	Yes	INPUT	NOMINAL	Replacement: Payment of Min Amount	3615
REP_personal_loan_f	COUNT	IMP_REP_personal_loan_f	0	INPUT	BINARY	Replacement: personal loan f	3630
REP_student_loan_f	COUNT	IMP_REP_student_loan_f	0	INPUT	BINARY	Replacement: student loan f	3630
REP_Total_EMI_per_month	MEDIAN	IMP_REP_Total_EMI_per_month	70.6	INPUT	INTERVAL	Replacement: Total EMI per month	0

Figure 7.2.7: Imputation Summary

5. The "Standard" and "Poor" classes in the target variable "REP_Credit_Score" are grouped together into "Standard to Poor" using the Replacement node as shown in Figure 7.2.8. This is to increase the performance of the machine learning model in classifying customers' credit score. (Figure 12.5.20)

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
REP_Credit_Score	Good		10000	C	Good	.
REP_Credit_Score	Poor	**Standard to Poor**	10000	C	Poor	.
REP_Credit_Score	Standard	**Standard to Poor**	10000	C	Standard	.
REP_Credit_Score	_UNKNOWN_	**_DEFAULT_**	.	C		.

Figure 7.2.8: Recategorizing target variable

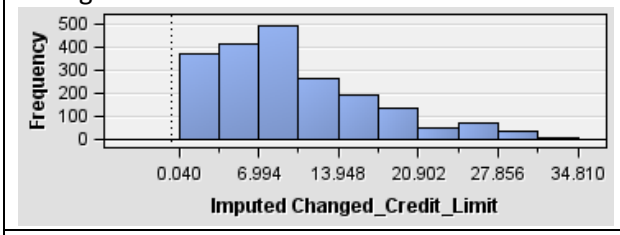
6. The interval variables are normalized using the log 10 transformation in the Transform node as shown in Figure 7.2.9 to transform all features into the same scale. (Figures 12.5.21 - 12.5.24)

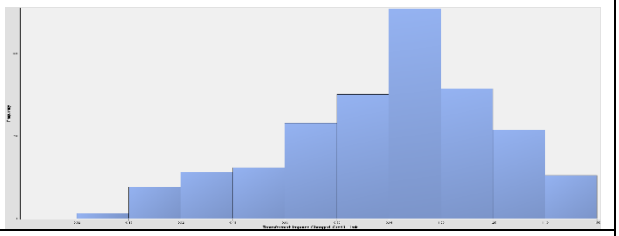
Name	Method	Number of Bins	Role	Level
IMP_Changed_Credit_Limit	**Default**	**4**	Rejected	Interval
IMP_REP_Age	**Default**	**4**	Rejected	Interval
IMP_REP_Amount_invested_monthly	**Default**	**4**	Rejected	Interval
IMP_REP_Annual_Income	**Default**	**4**	Rejected	Interval
IMP_REP_Credit_History_Age	**Default**	**4**	Rejected	Interval
IMP_REP_Credit_Mix	**Default**	**4**	Input	Ordinal
IMP_REP_Credit_Utilization_Ratio	**Default**	**4**	Rejected	Interval
IMP_REP_Delay_from_due_date	**Default**	**4**	Rejected	Interval
IMP_REP_Interest_Rate	**Default**	**4**	Rejected	Interval
IMP_REP_Month	**Default**	**4**	Input	Nominal
IMP_REP_Monthly_Balance	**Default**	**4**	Rejected	Interval
IMP_REP_Monthly_Inhand_Salary	**Default**	**4**	Rejected	Interval
IMP_REP_Num_Bank_Accounts	**Default**	**4**	Rejected	Interval
IMP_REP_Num_Credit_Card	**Default**	**4**	Rejected	Interval
IMP_REP_Num_Credit_Inquiries	**Default**	**4**	Rejected	Interval
IMP_REP_Num_of_Delayed_Payment	**Default**	**4**	Rejected	Interval
IMP_REP_Num_of_Loan	**Default**	**4**	Rejected	Interval
IMP_REP_Occupation	**Default**	**4**	Input	Nominal
IMP_REP_Outstanding_Debt	**Default**	**4**	Rejected	Interval
IMP_REP_Payment_Behaviour	**Default**	**4**	Input	Nominal
IMP_REP_Payment_of_Min_Amount	**Default**	**4**	Input	Nominal
IMP_REP_Total_EMI_per_month	**Default**	**4**	Rejected	Interval
IMP_REP_auto_loan_f	**Default**	**4**	Input	Binary
IMP_REP_creditbuilder_loan_f	**Default**	**4**	Input	Binary
IMP_REP_debt_con_loan_f	**Default**	**4**	Input	Binary
IMP_REP_home_equi_loan_f	**Default**	**4**	Input	Binary
IMP_REP_mg_loan_f	**Default**	**4**	Input	Binary
IMP_REP_not_spec_loan_f	**Default**	**4**	Input	Binary
IMP_REP_payday_loan_f	**Default**	**4**	Input	Binary
IMP_REP_personal_loan_f	**Default**	**4**	Input	Binary
IMP_REP_student_loan_f	**Default**	**4**	Input	Binary
REP_Credit_Score	**Default**	**4**	Rejected	Ordinal
REP_IMP_Changed_Credit_Limit	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Age	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Amount_invested_mont	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Annual_Income	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Credit_History_Age	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Credit_Utilization_R	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Delay_from_due_date	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Interest_Rate	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Monthly_Balance	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Monthly_Inhand_Salar	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Num_Bank_Accounts	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Num_Credit_Card	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Num_Credit_Inquiries	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Num_of_Delayed_Payme	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Num_of_Loan	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Outstanding_Debt	**Log 10**	**4**	Input	Interval
REP_IMP_REP_Total_EMI_per_month	**Log 10**	**4**	Input	Interval
REP_REP_Credit_Score	**Default**	**4**	Target	Ordinal

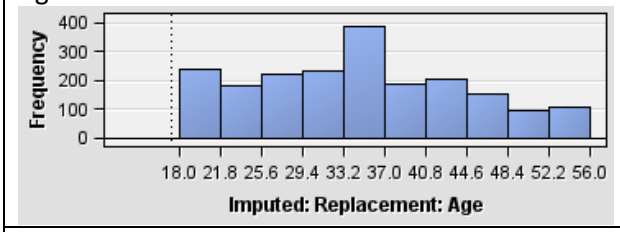
Figure 7.2.9: Log 10 transformation

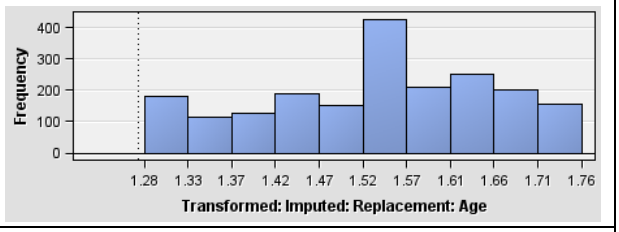
Figure 7.2.10 below shows the comparison of before and after the log 10 transformation normalization approach for some variables.

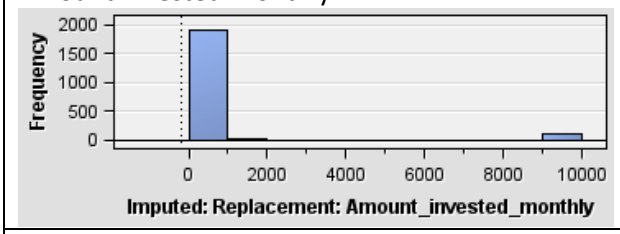
Before Transformation	After Transformation
Change in Credit Limit	
Age	
Amount Invested Monthly	
Annual Income	
Credit History Age	
Credit Utilization Ratio	
Delay from Due Date	

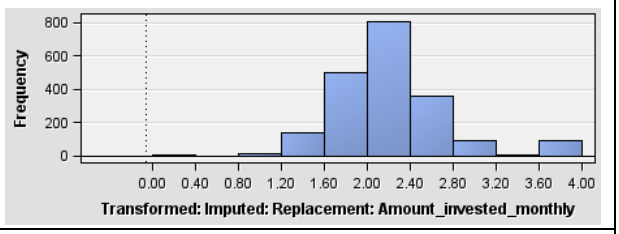


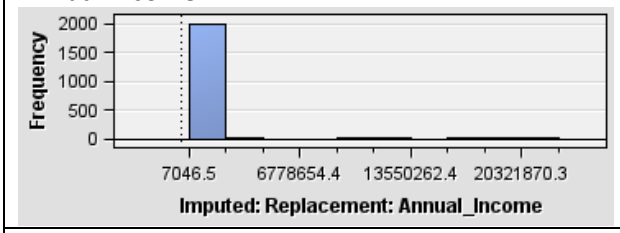


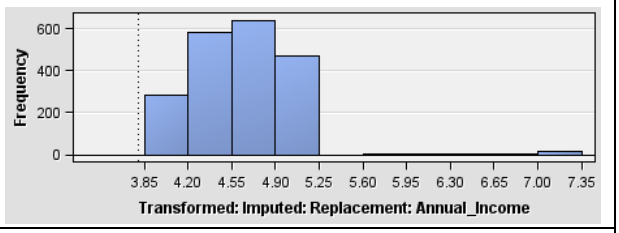


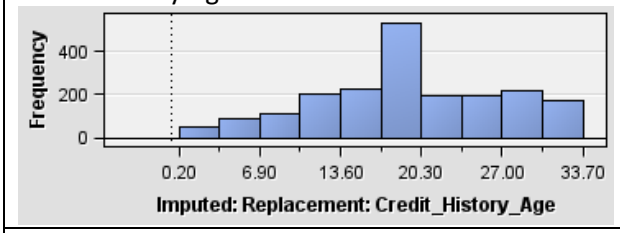


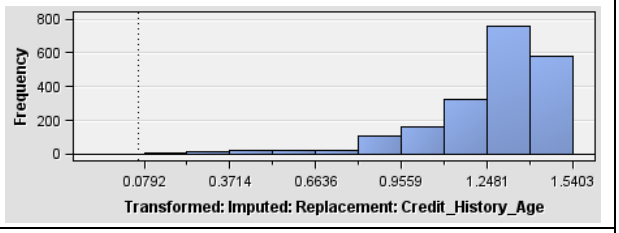


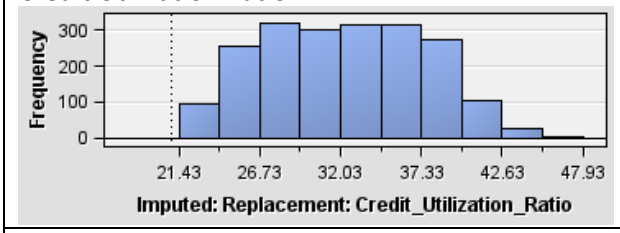


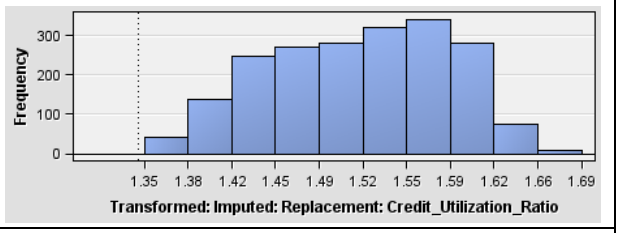


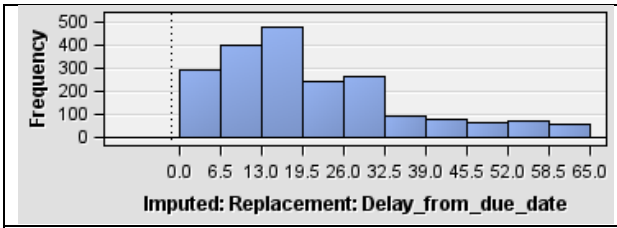


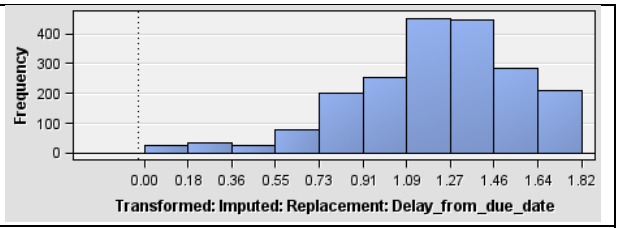




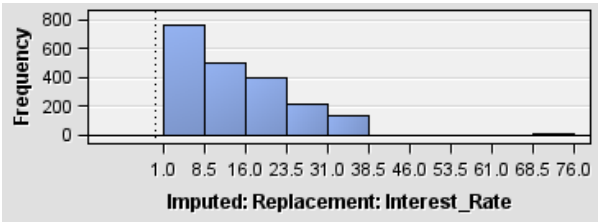


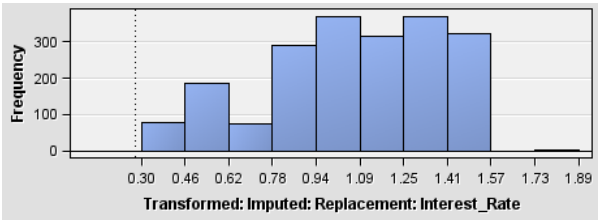




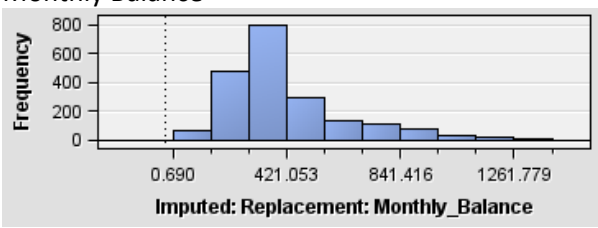


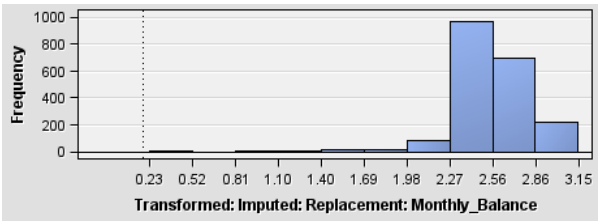
Interest Rate



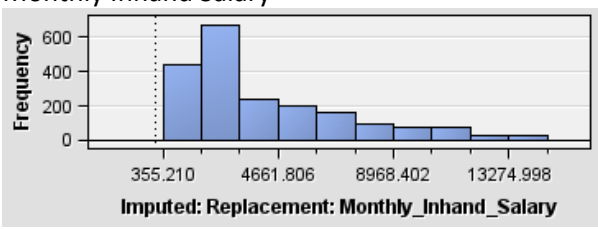


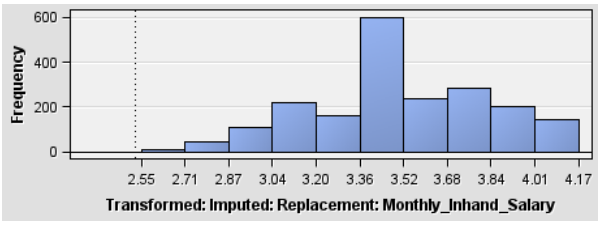
Monthly Balance



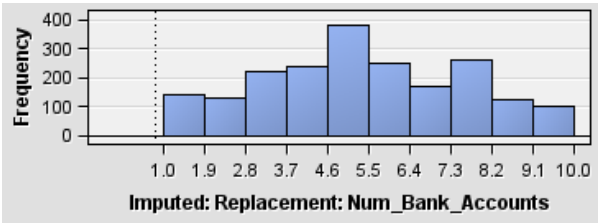


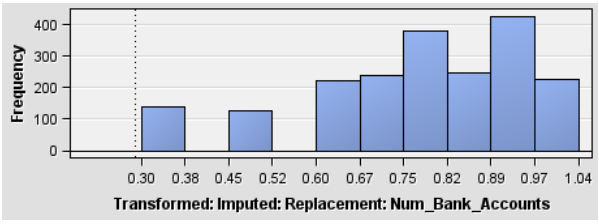
Monthly Inhand Salary



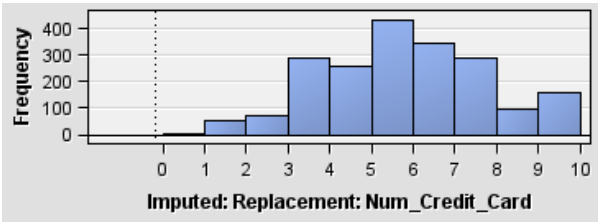


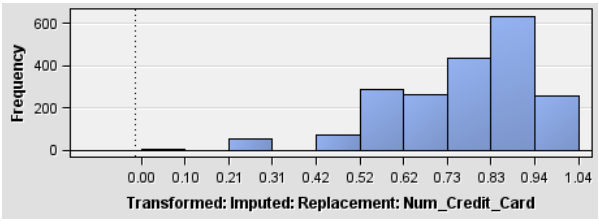
Number of Bank Accounts





Number of Credit Card





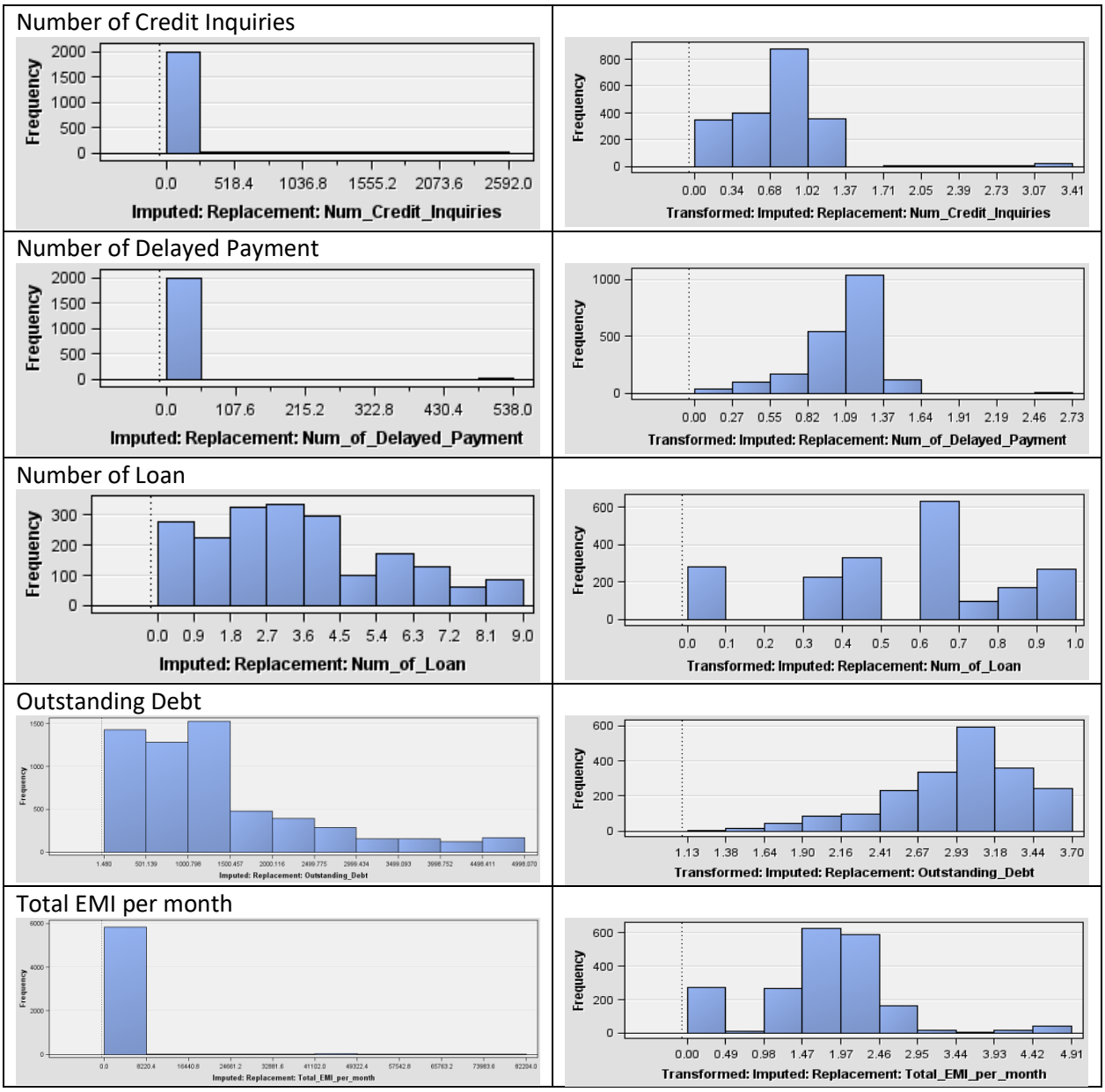


Figure 7.2.10: Log 10 transformation graphs comparison

A new dataset with new variables is formed, as shown in Figure 7.2.11. As observed, there are 31 Input role variables, with different levels that consist of 4 nominal variables, 17 interval variables, 1 ordinal variable, and 9 binary variables while 1 ordinal variable as target role in the dataset.

Name	Sample Role	Role	Level
IMP_REP_Credit_Mix	**Default**	Input	Ordinal
IMP_REP_Month	**Default**	Input	Nominal
IMP_REP_Occupation	**Default**	Input	Nominal
IMP_REP_Payment_Behaviour	**Default**	Input	Nominal
IMP_REP_Payment_of_Min_Amount	**Default**	Input	Nominal
IMP_REP_auto_loan_f	**Default**	Input	Binary
IMP_REP_creditbuilder_loan_f	**Default**	Input	Binary
IMP_REP_debt_con_loan_f	**Default**	Input	Binary
IMP_REP_home_equi_loan_f	**Default**	Input	Binary
IMP_REP_mg_loan_f	**Default**	Input	Binary
IMP_REP_not_spec_loan_f	**Default**	Input	Binary
IMP_REP_payday_loan_f	**Default**	Input	Binary
IMP_REP_personal_loan_f	**Default**	Input	Binary
IMP_REP_student_loan_f	**Default**	Input	Binary
LG10_REP_IMP_Changed_Credit_Limi	**Default**	Input	Interval
LG10_REP_IMP_REP_Age	**Default**	Input	Interval
LG10_REP_IMP_REP_Amount_invested	**Default**	Input	Interval
LG10_REP_IMP_REP_Annual_Income	**Default**	Input	Interval
LG10_REP_IMP_REP_Credit_History_	**Default**	Input	Interval
LG10_REP_IMP_REP_Credit_Utilizat	**Default**	Input	Interval
LG10_REP_IMP_REP_Delay_from_due_	**Default**	Input	Interval
LG10_REP_IMP_REP_Interest_Rate	**Default**	Input	Interval
LG10_REP_IMP_REP_Monthly_Balance	**Default**	Input	Interval
LG10_REP_IMP_REP_Monthly_Inhand_	**Default**	Input	Interval
LG10_REP_IMP_REP_Num_Bank_Accoun	**Default**	Input	Interval
LG10_REP_IMP_REP_Num_Credit_Card	**Default**	Input	Interval
LG10_REP_IMP_REP_Num_Credit_Inqu	**Default**	Input	Interval
LG10_REP_IMP_REP_Num_of_Delayed_	**Default**	Input	Interval
LG10_REP_IMP_REP_Num_of_Loan	**Default**	Input	Interval
LG10_REP_IMP_REP_Outstanding_Deb	**Default**	Input	Interval
LG10_REP_IMP_REP_Total_EMI_per_m	**Default**	Input	Interval
REP_REP_Credit_Score	**Default**	Target	Ordinal

Figure 7.2.11 Cleaned dataset

7. Data resampling is performed to ensure that the class distribution is equally distributed. As a result, there are 10,000 customers with good and standard to poor credit scores as shown in Figure 7.2.12. (Figures 12.5.25 - 12.5.26)

Summary Statistics for Class Targets
(maximum 500 observations printed)

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
REP_REP_Credit_Score	.	Good	10000	33.3333	Replacement: Replacement: Credit_Score
REP_REP_Credit_Score	.	Standard to Poor	20000	66.6667	Replacement: Replacement: Credit_Score

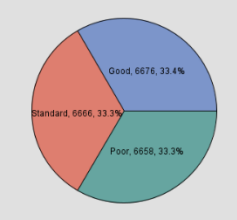
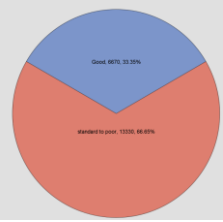
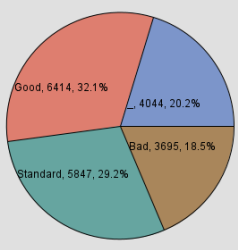
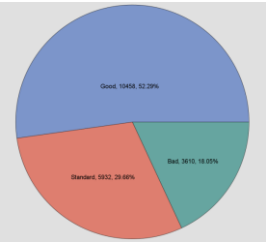
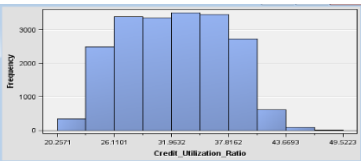
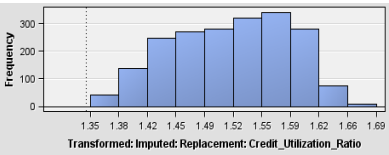
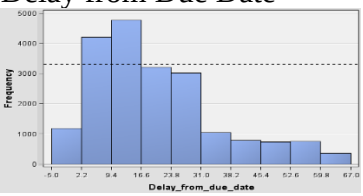
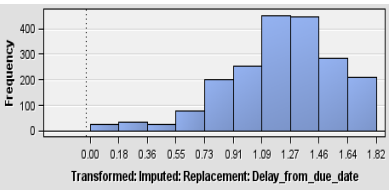
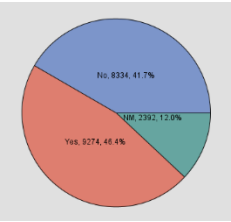
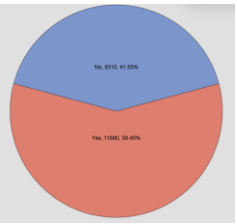
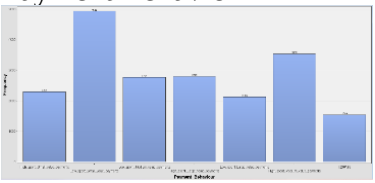
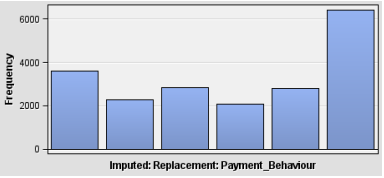
Data=SAMPLE

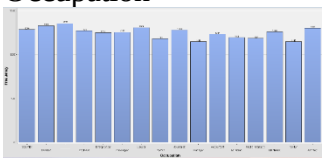
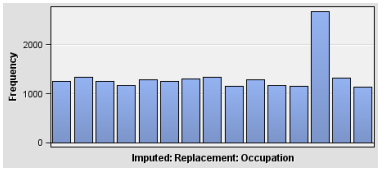
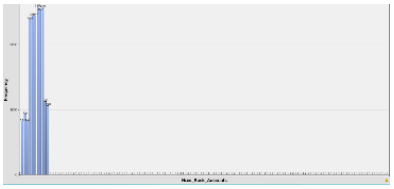
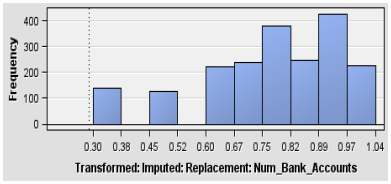
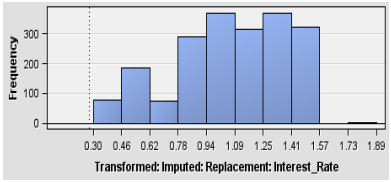
Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
REP_REP_Credit_Score	.	Good	10000	50	Replacement: Replacement: Credit_Score
REP_REP_Credit_Score	.	Standard to Poor	10000	50	Replacement: Replacement: Credit_Score

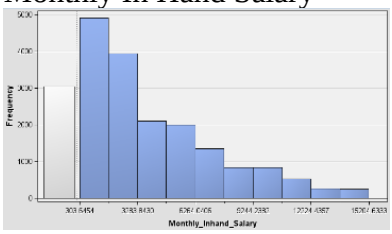
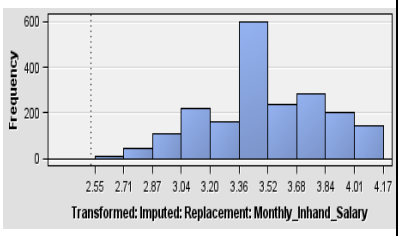
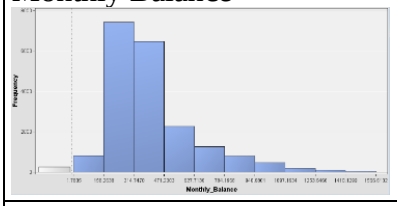
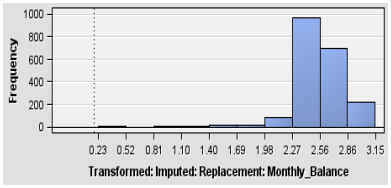
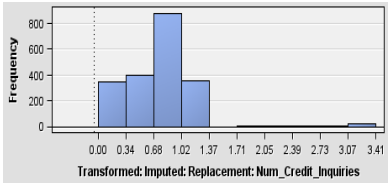
Figure 7.2.12: Sample Dataset

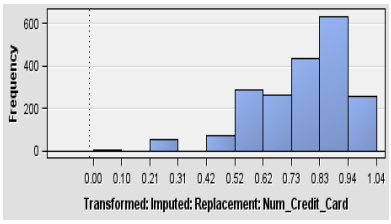
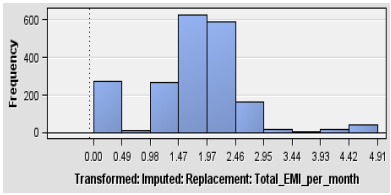
7.3 Data Exploration on Cleaned Dataset

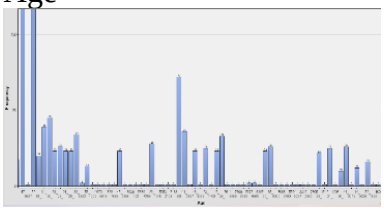
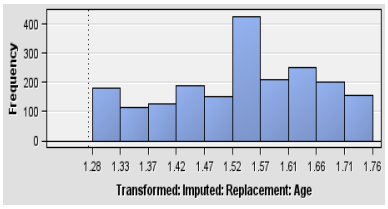
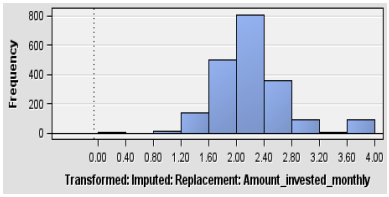
After data cleaning, each of the variables is explored as showcased in Figure 5.3 below.

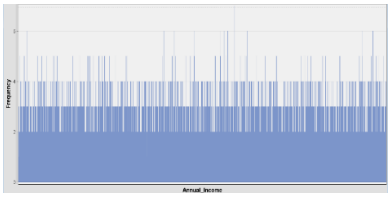
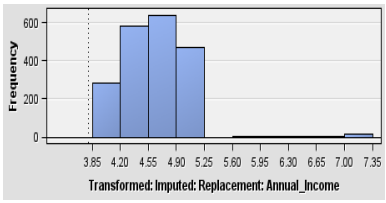
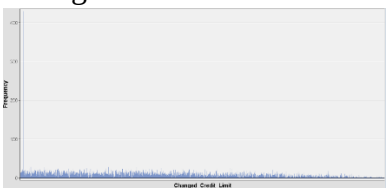
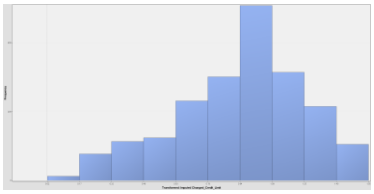
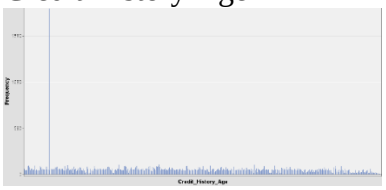
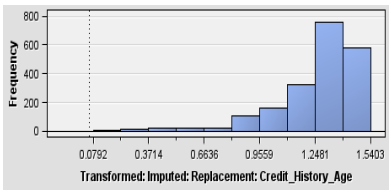
Variables - Before Modify	After Modify	Explanation
Credit Score 		The labels of target variable have combined 'standard' and 'poor' to form a new label, named 'Standard to Poor'. This provides a higher accuracy when predicting the outcome based on the independent class variables.
Credit Mix 		Credit Mix variable was imputed with Count value, resulting in no Missing value in the dataset.
Credit Utilization Ratio 		Normalization using log10 transformation is applied to Credit Utilization Ratio variable to scale down the dataset.
Delay from Due Date 		Normalization using log10 transformation is applied to Delay from Due Date variable to scale down the dataset.
Payment of Minimum Amount 		Payment of Minimum Amount is imputed with Count value, leaving no missing values in the dataset.
Payment Behavior 		Payment behavior with invalid values was imputed with Count value, i.e. Low Spent Small Value Payment, eliminating those invalid values.

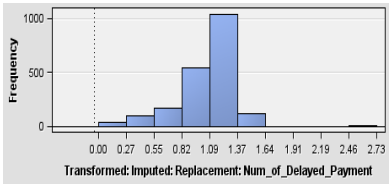
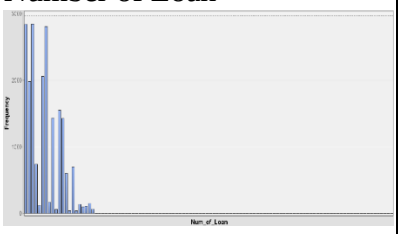
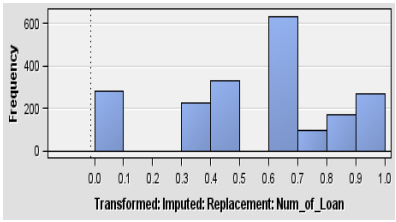
Occupation 		Unknown values from occupation were imputed with Count value, i.e. Scientist, eliminating those unknown values.
Number of Bank Accounts 		Dirty and noisy data are cleaned up by formatting character values into numeric, imputing based on median value and normalizing using log 10 transformation.
Interest Rate 		Dirty and noisy data are cleaned up by setting replacement upper limit of 100% using missing as replace method, followed by imputing data using median value and lastly to normalize data using log 10 transformation approach.

Monthly In Hand Salary 		Missing values are imputed using median value, thus showing no missing value after the imputation. Data is normalized to form a smooth bell-curved shape distribution.
Monthly Balance 		Missing value found in monthly balance is imputed based on median value and normalized using log 10 transformation.
Number of Credit Inquiries 		Missing value found in number of credit inquiries is imputed based on median value and normalized using log 10 transformation.
Number of Credit Cards		

		Outliers found in number of credit cards were removed by setting replacement upper limit value of 15 with missing replace method. It is then imputed using median value and normalized using log 10 transformation.
Total EMI per Month 		Total EMI was set with replacement lower limit value of 0 with missing replace method. It is imputed using median value, followed by normalisation.

Age 		Dirty data is formatted from character to numeric value. It is then set with replacement lower limit of 18 and upper limit of 75, that is the age limit to apply for loans. The data is then imputed using median value and is normalized.
Amount Invested Monthly 		Dirty data is formatted from character to numeric value. It is then set with replacement lower limit of 0 as amount cannot be less than 0. The data is then imputed using median value and is normalized.
Annual Income		Dirty data is formatted from character to numeric value. Annual income is then set with replacement lower limit of 0 as

		amount cannot be less than 0. The data is then imputed using median value and is normalized.
Change Credit Limit 		Dirty data is formatted from character to numeric value. The variable is then imputed using median value and is normalized, forming a bell-curve shape distribution.
Credit History Age 		Dirty data is formatted from character to numeric value. It is then set with a replacement lower limit of 0. The data is then imputed using median value and is normalized, forming a left-skewed distribution.

Number of Delayed Payment 		Dirty data is formatted from character to numeric value. It is then set with replacement lower limit of 0 and upper limit of 700 as data beyond 700 is deemed as outliers. The data is then imputed using median value and is normalized.
Number of Loan 		Dirty data is formatted from character to numeric value. It is then set with a replacement lower limit of 0 and upper limit of 15. The data is then imputed using median value and is normalized.
Outstanding Debt		Dirty data is formatted from character to numeric value. It is then set with replacement lower limit of 0 as amount cannot be less

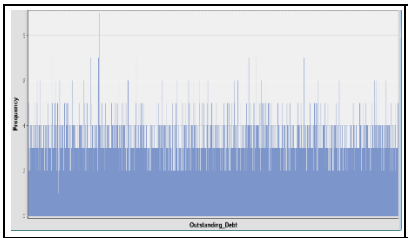
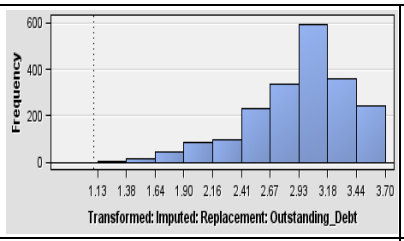
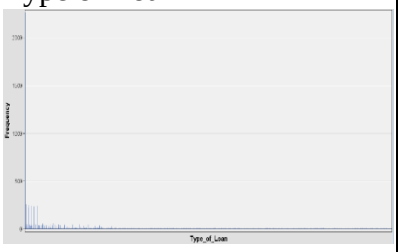
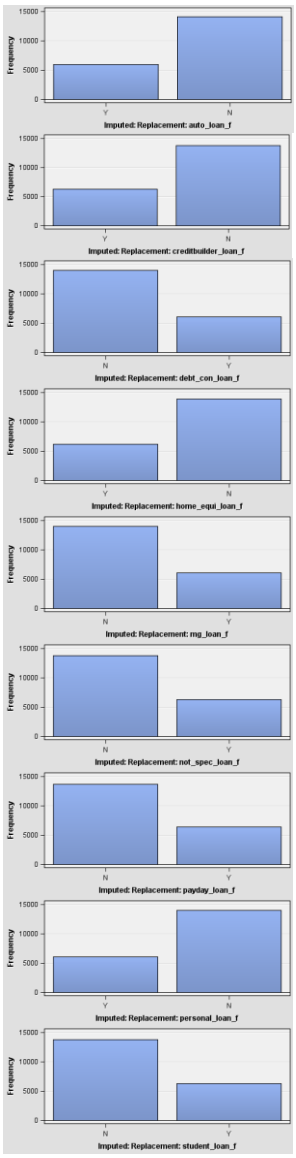
		than 0. The outstanding debt variable is then imputed using median value and is normalized, forming a left-skewed distribution.
Type of Loan 		The dirty data is cleaned up, and different types of loans are broken down into different loan flags, such as auto loan, credit builder loan, personal loan, student loan etc., forming 9 new variables.

Figure 5.3: Comparison of each variable (before VS after data cleaning)

The relationship between the input variables is explored through correlation matrix under Figure 5.4. It can be seen that annual income and monthly inhand salary has a relative high positive correlation of 0.76, followed by number of loan and total EMI per month, which is 0.59. Delay from due date and number of bank accounts has a positive correlation of 0.49. From the negative correlation view, interest rate and credit history age correlation value is -0.48, followed by number

of loan and credit history age with correlation value of -0.46. Full correlation matrix table as shown in Appendix (Figure 12.5.27)

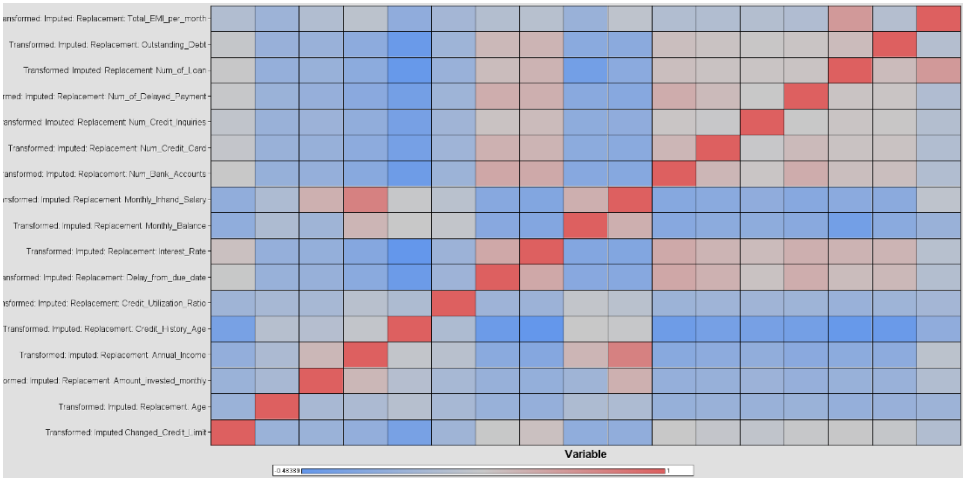


Figure 5.4 Correlation Matrix

Looking from the chi-square bar plots in Figure 5.5, Credit Mix and Payment of Minimum Amount variables show a strong significant association, which might have a potentially meaningful relationship with the Credit Score target variable. Meanwhile, Occupation and Month shows a weaker association between the variables.

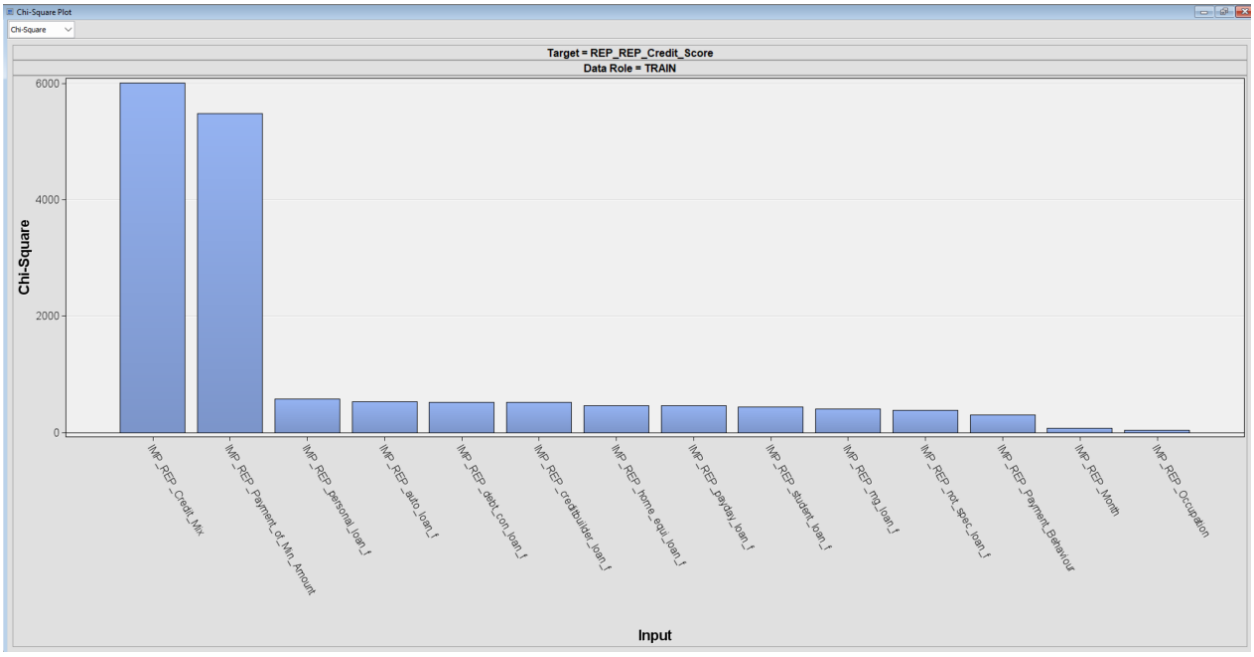


Figure 5.5 Chi-Square Bar Plots between input variables VS target variable

7.4 Data Partitioning

The data partition node was used to split the dataset into training and validation sets at a split ratio of 50:50 as shown in Figure 7.4.1, along with the results in Figure 7.4.2. (Figures 12.6.1 - 12.6.3 show the steps to partition the data.)

Property	Value
General	
Node ID	Part2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes

Figure 7.4.1: Data Partition Ratio

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
REP_REP_Credit_Score	.	Good	10000	50	Replacement: Replacement: Credit_Score
REP_REP_Credit_Score	.	Standard to Poor	10000	50	Replacement: Replacement: Credit_Score

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
REP_REP_Credit_Score	.	Good	5000	50	Replacement: Replacement: Credit_Score
REP_REP_Credit_Score	.	Standard to Poor	5000	50	Replacement: Replacement: Credit_Score

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
REP_REP_Credit_Score	.	Good	5000	50	Replacement: Replacement: Credit_Score
REP_REP_Credit_Score	.	Standard to Poor	5000	50	Replacement: Replacement: Credit_Score

Figure 7.4.2: Data Partition Output

8 SEMMA: Model

The training dataset consists of 50% of the data used for the model building. The models used in this project to classify the credit score are the Decision Tree (DT), Logistic Regression (LR), Gradient Boosting (GB), and Least-Angle Regression (LARS).

8.1 Decision Tree Model

A supervised learning technique which is preferred to solve classification problems is known as the Decision Tree (DT). It is a tree-like structure with root nodes, interval nodes, leaf nodes and branches. Internal node is the representation of the features, the branch is the representation of the decision rules, and the leaf node is the representation of the outcomes. The Decision Tree usually mimics the human thinking ability to make decisions, hence it is easy to understand.

The Decision Tree built using our dataset is shown in Figure 8.1.1 below.

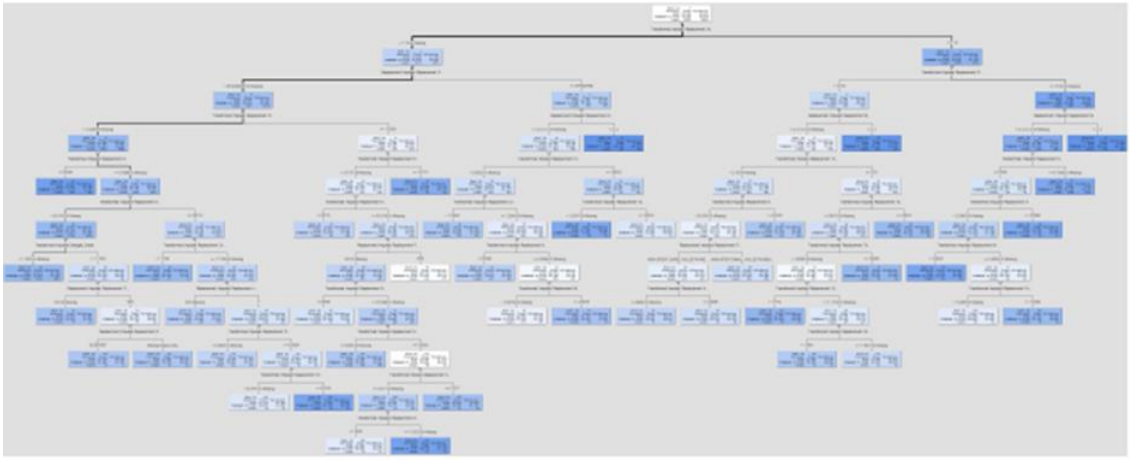


Figure 8.1.1: Decision Tree

Some interesting points while developing the Decision Tree are:

1. The interest rate has the highest level of information gain. Thus, it is selected as the root node.
2. The Total EMI per month feature has the highest number of splitting.
3. Mostly if the interest rate is lesser, then the credit score is ‘standard to poor’.

Below are the decision rules:

- If Interest Rate < 1.13004 or Missing, Month is one of 1, 3, 2 and Credit Mix >= STANDARD then Credit Score = Standard to Poor
- If Outstanding Debt < 3.10396, Interest Rate >= 1.13004 and Month IS ONE OF 1, 3, 2 then Credit Score = Standard to Poor
- If Outstanding Debt >= 3.10396 or Missing, Interest Rate >= 1.13004 and Month is one of 1, 3, 2 then Credit Score = Standard to Poor
- If Num Credit Card < 0.53959, Interest Rate < 1.13004 or Missing, Delay from due date < 1.24286 or Missing and Credit Mix <= GOOD or Missing then Credit Score = Good
- If Outstanding Debt >= 3.17146, Interest Rate < 1.13004 or Missing, Delay from due date >= 1.24286 and Credit Mix <= GOOD or Missing then Credit Score = Standard to Poor
- If Outstanding Debt >= 3.10396 or Missing, Num Credit Inquiries >= 0.73856 or Missing, Interest Rate >= 1.13004 and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score=Standard to Poor
- If Outstanding Debt < 3.17146 or Missing, Num of Delayed Payment < 0.81162, Interest Rate < 1.13004 or Missing, Delay from due date >= 1.24286 and Credit Mix <= GOOD or Missing then Credit Score = Standard to Poor
- If Total EMI per month < 2.68316 or Missing, Interest Rate < 1.13004 or Missing, Credit History Age < 1.28443, Month is one of 7, 4, 6, 5, 8 or Missing and Credit Mix >= STANDARD then Credit Score = Standard to Poor
- If Total EMI per month < 3.26778, Total EMI per month >= 2.68316 or Missing, Interest Rate < 1.13004 or Missing, Month is one of 7, 4, 6, 5, 8 or Missing and Credit Mix >= STANDARD then Credit Score = Good

- If Total EMI per month >= 3.26778, Interest Rate < 1.13004 or Missing, Month is one of 7, 4, 6, 5, 8 or Missing and Credit Mix >= STANDARD then Credit Score = Standard to Poor
- If Total EMI per month < 2.17995 or Missing, Outstanding Debt < 3.10396, Num Bank Accounts >= 0.87409, Interest Rate >= 1.13004 and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Standard to Poor
- If Total EMI per month >= 4.3478, Outstanding Debt < 3.10396, Interest Rate >= 1.13004 and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Standard to Poor
- If Outstanding Debt >= 3.10396 or Missing, Num Credit Inquiries < 0.73856, Interest Rate >= 1.13004, Amount invested monthly >= 2.59556 and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Good
- If Num Credit Card < 0.81162, Num Credit Card >= 0.53959 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date < 1.24286 or Missing, Changed Credit Limit < 1.12024 or Missing and Credit Mix <= GOOD or Missing then Credit Score = Good
- If Num Credit Card >= 0.81162, Interest Rate < 1.13004 or Missing, Delay from due date < 1.24286 or Missing, Credit History Age < 1.17492 and Credit Mix <= GOOD or Missing then Credit Score = Standard to Poor
- If Outstanding Debt < 3.17146 or Missing, Num of Delayed Payment >= 0.81162 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date >= 1.24286, Payment of Min Amount is one of YES and Credit Mix <= GOOD or Missing then Credit Score = Standard to Poor
- If Total EMI per month < 2.68316 or Missing, Num Credit Inquiries < 0.53959, Interest Rate < 1.13004 or Missing, Credit History Age >= 1.28443 or Missing, Month is one of 7, 4, 6, 5, 8 or Missing and Credit Mix >= STANDARD then Credit Score = Standard to Poor
- If Total EMI per month < 2.17995 or Missing, Outstanding Debt < 3.10396, Num Bank Accounts < 0.87409 or Missing, Interest Rate >= 1.13004, Payment Behaviour is one of HIGH SPENT SMALL VALUE PAYMENT, LOW SPENT SMALL VALUE PAYMENT, LOW SPENT MEDIUM VALUE PAYMENT, LOW SPENT LARGE VALUE PAYMENT or Missing and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Standard to Poor
- If Total EMI per month < 4.3478, Total EMI per month >= 2.49057, Outstanding Debt < 3.10396, Interest Rate >= 1.13004 and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Good
- If Outstanding Debt >= 3.10396 or Missing, Num Credit Inquiries < 0.73856, Monthly Inhand Salary < 3.48237, Interest Rate >= 1.13004, Amount invested monthly < 2.59556 or Missing and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Standard to Poor
- If Num Credit Card < 0.81162, Num Credit Card >= 0.53959 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date < 1.24286 or Missing, Changed Credit Limit >= 1.12024, Payment of Min Amount is one of NO or Missing and Credit Mix <= GOOD or Missing then Credit Score=Good
- If Num Credit Card >= 0.81162, Interest Rate < 1.13004 or Missing, Delay from due date < 1.24286 or Missing, Credit History Age >= 1.17492 or Missing, not_spec_loan_f is one of N or Missing and Credit Mix <= GOOD or Missing then Credit Score = Good
- If Outstanding Debt < 3.17146 or Missing, Num of Delayed Payment >= 0.81162 or Missing, Num Bank Accounts < 0.53959, Interest Rate < 1.13004 or Missing, Delay from

due date >= 1.24286, Payment of Min Amount is one of NO or Missing and Credit Mix <= GOOD or Missing then Credit Score = Standard to Poor

- If Total EMI per month < 2.68316 or Missing, Num Credit Inquiries >= 0.53959 or Missing, Monthly Inhand Salary < 3.96792 or Missing, Interest Rate < 1.13004 or Missing, Credit History Age >= 1.28443 or Missing, Month is one of 7, 4, 6, 5, 8 or Missing and Credit Mix >= STANDARD then Credit Score=Standard to Poor
- If Total EMI per month < 2.68316 or Missing, Num Credit Inquiries >= 0.53959 or Missing, Monthly Inhand Salary >= 3.96792, Interest Rate < 1.13004 or Missing, Credit History Age >= 1.28443 or Missing, Month is one of 7, 4, 6, 5, 8 or Missing and Credit Mix >= STANDARD then Credit Score = Good
- If Total EMI per month < 2.17995 or Missing, Outstanding Debt < 3.10396, Num Bank Accounts < 0.87409 or Missing, Interest Rate >= 1.13004, Annual Income < 4.64086 or Missing, Payment Behaviour is one of HIGH SPENT LARGE VALUE PAYMENT, HIGH SPENT MEDIUM VALUE PAYMENT and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Good
- If Total EMI per month < 2.17995 or Missing, Outstanding Debt < 3.10396, Num Bank Accounts < 0.87409 or Missing, Interest Rate >= 1.13004, Annual Income >= 4.64086, Payment Behaviour is one of HIGH SPENT LARGE VALUE PAYMENT, HIGH SPENT MEDIUM VALUE PAYMENT and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Standard to Poor
- If Total EMI per month < 2.49057, Total EMI per month >= 2.17995 or Missing, Outstanding Debt < 3.10396, Interest Rate >= 1.13004, Credit History Age < 1.17742 and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Standard to Poor
- If Total EMI per month < 1.82859 or Missing, Outstanding Debt >= 3.10396 or Missing, Num Credit Inquiries < 0.73856, Monthly Inhand Salary >= 3.48237 or Missing, Interest Rate >= 1.13004, Amount invested monthly < 2.59556 or Missing and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Good
- If Total EMI per month >= 1.82859, Outstanding Debt >= 3.10396 or Missing, Num Credit Inquiries < 0.73856, Monthly Inhand Salary >= 3.48237 or Missing, Interest Rate >= 1.13004, Amount invested monthly < 2.59556 or Missing and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Standard to Poor
- If Num Credit Card < 0.81162, Num Credit Card >= 0.53959 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date < 1.24286 or Missing, Changed Credit Limit >= 1.12024, Payment of Min Amount is one of YES, Occupation is one of SCIENTIST and Credit Mix <= GOOD or Missing then Credit Score = Good
- If Num Credit Card < 0.81162, Num Credit Card >= 0.53959 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date < 1.24286 or Missing, Changed Credit Limit >= 1.12024, Payment of Min Amount is one of YES, Occupation equals Missing and Credit Mix <= GOOD or Missing then Credit Score = Standard to Poor
- If Outstanding Debt < 3.09235 or Missing, Num Credit Card >= 0.81162, Interest Rate < 1.13004 or Missing, Delay from due date < 1.24286 or Missing, Credit History Age >= 1.17492 or Missing, not_spec_loan_f is one of Y and Credit Mix <= GOOD or Missing then Credit Score = Good
- If Outstanding Debt < 3.17146 or Missing, Num of Delayed Payment >= 0.81162 or Missing, Num Bank Accounts >= 0.53959 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date >= 1.24286, Annual Income < 5.03928 or Missing, Payment of Min

Amount is one of NO or Missing and Credit Mix <= GOOD or Missing then Credit Score=Good

- If Total EMI per month < 2.49057, Total EMI per month >= 2.17995 or Missing, Outstanding Debt < 3.10396, Interest Rate < 1.19011, Interest Rate >= 1.13004, Credit History Age >= 1.17742 or Missing and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Standard to Poor
- If Total EMI per month < 2.49057, Total EMI per month >= 2.17995 or Missing, Outstanding Debt < 3.10396, Interest Rate >= 1.19011 or Missing, Credit History Age >= 1.17742 or Missing and Month is one of 7, 4, 6, 5, 8 or Missing then Credit Score = Good
- If Outstanding Debt >= 3.09235, Num Credit Card >= 0.81162, Interest Rate < 0.87409 or Missing, Delay from due date < 1.24286 or Missing, Credit History Age >= 1.17492 or Missing, not_spec_loan_f is one of Y and Credit Mix <= GOOD or Missing then Credit Score=Good
- If Outstanding Debt >= 3.09235, Num Credit Card >= 0.81162, Interest Rate < 1.13004, Interest Rate >= 0.87409, Delay from due date < 1.24286 or Missing, Credit History Age >= 1.17492 or Missing, not_spec_loan_f is one of Y and Credit Mix <= GOOD or Missing then Credit Score = Standard to Poor
- If Outstanding Debt < 3.17146 or Missing, Num of Delayed Payment >= 0.81162 or Missing, Num Bank Accounts >= 0.53959 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date >= 1.24286, Annual Income >= 5.1377, Payment of Min Amount is one of NO or Missing and Credit Mix <= GOOD or Missing then Credit Score = Good
- If Outstanding Debt < 3.17146 or Missing, Num of Delayed Payment >= 0.81162 or Missing, Num Bank Accounts >= 0.53959 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date < 1.33232, Delay from due date >= 1.24286, Annual Income < 5.1377, Annual Income >= 5.03928 or Missing, Payment of Min Amount is one of NO or Missing and Credit Mix <= GOOD or Missing then Credit Score = Good
- If Outstanding Debt < 3.17146 or Missing, Num of Delayed Payment >= 0.81162 or Missing, Num Bank Accounts >= 0.53959 or Missing, Interest Rate < 1.13004 or Missing, Delay from due date >= 1.33232 or Missing, Annual Income < 5.1377, Annual Income >= 5.03928 or Missing, Payment of Min Amount is one of NO or Missing and Credit Mix <= GOOD or Missing then Credit Score = Standard to Poor

The decision tree model is built using the training dataset and some of the properties are set accordingly to archive the best accuracy. The adjusted properties are shown in Figure 8.1.2 below.

.. Property	Value
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	1.0
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	10
Minimum Categorical Size	5
Node	
Leaf Size	10
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000

Figure 8.1.2: Adjusted Property of Decision Tree

The misclassification rate of the validation set is 0.1599 which gives an accuracy of 84.01% as outlined in Figure 8.1.3.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_NOBS_	Sum of Frequencies	10000	10000	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MISC_	Misclassification Rate	0.1521	0.1599	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MAX_	Maximum Absolute Error	0.996094	0.996094	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_SSE_	Sum of Squared Errors	2314.401	2398.876	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_ASE_	Average Squared Error	0.11572	0.119944	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_RASE_	Root Average Squared Error	0.340176	0.346329	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DIV_	Divisor for ASE	20000	20000	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DFT_	Total Degrees of Freedom	10000		

Figure 8.1.3: Statistical Output of Decision Tree

Based on Figure 8.1.4, the three variables with the highest important values are the Interest rate with 1.0000, Credit Mix with 0.5454, and Month with 0.2947.

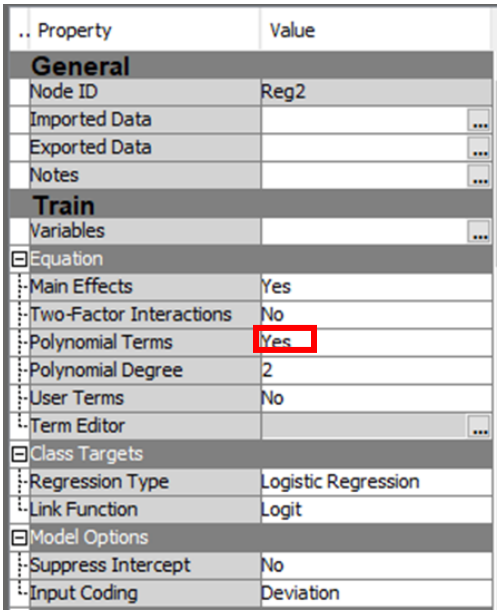
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LG10_IMP_REP_Interest_Rate	Transformed: Imputed: Replacement: Interest_Rate	3	1.0000	1.0000	1.0000
REP_IMP_REP_Credit_Mix	Replacement: Imputed: Replacement: Credit_Mix	1	0.5454	0.5240	0.9608
REP_IMP_REP_Month	Replacement: Imputed: Replacement: Month	3	0.2947	0.3028	1.0273
LG10_IMP_REP_Outstanding_Debt	Transformed: Imputed: Replacement: Outstanding_Debt	3	0.2779	0.2603	0.9368
LG10_IMP_REP_Delay_from_due_date	Transformed: Imputed: Replacement: Delay_from_due_date	2	0.2450	0.2829	1.1549
LG10_IMP_REP_Total_EMI_per_month	Transformed: Imputed: Replacement: Total_EMI_per_month	6	0.1938	0.1617	0.8346
LG10_IMP_REP_Num_Credit_Card	Transformed: Imputed: Replacement: Num_Credit_Card	2	0.1902	0.1865	0.9808
LG10_IMP_REP_Credit_History_Age	Transformed: Imputed: Replacement: Credit_History_Age	3	0.1214	0.0806	0.6639
LG10_IMP_REP_Num_of_Delayed_Paym	Transformed: Imputed: Replacement: Num_of_Delayed_Payment	1	0.1199	0.1096	0.9148
LG10_IMP_REP_Annual_Income	Transformed: Imputed: Replacement: Annual_Income	3	0.1114	0.0746	0.6703
REP_IMP_REP_Payment_of_Min_Amoun	Replacement: Imputed: Replacement: Payment_of_Min_Amount	2	0.1094	0.1159	1.0593
LG10_IMP_REP_Num_Credit_Inquirie	Transformed: Imputed: Replacement: Num_Credit_Inquiries	2	0.1085	0.1216	1.1204
LG10_IMP_REP_Num_Bank_Accounts	Transformed: Imputed: Replacement: Num_Bank_Accounts	2	0.1013	0.1031	1.0176
LG10_IMP_REP_Monthly_Inhand_Sala	Transformed: Imputed: Replacement: Monthly_Inhand_Salary	2	0.0798	0.0537	0.6736
REP_IMP_REP_Payment_Behaviour	Replacement: Imputed: Replacement: Payment_Behaviour	1	0.0745	0.0000	0.0000
LG10_IMP_REP_Amount_invested_mon	Transformed: Imputed: Replacement: Amount_invested_monthly	1	0.0728	0.0000	0.0000
LG10_IMP_Changed_Credit_Limit	Transformed: Imputed Changed_Credit_Limit	1	0.0657	0.0506	0.7692
REP_IMP_REP_Occupation	Replacement: Imputed: Replacement: Occupation	1	0.0564	0.0000	0.0000
REP_IMP_REP_not_spec_loan_f	Replacement: Imputed: Replacement: not_spec_loan_f	1	0.0403	0.0000	0.0000
LG10_IMP_REP_Credit_Utilization_	Transformed: Imputed: Replacement: Credit_Utilization_Ratio	0	0.0000	0.0000	.
LG10_IMP_REP_Age	Transformed: Imputed: Replacement: Age	0	0.0000	0.0000	.
REP_IMP_REP_debt_con_loan_f	Replacement: Imputed: Replacement: debt_con_loan_f	0	0.0000	0.0000	.
REP_IMP_REP_auto_loan_f	Replacement: Imputed: Replacement: auto_loan_f	0	0.0000	0.0000	.
LG10_IMP_REP_Monthly_Balance	Transformed: Imputed: Replacement: Monthly_Balance	0	0.0000	0.0000	.
REP_IMP_REP_payday_loan_f	Replacement: Imputed: Replacement: payday_loan_f	0	0.0000	0.0000	.
REP_IMP_REP_home_equi_loan_f	Replacement: Imputed: Replacement: home_equi_loan_f	0	0.0000	0.0000	.
REP_IMP_REP_mg_loan_f	Replacement: Imputed: Replacement: mg_loan_f	0	0.0000	0.0000	.
REP_IMP_REP_creditbuilder_loan_f	Replacement: Imputed: Replacement: creditbuilder_loan_f	0	0.0000	0.0000	.
REP_IMP_REP_student_loan_f	Replacement: Imputed: Replacement: student_loan_f	0	0.0000	0.0000	.
REP_IMP_REP_personal_loan_f	Replacement: Imputed: Replacement: personal_loan_f	0	0.0000	0.0000	.
LG10_IMP_REP_Num_of_Loan	Transformed: Imputed: Replacement: Num_of_Loan	0	0.0000	0.0000	.

Figure 8.1.4: Variable Importance

8.2 Logistic Regression Model

Logistic Regression (LR), also known as logit model is a type of statistical model used in classification and predictive analytics. The value of the dependent variable is bounded between 0 and 1 due to the probabilistic outcome. A logit transformation is performed to the odds in the Logistic Regression, which is the probability of success divided by the probability of failure. There are three types of logistics regressions which are binary Logistic Regression, multinomial Logistic Regression, and ordinal Logistic Regression.

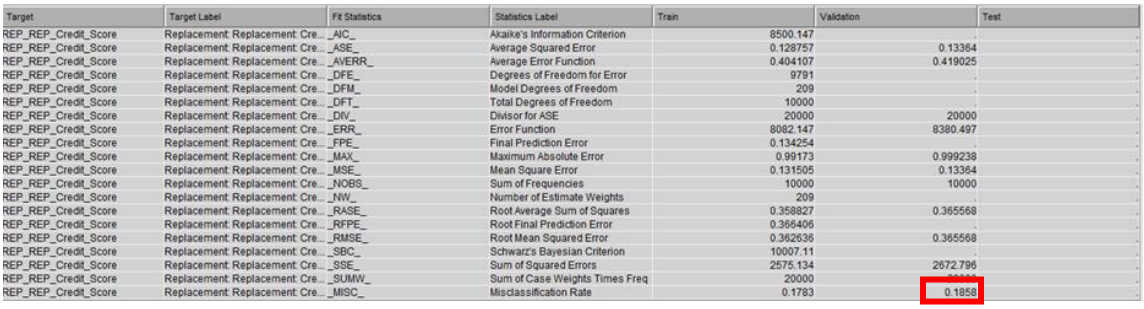
The Logistic Regression model is run using the Regression node where some properties are set as below:



Property	Value
General	
Node ID	Reg2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation

Figure 8.2.1: Logistic Regression Adjusted Property

The misclassification rate of this Logistic Regression model for the validation dataset is 0.1858. Based on the misclassification value, the accuracy obtained is 81.42%.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_AIC_	Akaike's Information Criterion	8500.147		
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_ASE_	Average Squared Error	0.128757	0.13364	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_AVERR_	Average Error Function	0.404107	0.419025	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DFE_	Degrees of Freedom for Error	9791	.	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DFM_	Model Degrees of Freedom	209	.	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DFT_	Total Degrees of Freedom	10000		
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DIV_	Divisor for ASE	20000	20000	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_ERR_	Error Function	8082.147	8380.497	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_FPE_	Final Prediction Error	0.134254	.	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MAX_	Maximum Absolute Error	0.99173	0.999238	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MSE_	Mean Square Error	0.131505	0.13364	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_NOBS_	Sum of Frequencies	10000	10000	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_NW_	Number of Estimate Weights	209	.	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_RASE_	Root Average Sum of Squares	0.358827	0.365568	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_RFPE_	Root Final Prediction Error	0.366406	.	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_RMSE_	Root Mean Squared Error	0.362636	0.365568	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_SBC_	Schwarz's Bayesian Criterion	10007.11	.	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_SSE_	Sum of Squared Errors	2575.134	2672.796	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_SUMW_	Sum of Case Weights Times Freq	20000		
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MISC_	Misclassification Rate	0.1783	0.1858	

Figure 8.2.2: Statistical Output of Logistic Regression

8.3 Gradient Boosting Model

Gradient is the term used to describe the error, or residual, that results from model construction. Boosting means to get better. Error can be improved (reduced) progressively by using gradient boosting. Gradient Boosting and random forest are comparable in certain ways, however random forest is a bagging (bootstrap aggregating) approach that mixes the output of various decision trees to get the forecast. To obtain the average forecast across all trees in a bagging technique, trees are typically generated in parallel using a sample of the original data for each tree. On the other hand, predictions made by gradient boosting have a different format. In order to obtain predictions, boosting uses a sequential method as opposed to parallelizing the tree-building process. Gradient boosting is a technique where each decision tree boosts (improves) the error of the previous decision tree by predicting its error (Ayyadevara, 2018).

The Gradient Boosting model is run using the Gradient Boosting node where some properties are set as below:

Series Options	
N Iterations	700
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	3
Minimum Categorical S	5
Reuse Variable	3
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate R	0
Split Size	.
Split Search	

Figure 8.3.1: Gradient Boosting Adjusted Property

The misclassification rate of this Gradient Boosting model for the validation dataset is 0.1538. Based on the misclassification value, the accuracy obtained is 84.62%.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_NOBS_	Sum of Frequencies	10000	10000	.
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_SUMW_	Sum of Case Weights Times Freq	20000	20000	.
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MISC_	Misclassification Rate	0.0717	0.1538	.
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MAX_	Maximum Absolute Error	0.940252	0.9973	.
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_SSE_	Sum of Squared Errors	1179.768	2240.845	.
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_ASE_	Average Squared Error	0.058988	0.112042	.
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_RASE_	Root Average Squared Error	0.242875	0.334727	.
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DIV_	Divisor for ASE	20000	20000	.
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DFT_	Total Degrees of Freedom	10000	.	.

Figure 8.3.2: Statistical Output of Gradient Boosting

8.4 Least-Angle Regression (LARS) model

Efron et al. (2004) introduced least angle regression; an algorithm used in regression for high-dimensional data. The LARS node may effectively generate LASSO solutions by modifying the LAR algorithm, which is useful for choosing the best-fitting model. The LAR algorithm generates a series of regression models, similar to the forward selection method used with regression models but sometimes may be more accurate in predictions. With each stage, a new model parameter is added. Once all parameters have been incorporated into the model, the succession of models comes to an end at the full least squares solution. A significant number of potential model inputs (independent or explanatory variables) are typically present in data mining databases and can be utilized to forecast the value of a particular objective (a dependent or response variable). The LARS node is capable of model fitting and variable selection. The LARS node picks the variables in a continuous manner when used for variable selection, as the coefficients for each chosen variable increase from zero to the variable's least square estimates.

The LARS model is run using the LARS node where some properties are set as below:

Modeling Techniques	
Use Class Inputs	Yes
Intercept	Yes
Variable Selection Me	LASSO
Model Selection Crite	Cross Validation
Path Stopping Criterio	Maximum Steps
Maximum Steps	600
Cross Validation Optio	
Cross Validation	Random
CV Fold	200
Seed	12345

Figure 8.4.1: LARS Adjusted Property

The misclassification rate of this Gradient Boosting model for the validation dataset is 0.1833. Based on the misclassification value, the accuracy obtained is 81.67%.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_ASE_	Average Squared Error	0.137976	0.13791	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DIV_	Divisor for ASE	20000	20000	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MAX_	Maximum Absolute Error	0.988107	0.974172	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_NOBS_	Sum of Frequencies	10000	10000	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_RASE_	Root Average Squared Error	0.371452	0.371362	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_SSE_	Sum of Squared Errors	2759.53	2758.192	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_DISF_	Frequency of Classified Cases	10000	10000	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_MISC_	Misclassification Rate	0.1856	0.1833	
REP_REP_Credit_Score	Replacement: Replacement: Cre...	_WRONG_	Number of Wrong Classifications	1856	1833	

Figure 8.4.2: Statistical Output of Least-Angle Regression

9 SEMMA: Assess

The performance between the four chosen models which are the Decision Tree (DT), Logistic Regression (LR), Gradient Boosting (GB), and Least-Angle Regression (LARS) are compared using several evaluation metrics, including misclassification rate, accuracy, precision, recall, F1-score, and ROC-AUC.

Based on Figure 9.1, it is observed that the Gradient Boosting has the highest True Positive and True Negative value for both the training and validation sets. The confusion matrices presented in Figure 9.2 describe the performance of each algorithm on training and validation sets. These matrices serve as the basis for calculating important evaluation metrics such as precision, recall, accuracy, and F1-score.

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg2	Regression Model	TRAIN	REP_REP_Credit_Score	Replacement: Replacement: Credit_Score	1042	4259	741	3958
Reg2	Regression Model	VALIDATE	REP_REP_Credit_Score	Replacement: Replacement: Credit_Score	1068	4210	790	3932
Tree3	Decision Tree Model	TRAIN	REP_REP_Credit_Score	Replacement: Replacement: Credit_Score	894	4373	627	4106
Tree3	Decision Tree Model	VALIDATE	REP_REP_Credit_Score	Replacement: Replacement: Credit_Score	917	4318	682	4083
Boost2	Gradient Boosting Model	TRAIN	REP_REP_Credit_Score	Replacement: Replacement: Credit_Score	497	4780	220	4503
Boost2	Gradient Boosting Model	VALIDATE	REP_REP_Credit_Score	Replacement: Replacement: Credit_Score	978	4440	560	4022
LARS	LARS	TRAIN	REP_REP_Credit_Score	Replacement: Replacement: Credit_Score	1087	4231	769	3913
LARS	LARS	VALIDATE	REP_REP_Credit_Score	Replacement: Replacement: Credit_Score	1075	4242	758	3925

Figure 9.1: Classification Table

Decision Tree Model

		Training Set Outcome	
		P	N
Target	P	4106	894
	N	627	4373

		Validation Set Outcome	
		P	N
Target	P	4083	917
	N	682	4318

Logistic Regression Model

		Training Set Outcome	
		P	N
Target	P	3958	1042
	N	741	4259

		Validation Set Outcome	
		P	N
Target	P	3952	1068
	N	790	4210

Gradient Boosting Model

		Training Set Outcome	
		P	N
Target	P	4503	497
	N	220	4780

		Validation Set Outcome	
		P	N
Target	P	4022	978
	N	560	4440

LARS Model

		Training Set Outcome	
		P	N
Target	P	3913	1087
	N	769	4231

		Validation Set Outcome	
		P	N
Target	P	3925	1075
	N	758	4242

Figure 9.2 Confusion Matrices

The observation made in Figure 9.1 coincides and is better represented by the misclassification rates as shown in Figure 9.3.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Boost2	Boost2	Gradient Bo...	REP_REP_...	Replaceme...	0.1538
	Tree3	Tree3	Decision Tr...	REP_REP_...	Replaceme...	0.1599
	LARS	LARS	LARS	REP_REP_...	Replaceme...	0.1833
	Reg2	Reg2	Regression...	REP_REP_...	Replaceme...	0.1858

Figure 9.3: Model Performance

Table 9.1 shows the precision, recall, accuracy, and F1-score for each model. The Gradient Boosting model outperformed the other models with a misclassification rate of 0.1538, and accuracy score of 84.62%. With a slightly different in the misclassification rate, 0.1599, the Decision Tree has an accuracy of 84.01%. The other two models, LARS and Logistic Regression have an accuracy of 81.67% and 81.42% respectively. Apart from that, Gradient Boosting also has the highest precision and F1-score, whereas Decision Tree has the highest recall, followed by Gradient Boosting with the second highest recall value.

Models	**DT**	**LR**	**GB**	**LARS**
Accuracy	84.01%	81.42%	84.62%	81.67%
Precision	0.8569	0.8327	0.8778	0.8381
Recall	0.8166	0.7864	0.8044	0.7850
F1-Score	0.8362	0.8089	0.8395	0.8107

Table 9.1: Accuracy, Precision, Recall and F1-Score of the Models

Figure 9.4 shows that Gradient Boosting has the highest AUC followed by DT, Logistic Regression, and LARS with the lowest AUC. An excellent model has an AUC close to one, indicating that it has a high level of separability, so the higher the AUC, the better the model.

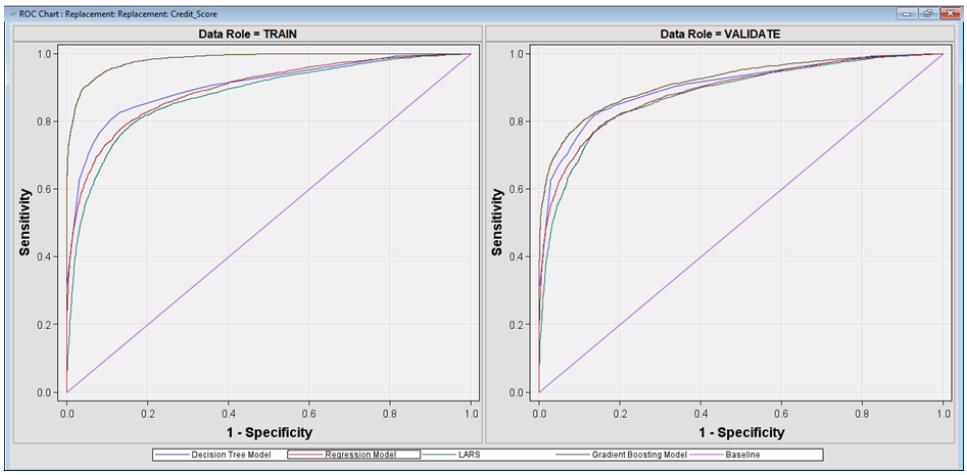


Figure 9.4: ROC Curve

Overall, the Gradient Boosting model is the best model for this credit score classification problem. It has the highest accuracy, precision, F1-score, and AUC as well as the second highest recall value.

10 Conclusion

In the data exploration stage, there are variables of several error types that required data preprocessing and cleaning. They are then modified using different methods such as imputation and transformation, which are illustrated in Table 10.1.

Variable	Data Type Error	Replacement (Y/N)	Imputation (Method)	Log10 Transformation (Y/N)	Other modification
Name	Incomplete				Dropped
Age	Noisy	Y	Median	Y	-
Occupation	Noisy, Incomplete	Y	Count		-
Annual_Income	Noisy	Y	Median	Y	-
Monthly_Inhand_Salary	Incomplete	Y	Median	Y	-
Num_Bank_Accounts	Noisy	Y	Median	Y	-
Num_Credit_Card	Noisy	Y	Median	Y	-
Num_of_Loan	Noisy, Inconsistent	Y	Median	Y	-
Type_of_Loan	Incomplete				Split column
Num_of_Delayed_Payment	Incomplete, Noisy	Y	Median	Y	-
Changed_Credit_Limit	Incomplete, Noisy		Median	Y	-
Num_Credit_Inquiries	Incomplete	Y	Median	Y	-
Credit_Mix	Incomplete	Y	Count		-
Outstanding_Debt	Noisy	Y	Median	Y	-

Credit_History_Age	Incomplete, Noisy	Y	Median	Y	-
Payment_of_Min_Amount	Incomplete	Y	Count		-
Amount_invested_monthly	Incomplete, Noisy	Y	Median	Y	-
Payment_Behaviour	Noisy	Y	Count		-
Monthly_Balance	Incomplete	Y	Median	Y	-
Interest_Rate	Noisy	Y	Median	Y	-
Total_EMI_per_Month	Noisy	Y	Median	Y	-
Month	Incomplete	Y	Count		-
Credit_Utilization	-			Y	-
Delay_from_Due_Date	Incomplete	Y	Median	Y	-

Table 10.1 Data Modification Summary

After modifying, the variables in the initial dataset are reclassified and new variables are added into the dataset. A new dataset is formed with a total of 32 inputs and 1 target variable, with the variable types listed in Table 10.2 below.

Role	Types	Variable Count
Input	Nominal	4
	Interval	17
	Ordinal	1
	Binary	9
Target	Binary	1

Table 10.2 Variable Type and Role Summary

In the Modelling and Assessment Stage, several models are selected to classify the credit score of the customers based on their credit-related information (features) and these models are assessed based on various evaluation metrics as summarized in Table 10.3.

Model Type	Misclassification Rate	Accuracy Score (%)
Decision Tree (DT)	0.1599	84.01
Logistic Regression (LR)	0.1858	81.42
Gradient Boosting (GB)	0.1538	84.62
Least-Angle Regression (LARS)	0.1833	81.67

Table 10.3 Model Types and Performance

The following key findings were discovered during the data exploration phase by examining the relationships between features and target, as well as significant patterns within the modified data. These valuable insights highlight the factors that are closely associated with credit scores, thus allowing us to successfully achieve one of our project goals, which is to identify the key determinants of creditworthiness.

1. The most frequent payment behavior of the borrowers is low spending with small value payments and most of them are scientists.

2. Based on the relationship between the input variables, it shows that annual income and monthly in hand salary both have high positive correlation, followed by loan and total EMI per month. Both relationships are logical as annual income and monthly in hand salary are related to the borrower's income earned. When the loan amount is high, eventually total EMI to repay is high, given that the downpayment is constant. Meanwhile, the interest rate and credit history age have a negative correlation value, indicating that interest rate is lower when the credit history age is higher. This is due to lenders being able to track a lengthier pattern of good credit history of a borrower.
3. From an association standpoint between input variables and target variable using Chi-Square plot, credit mix and payment of minimum amount variables show higher correlation with credit score, which potentially shows greater variable importance in classifying the credit score.

By incorporating these findings into model development, we can ensure that the resulting model captures the important features that contribute to creditworthiness and provides more precise predictions on customers' credit scores.

The interesting findings in modelling and assessing phase are described below.

4. From the decision tree model, variables with the highest important values are interest rate, credit mix and the month. These variables demonstrate a stronger influence on the model's prediction compared to others. However, it should not be treated as a general statement about the variable importance in the real world.
5. In the decision tree model, the interest rate is selected as the root node as it has the highest level of information gain. In terms of splitting, total EMI per month has the highest splitting number, which indicates that the tree considers total EMI per month most informative and has highest discriminatory power for making splits in the data.

In this study, we have developed the credit score classification model based on customers' credit related information and assessed the best classification model by evaluating performance of different models on various metrices. Based on the result, we are able to justify that Gradient Boosting is the best credit score classification model as it provides the best accuracy with the lowest misclassification rate. All objectives of this study are successfully accomplished, including determining the key factors affecting the credit score, building credit classification model and assessing the best models.

11 Reference

Ayyadevara, V. K. (2018). Gradient Boosting Machine. *Pro Machine Learning Algorithms*, 117–134. https://doi.org/10.1007/978-1-4842-3564-5_6

Decision tree algorithm in Machine Learning - Javatpoint. www.javatpoint.com. (n.d.). https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, *32*(2), 407–451. https://www.jstor.org/stable/3448465

Einav, L., Jenkins, M. A., & Levin, J. (2013). The impact of credit scoring on consumer lending. The RAND Journal of Economics, 44(2), 249–274. https://doi.org/10.1111/1756-2171.12019

Hand, D. J., & Henley, W. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. Journal of the Royal Statistical Society, 160(3), 523–541. https://doi.org/10.1111/j.1467-985x.1997.00077.x

Jha, V. (2020, September 5). Semma model. GeeksforGeeks. https://www.geeksforgeeks.org/semma-model/

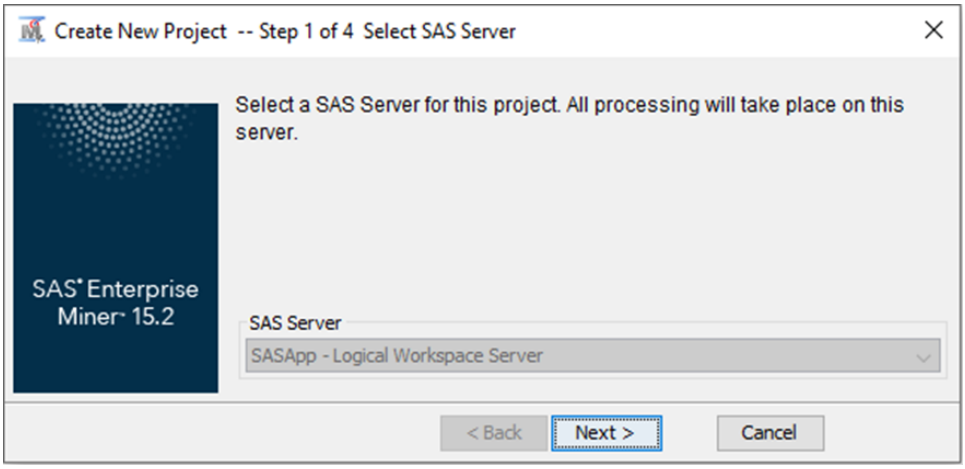
Li, X., Ying, W., Tuo, J., Li, B., & Liu, W. (2004). Applications of classification trees to consumer credit scoring methods in commercial banks. Systems, Man and Cybernetics. https://doi.org/10.1109/icsmc.2004.1401175

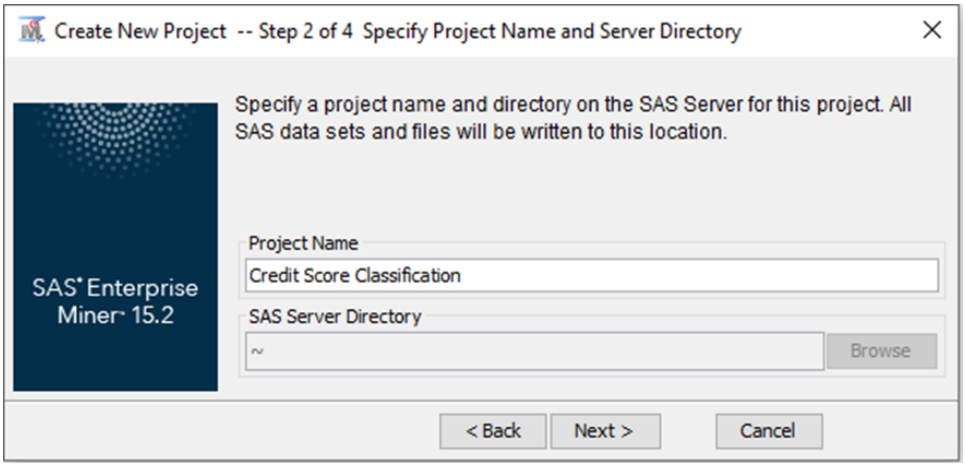
What is logistic regression?. IBM. (n.d.). https://www.ibm.com/topics/logistic-regression#:~:text=Related%20solutions-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables.

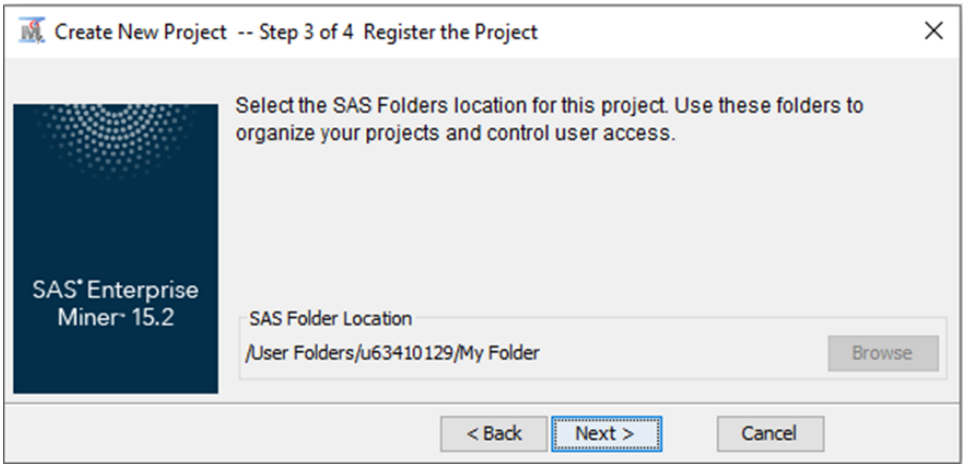
12 Appendix

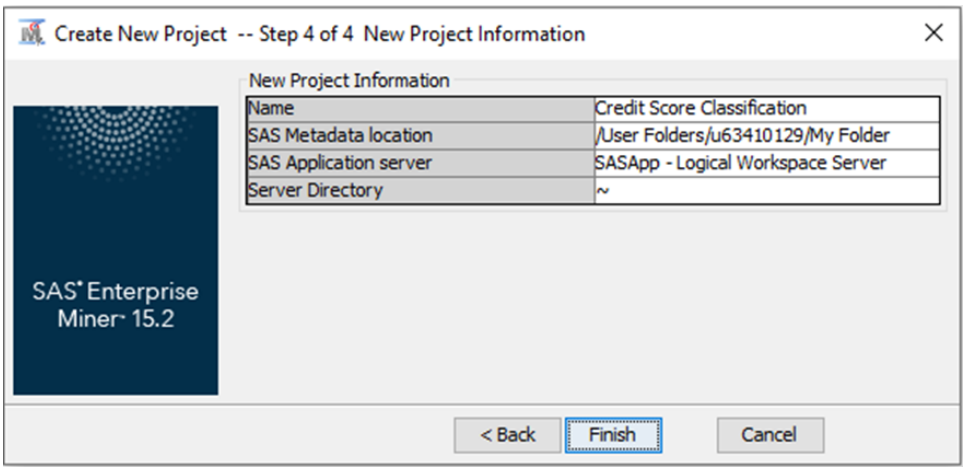
12.1 Project Setup

Figures 12.1.1 - 12.1.4 Create Project

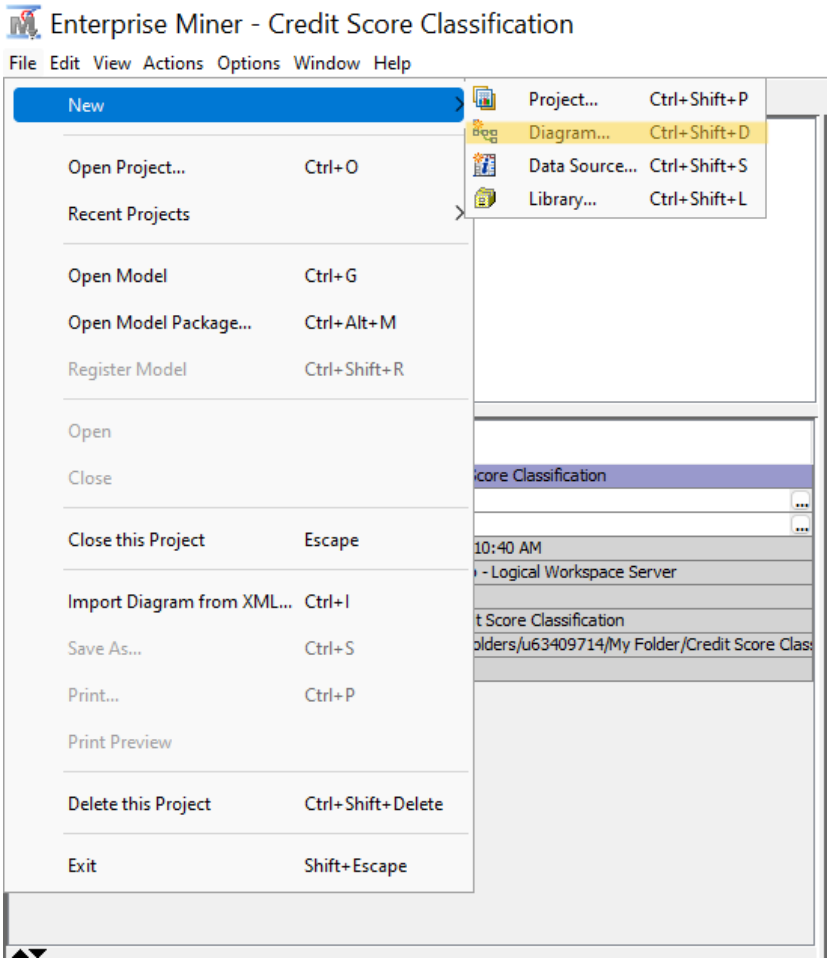


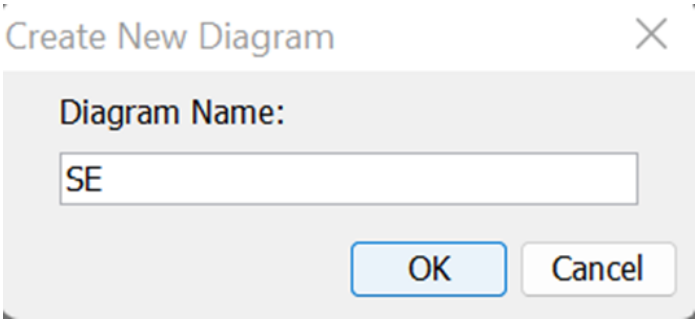




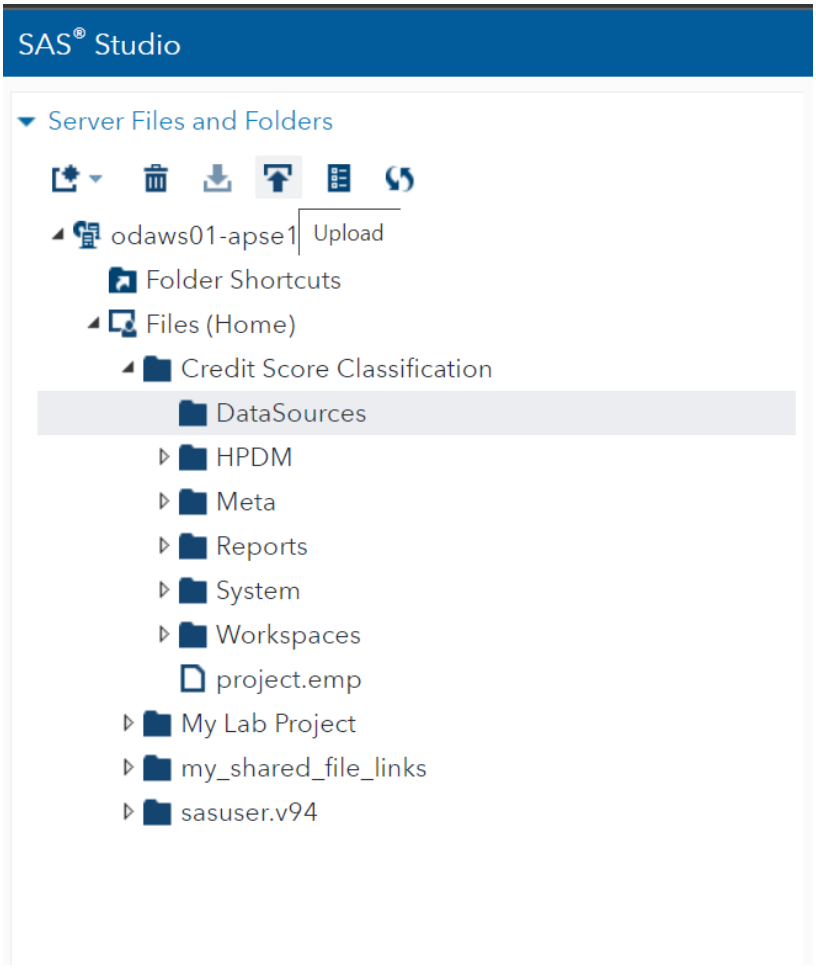


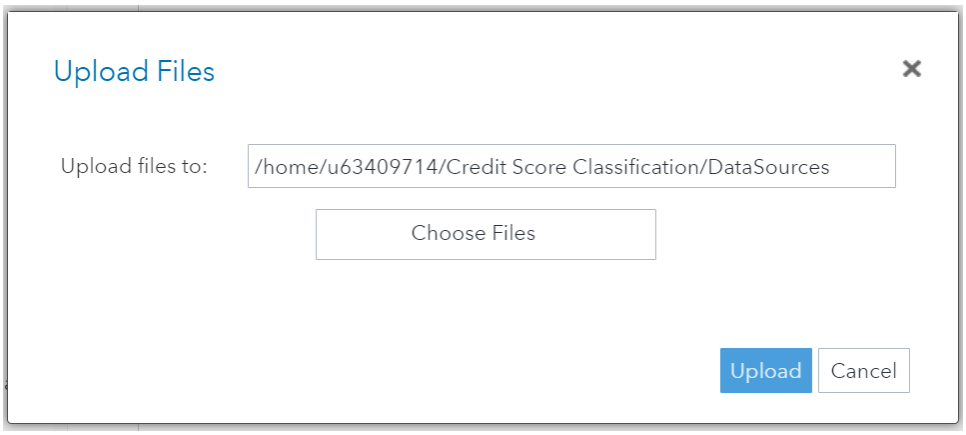
Figures 12.1.5 & 12.1.6 Create Diagram

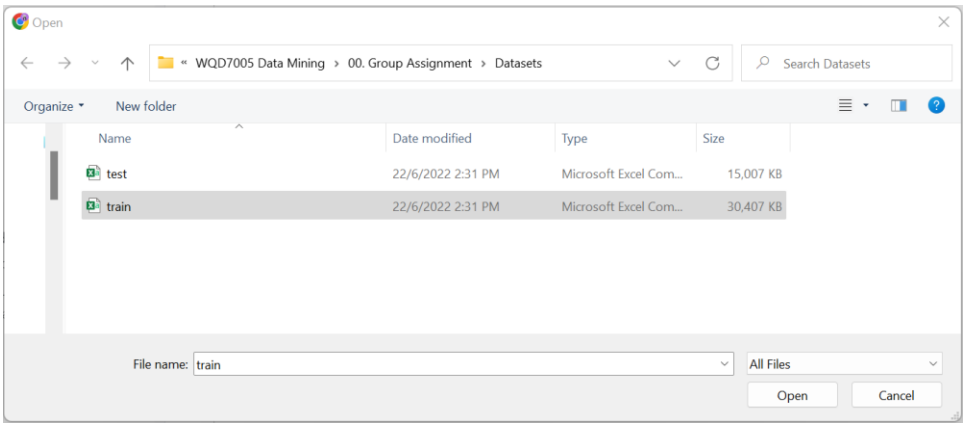


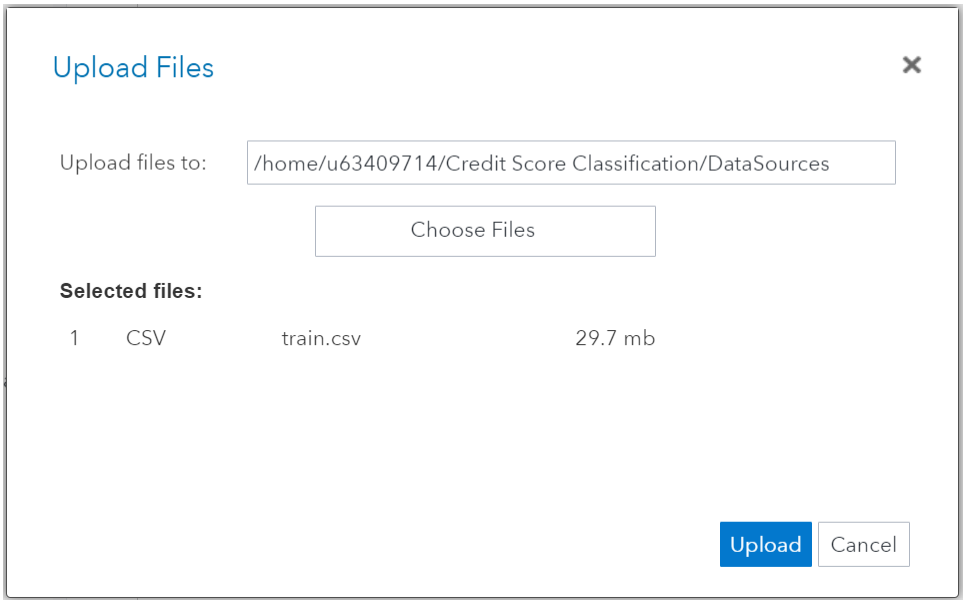


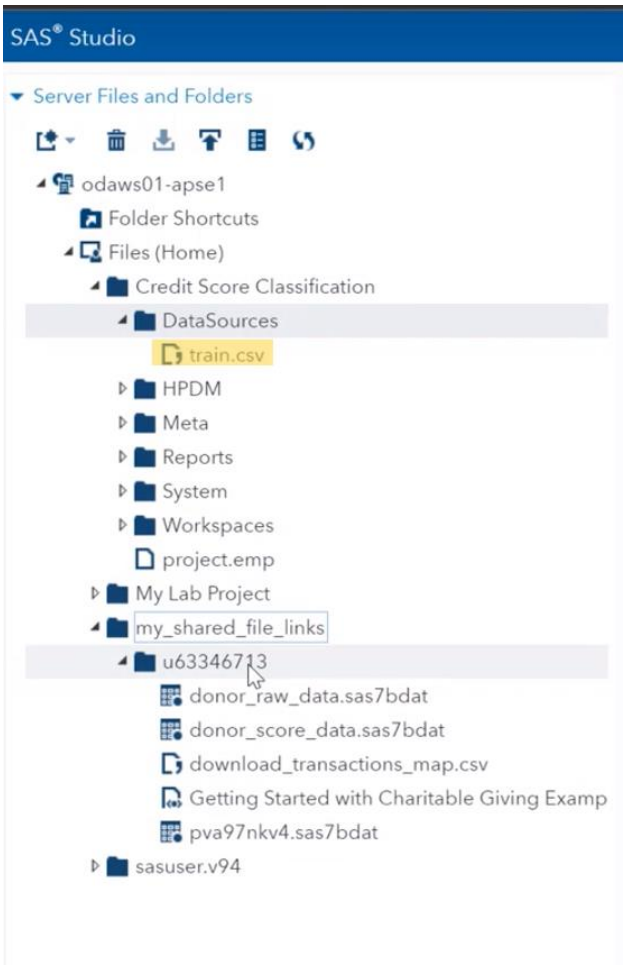
Figures 12.1.7 - 12.1.11 Upload Data to SAS Studio



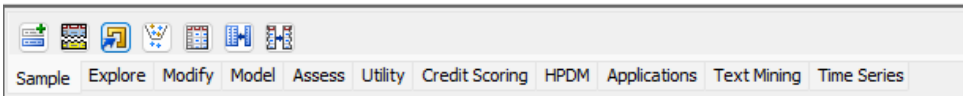

Upload Files
Upload files to:
/home/u63409714/Credit Score Classification/DataSources
Choose Files
Upload
Cancel


Open
WQD7005 Data Mining › 00. Group Assignment › Datasets
Search Datasets
Organize
New folder
Name
Date modified
Type
Size
test
22/6/2022 2:31 PM
Microsoft Excel Com...
15,007 KB
train
22/6/2022 2:31 PM
Microsoft Excel Com...
30,407 KB
File name: train
All Files
Open
Cancel

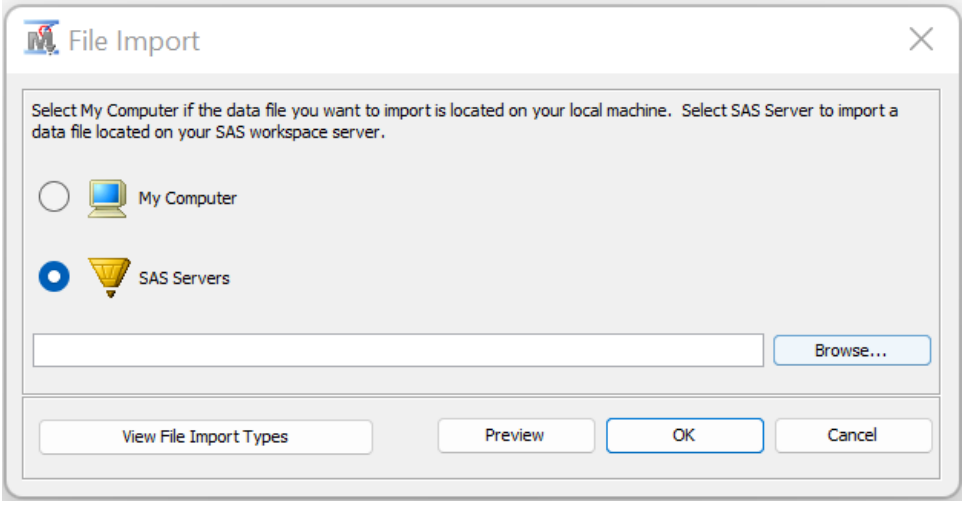

Upload Files
Upload files to:
/home/u63409714/Credit Score Classification/DataSources
Choose Files
Selected files:
1 CSV train.csv 29.7 mb
Upload
Cancel

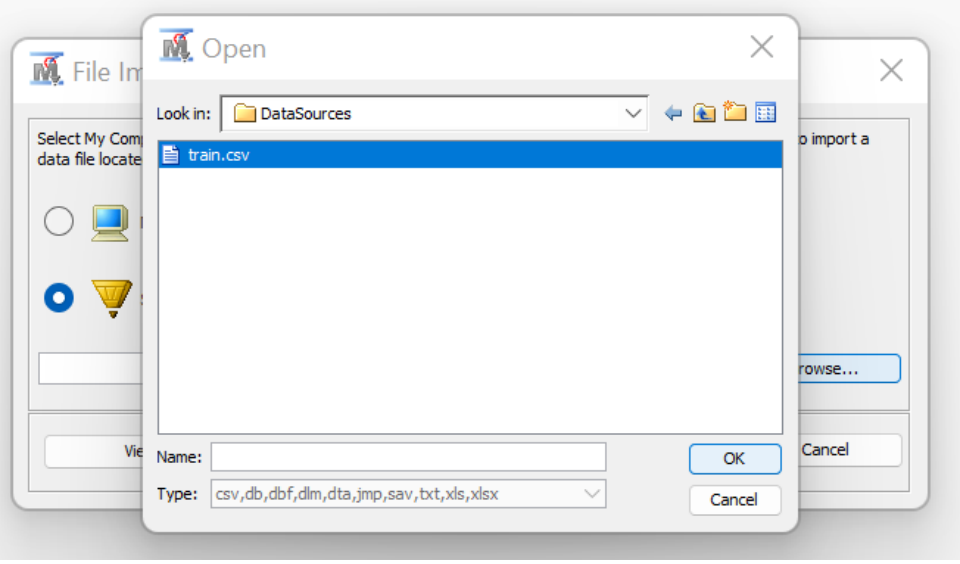


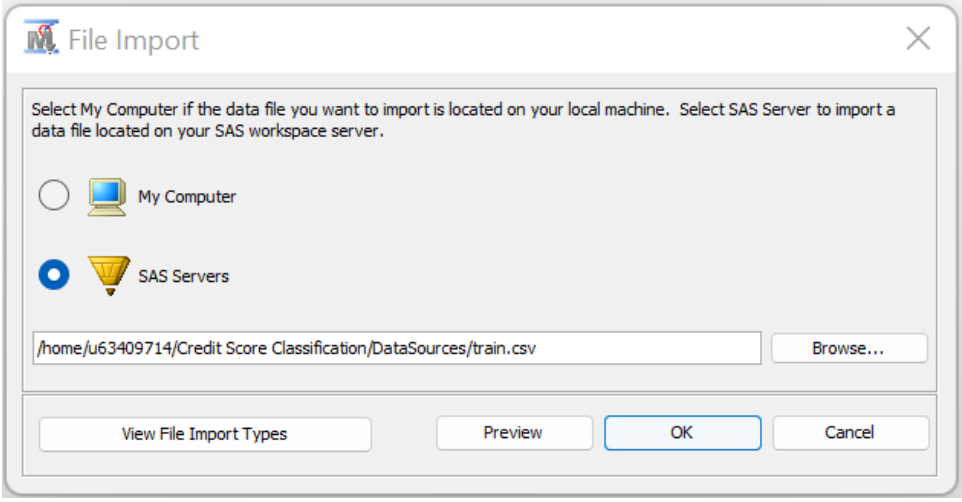
Figures 12.1.12 - 12.1.18 Import File to Enterprise Miner



Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	XLS
Advanced Advisor	No
Rerun	No
Score	
Role	Train

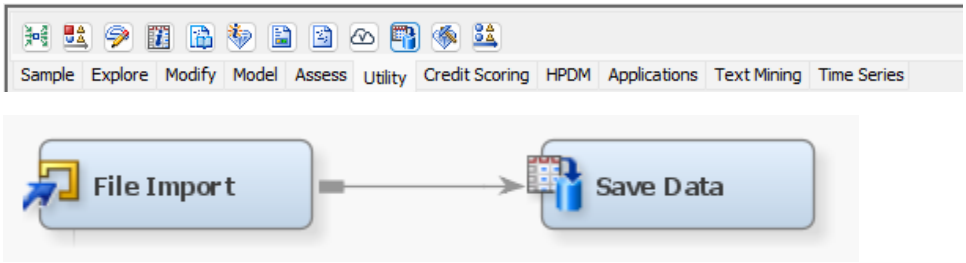


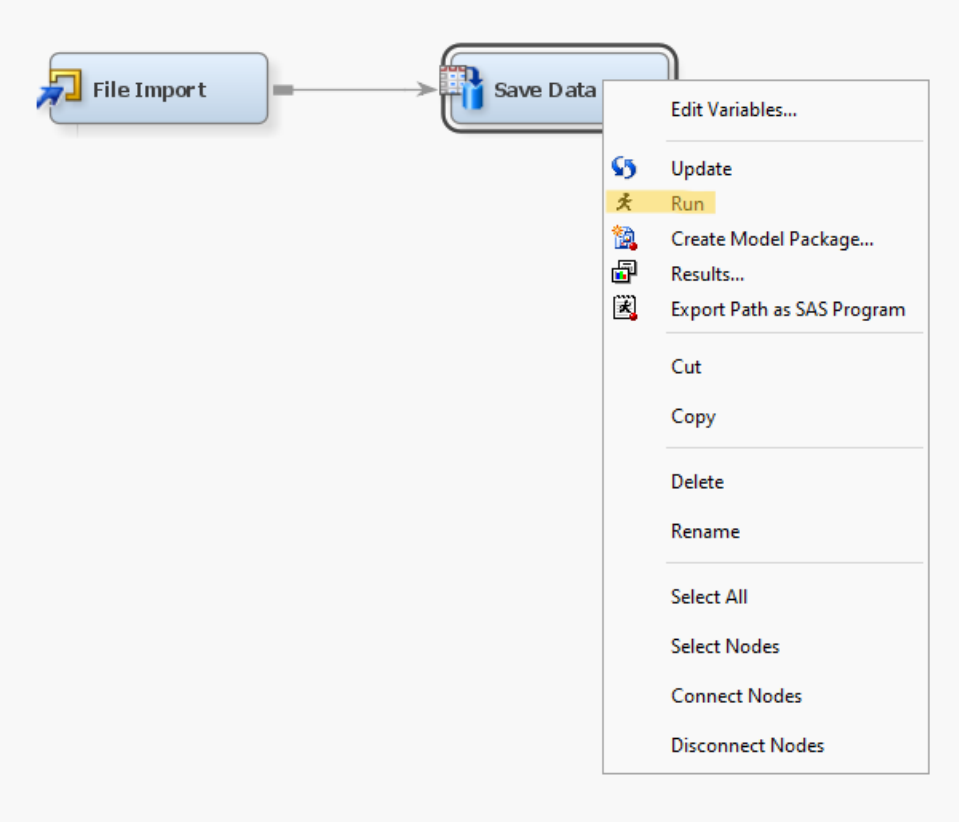

File Im
Open
Look in:
DataSources
train.csv
Name:
Type:
csv,db,dbf,dlm,dta,jmp,sav,txt,xls,xlsx
OK
Cancel
Cancel

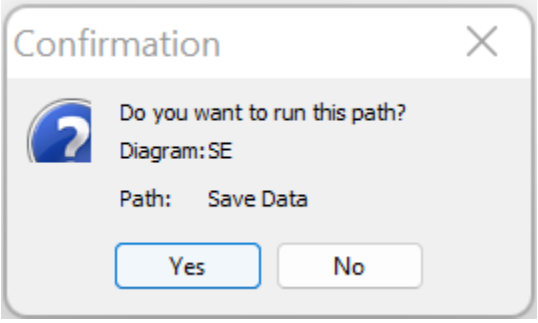

File Import
Select My Computer if the data file you want to import is located on your local machine. Select SAS Server to import a data file located on your SAS workspace server.
My Computer
SAS Servers
/home/u63409714/Credit Score Classification/DataSources/train.csv
Browse...
View File Import Types
Preview
OK
Cancel

Property	Value
General	
Node ID	FIMPORT
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Import File	/home/u63409714/Credit Score Classification/Dat
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Remote
File Type	csv
Advanced Advisor	No
Rerun	No
Score	
Role	Train

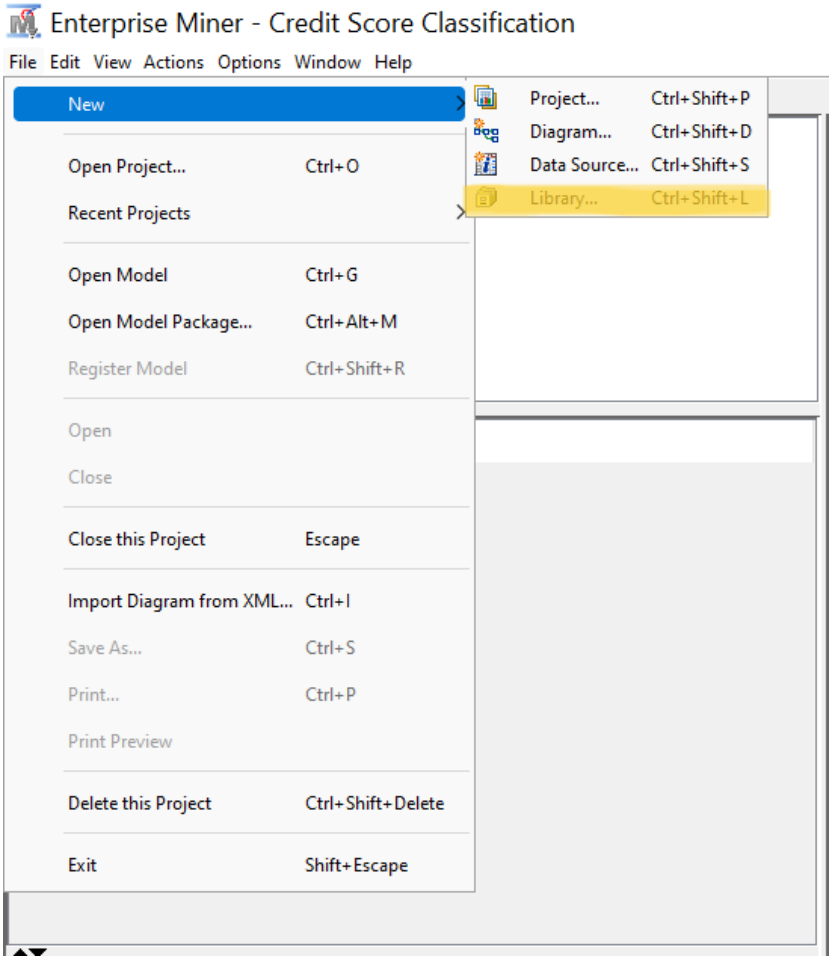
Figures 12.1.19 - 12.1.23 Save as SAS Table

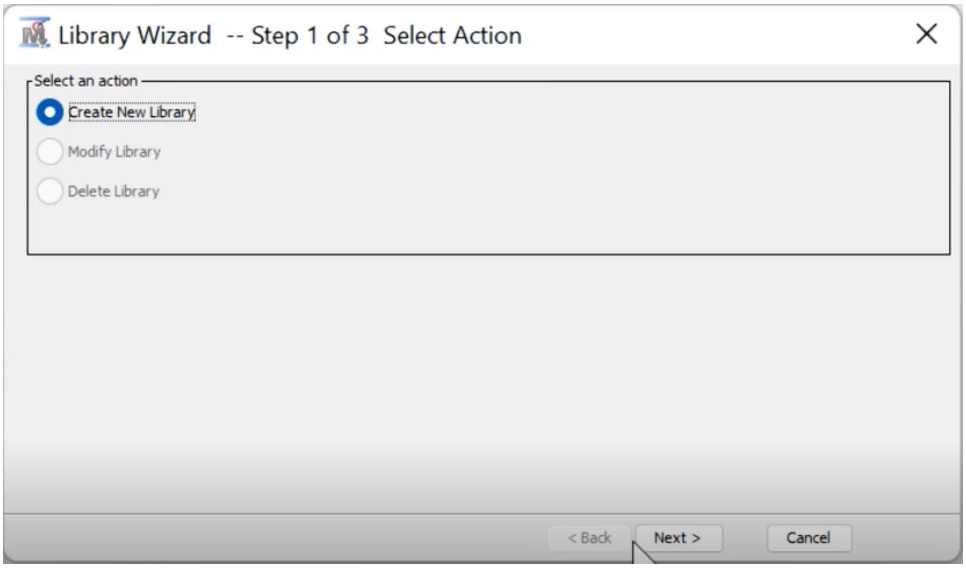


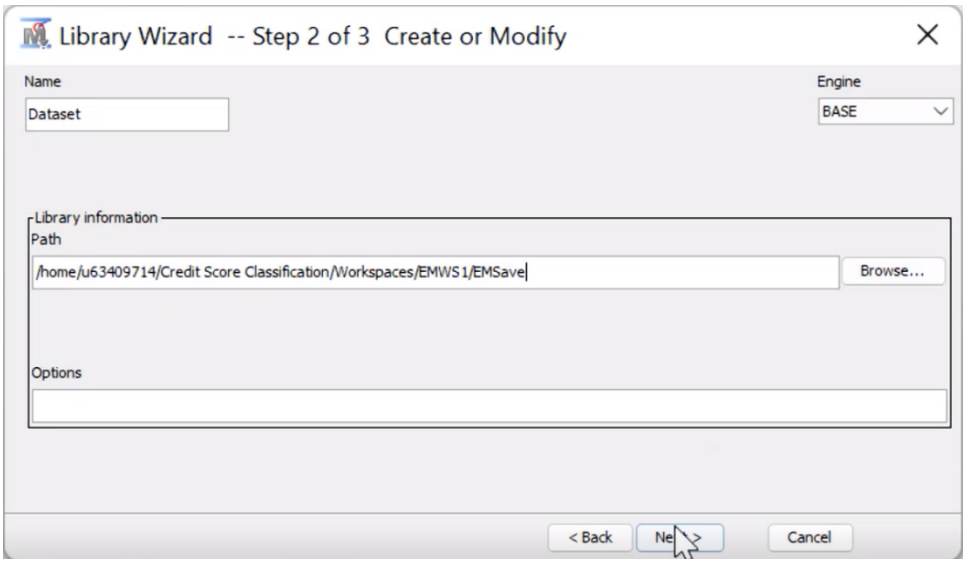


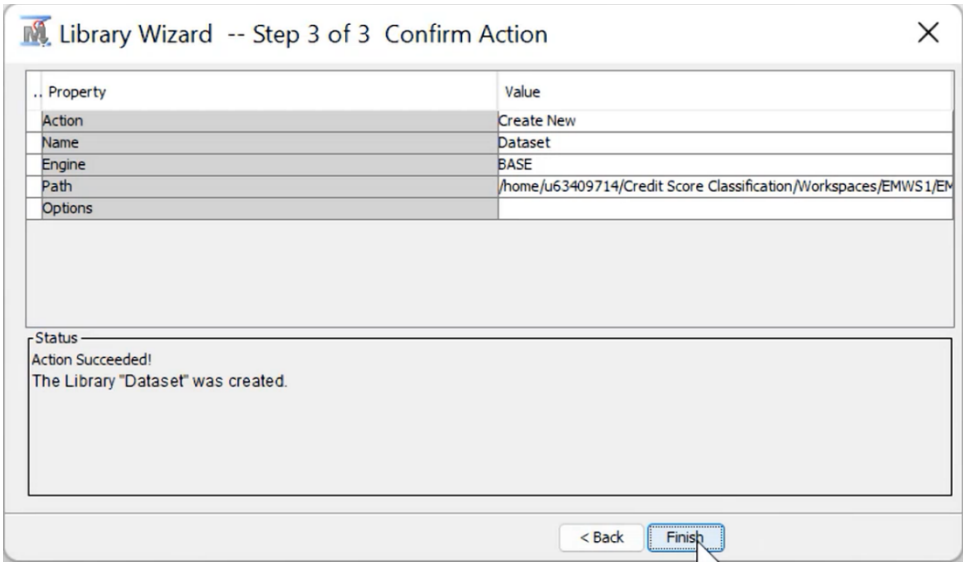


Figures 12.1.24 - 12.1.27 Create Library


Enterprise Miner - Credit Score Classification
File Edit View Actions Options Window Help
New
Open Project... Ctrl+O
Recent Projects
Open Model Ctrl+G
Open Model Package... Ctrl+Alt+M
Register Model Ctrl+Shift+R
Open
Close
Close this Project Escape
Import Diagram from XML... Ctrl+I
Save As... Ctrl+S
Print... Ctrl+P
Print Preview
Delete this Project Ctrl+Shift+Delete
Exit Shift+Escape
Project... Ctrl+Shift+P
Diagram... Ctrl+Shift+D
Data Source... Ctrl+Shift+S
Library... Ctrl+Shift+L

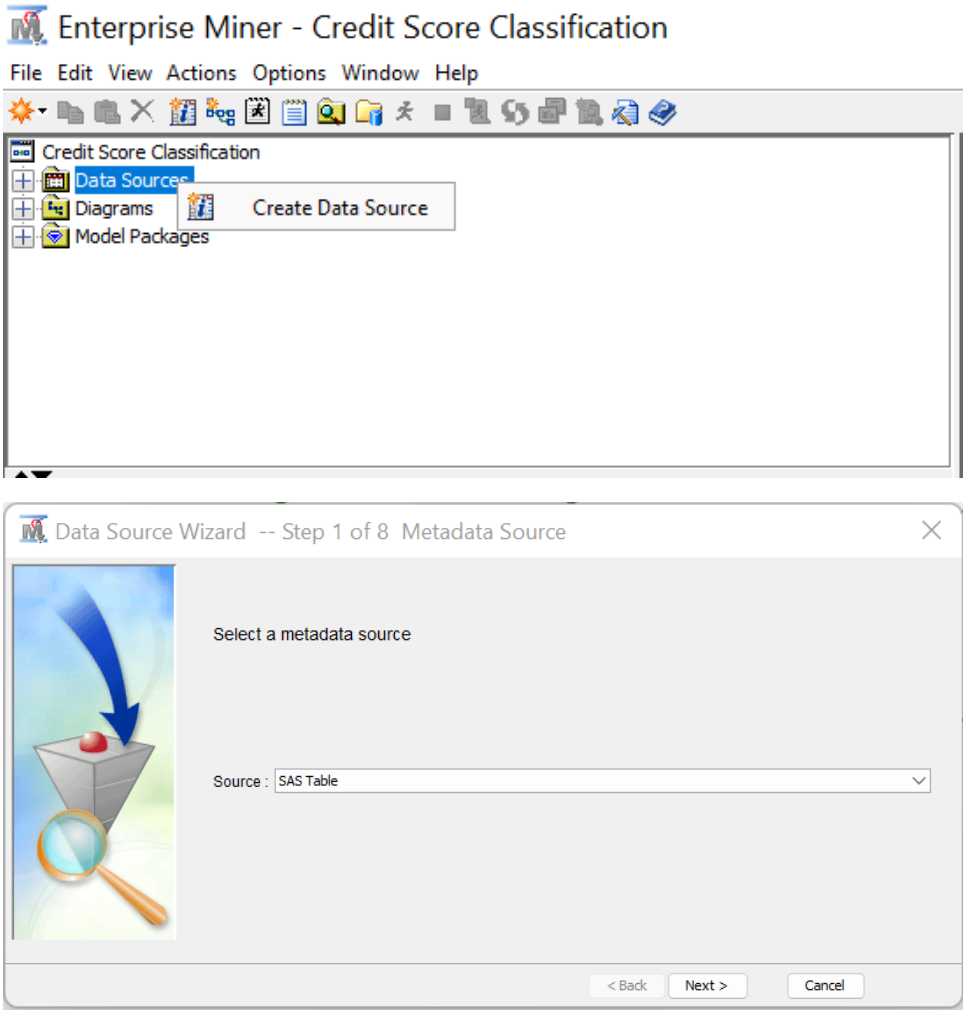

Library Wizard -- Step 1 of 3 Select Action
Select an action
Create New Library
Modify Library
Delete Library
< Back
Next >
Cancel



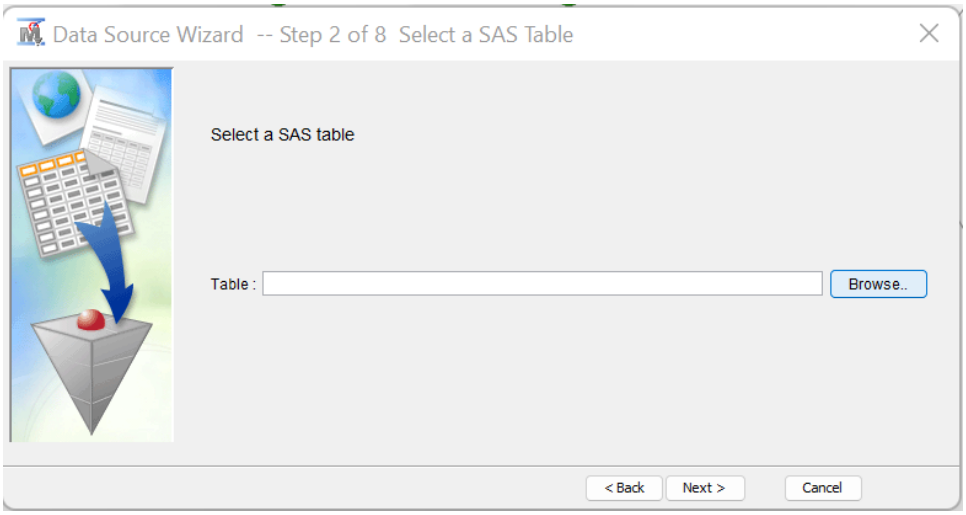


12.2 Sample

Figures 12.2.1 - 12.2.2 Create Data Source



Figures 12.2.3 - 12.2.6 Select SAS Table



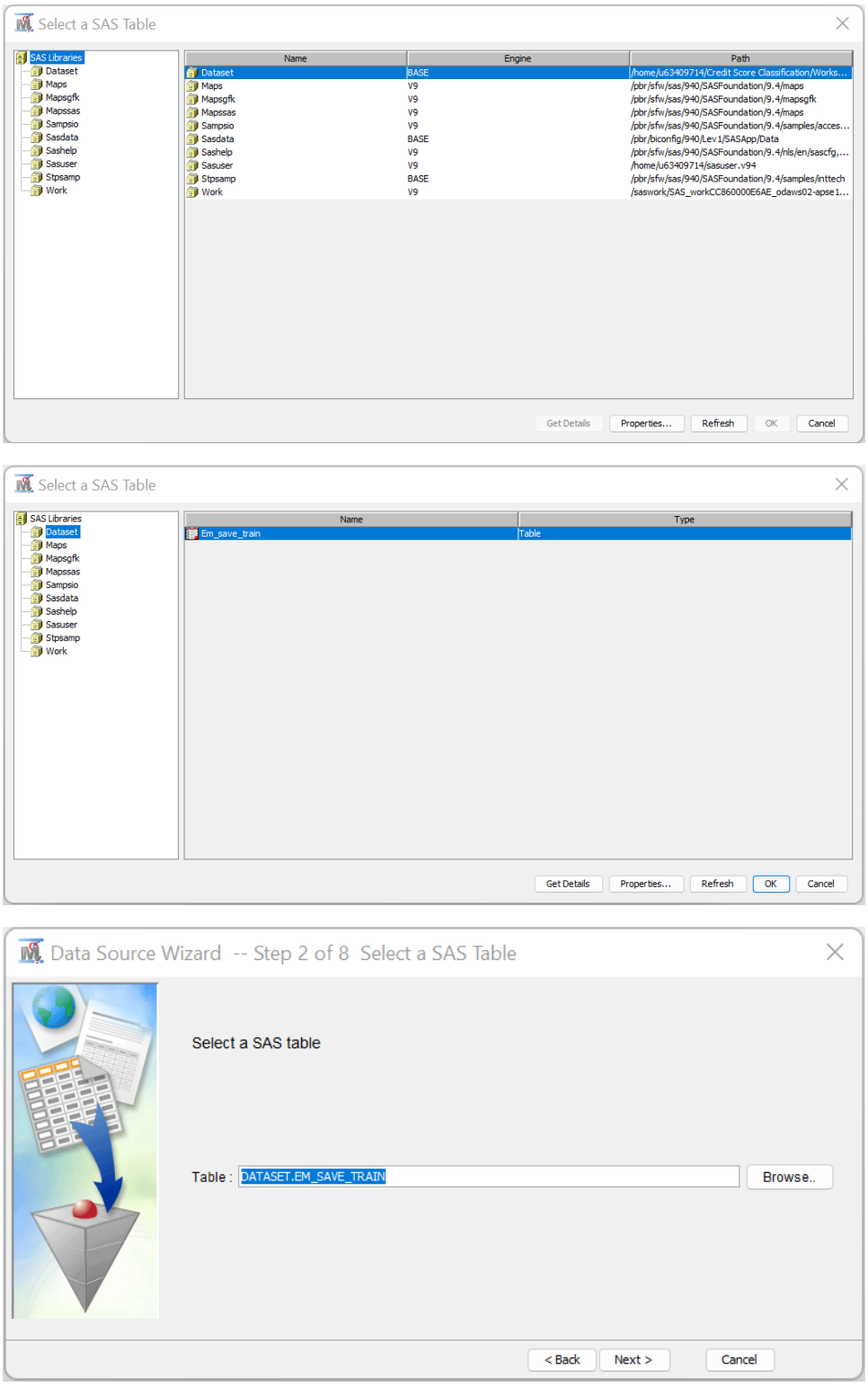


Figure 12.2.7 Information of Selected SAS Table

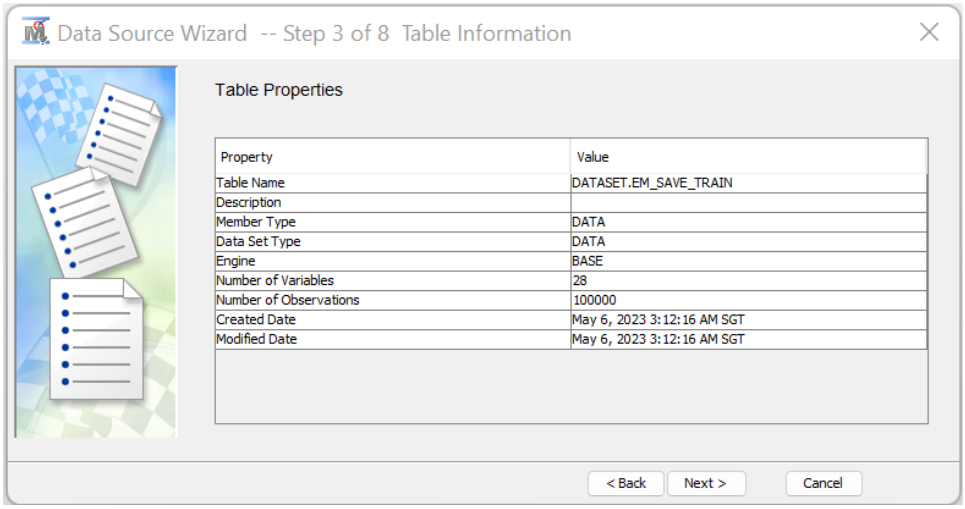


Figure 12.2.8 Metadata Advanced Setting

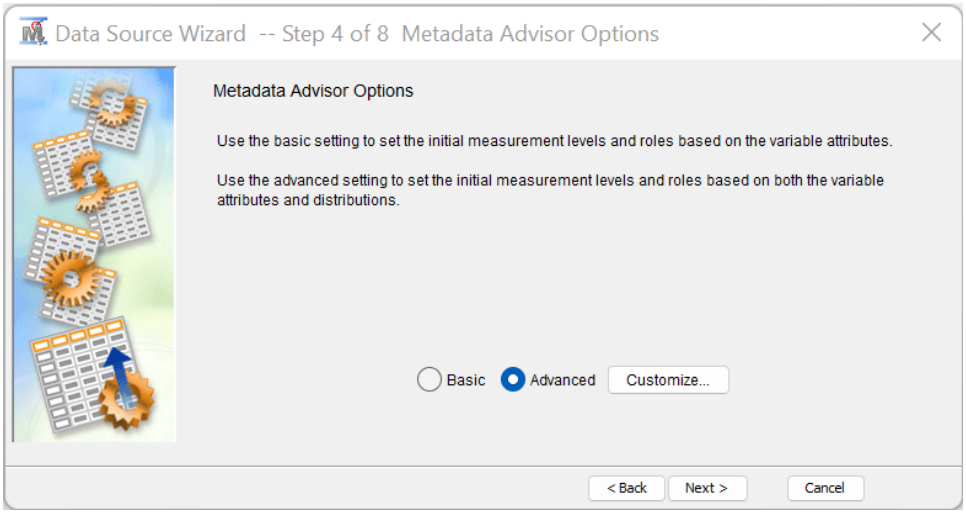


Figure 12.2.9 Default Metadata Setting

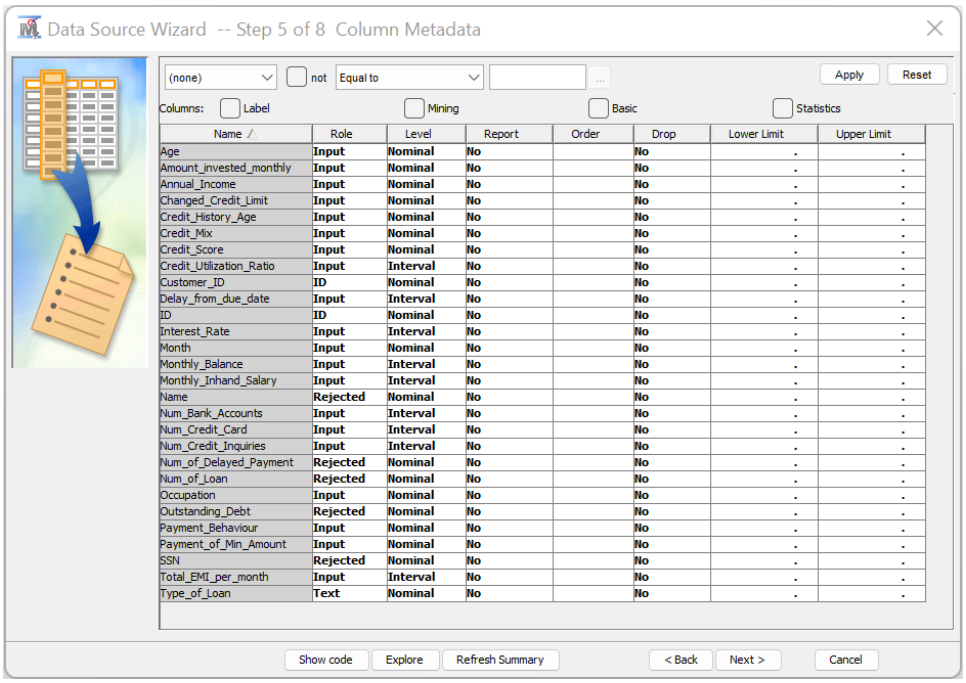


Figure 12.2.10 Modified Metadata Setting

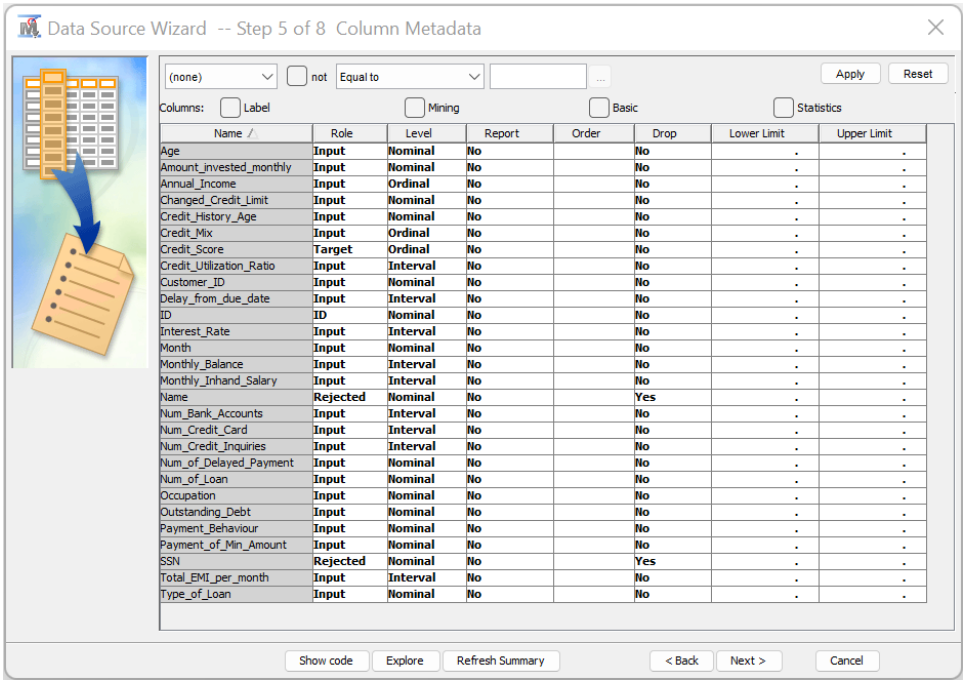


Figure 12.2.11 Unspecified Decision Processing

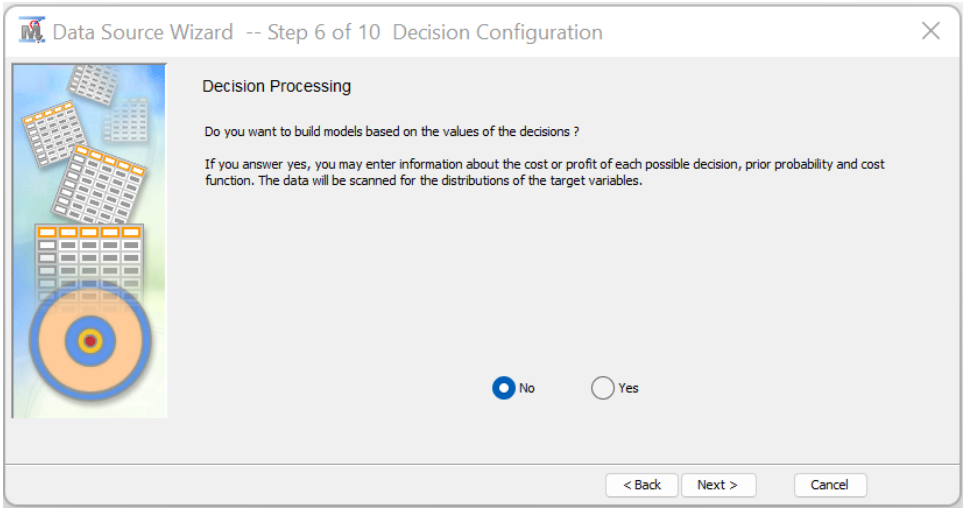


Figure 12.2.12 Did Not Create Sample Dataset

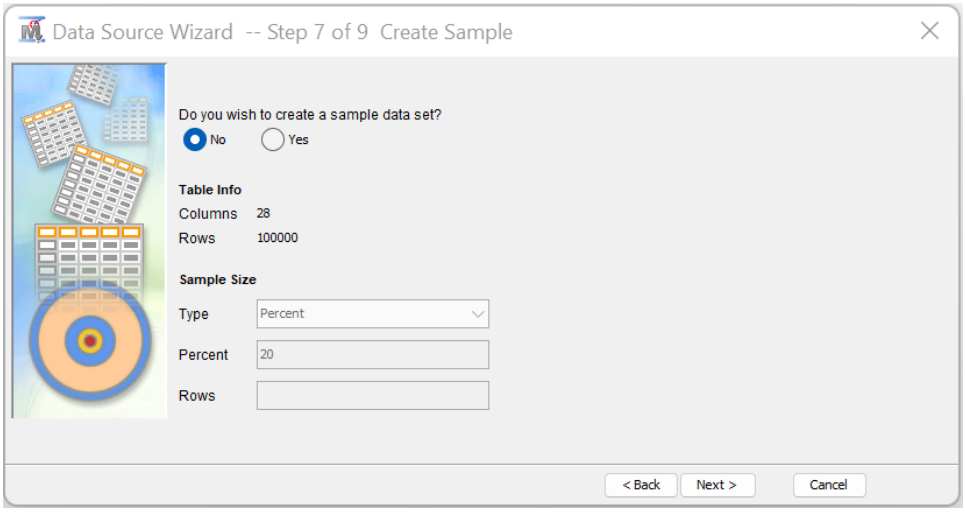


Figure 12.2.13 Name and Role of Data Source

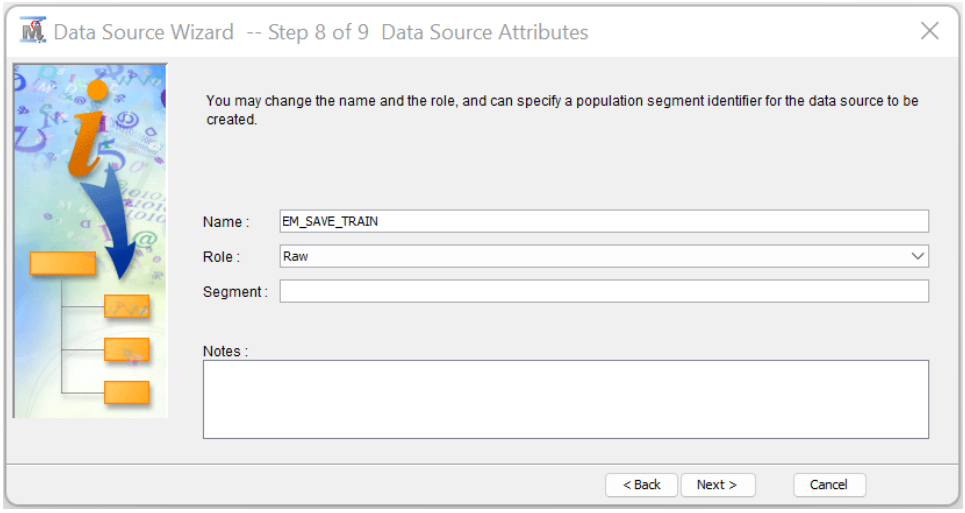
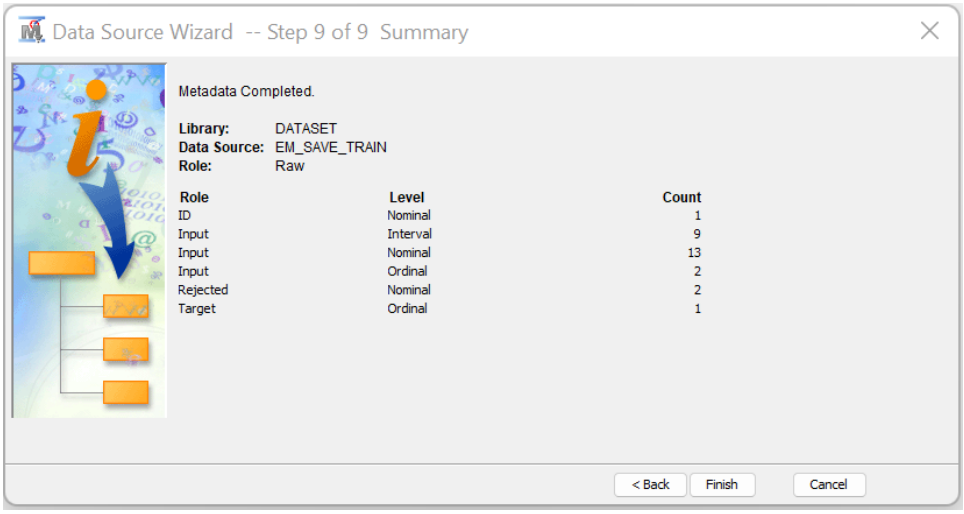
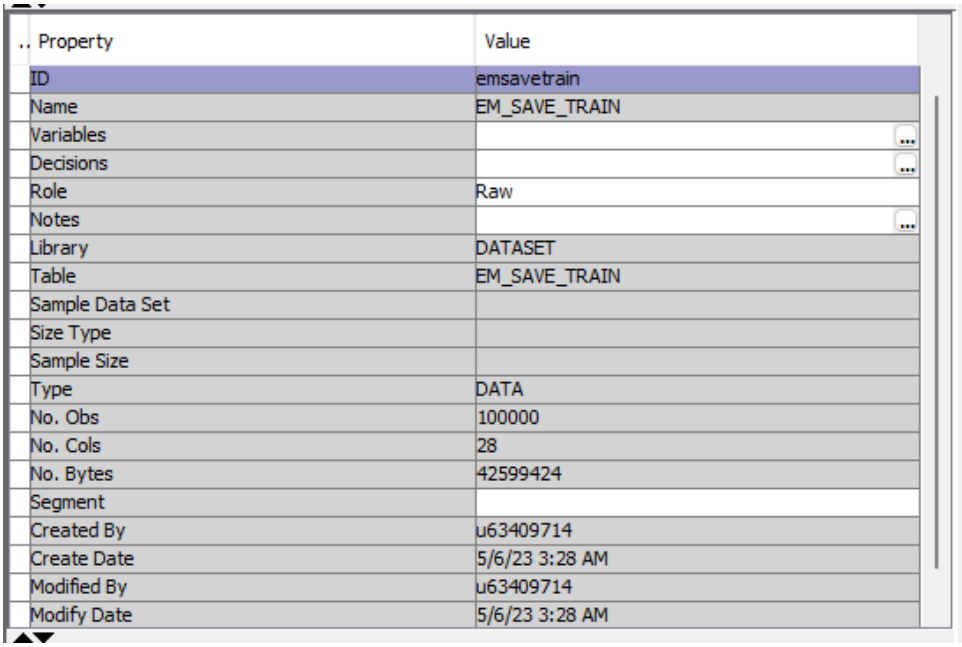


Figure 12.2.14 Data Source Summary

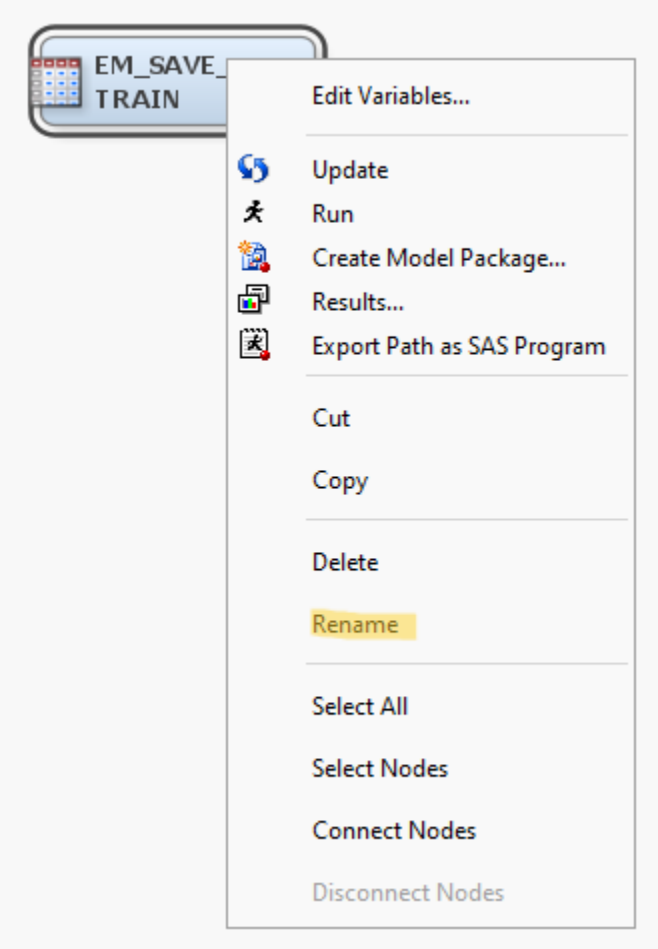


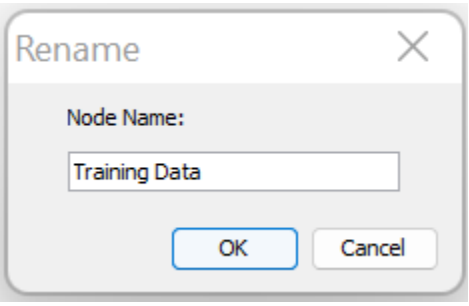
Figures 12.2.15 - 12.2.17 Use Created Data Source

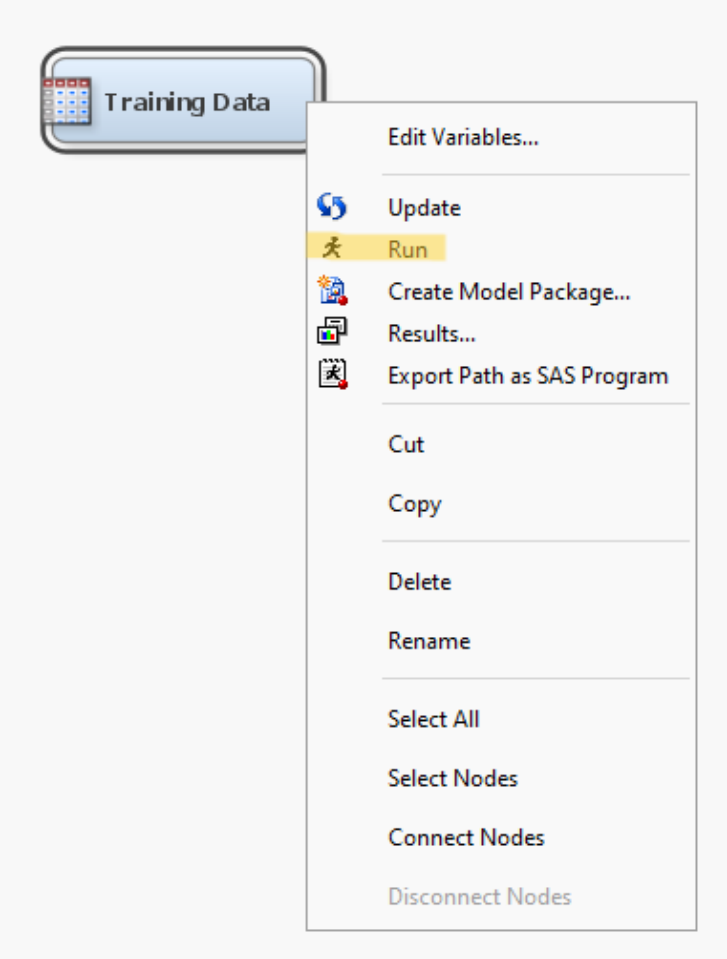


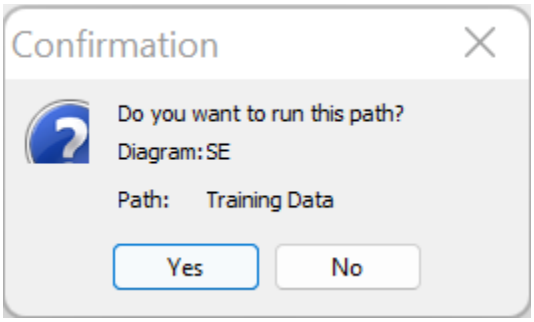
Property	Value
ID	emsavetrain
Name	EM_SAVE_TRAIN
Variables	
Decisions	
Role	Raw
Notes	
Library	DATASET
Table	EM_SAVE_TRAIN
Sample Data Set	
Size Type	
Sample Size	
Type	DATA
No. Obs	100000
No. Cols	28
No. Bytes	42599424
Segment	
Created By	u63409714
Create Date	5/6/23 3:28 AM
Modified By	u63409714
Modify Date	5/6/23 3:28 AM

Figures 12.2.18 - 12.2.22 Rename Data Source

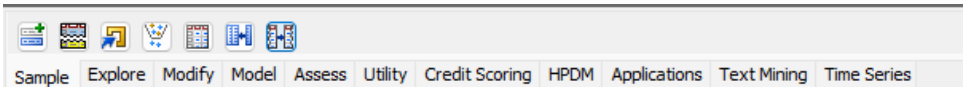








Figures 12.2.23 - 12.2.24 Sample Node in Sample Tab



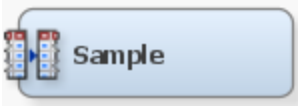
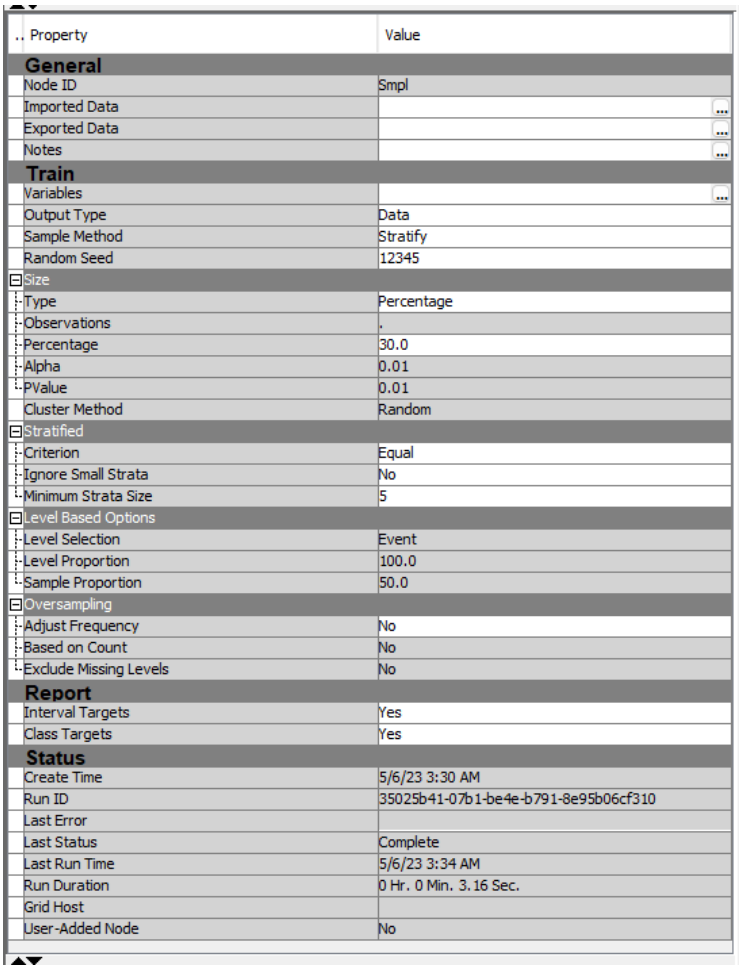


Figure 12.2.25 Default Setting of Sample Node

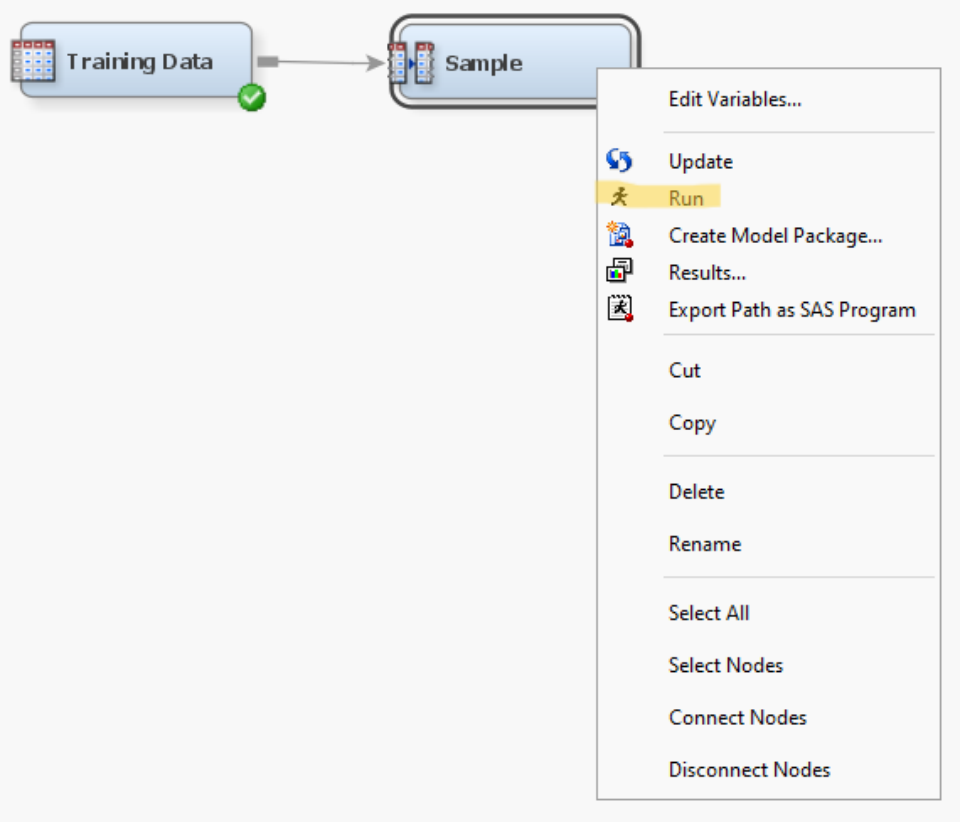
Property	Value
General	
Node ID	Smpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Sample Method	Default
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	10.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Proportional
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	5/8/23 10:09 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

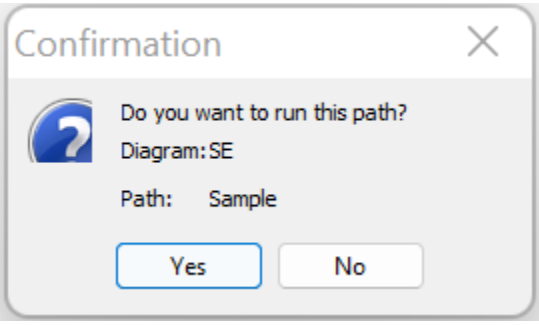
Figure 12.2.26 Modified Setting of Sample Node



Property	Value
General	
Node ID	Smpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Sample Method	Stratify
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	30.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	5/6/23 3:30 AM
Run ID	35025b41-07b1-be4e-b791-8e95b06cf310
Last Error	
Last Status	Complete
Last Run Time	5/6/23 3:34 AM
Run Duration	0 Hr. 0 Min. 3.16 Sec.
Grid Host	
User-Added Node	No

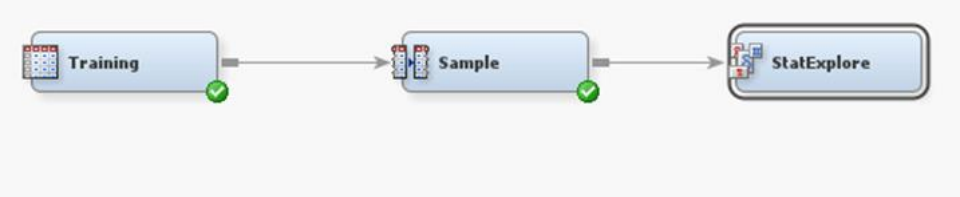
Figures 12.2.27 - 12.2.29 Perform Data Sampling

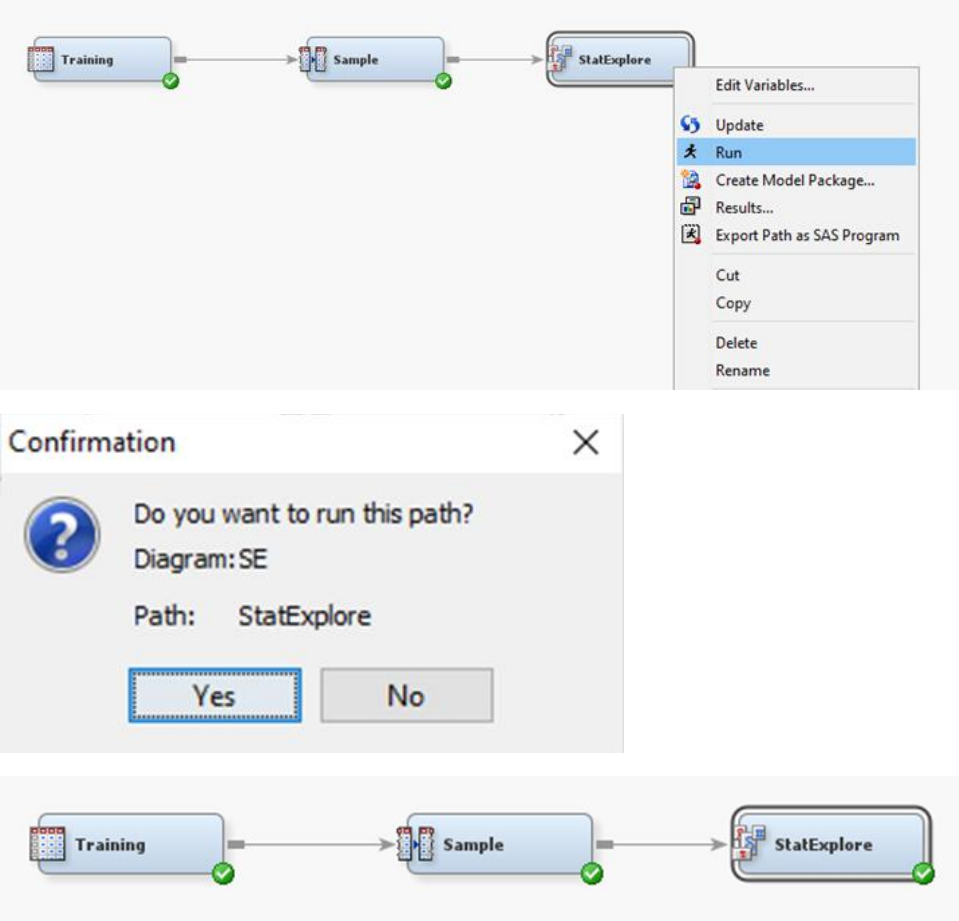




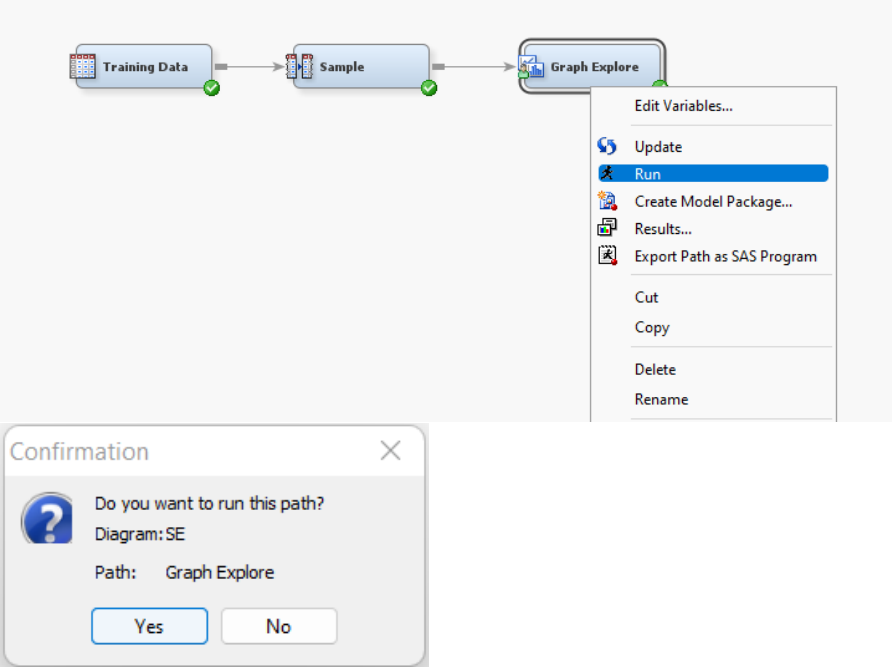
12.3 Explore

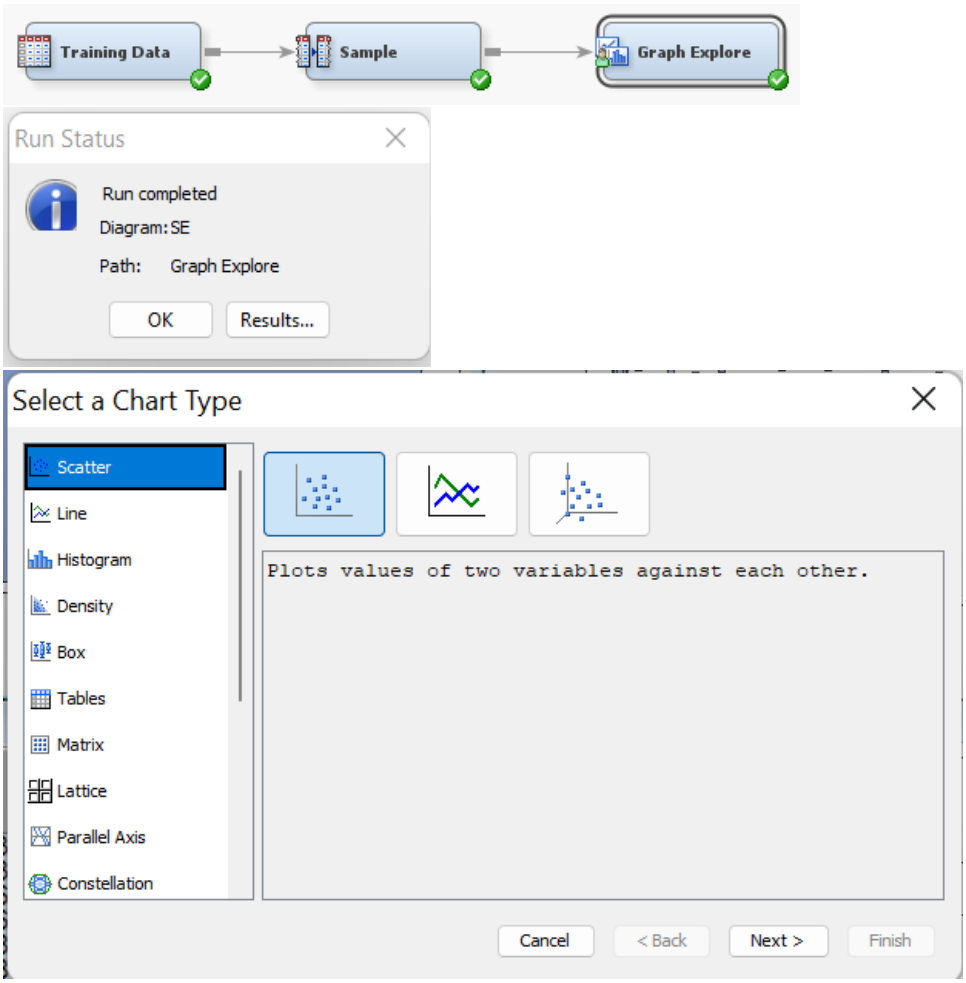
Figures 12.3.1 - 12.3.4 Generate summary statistics



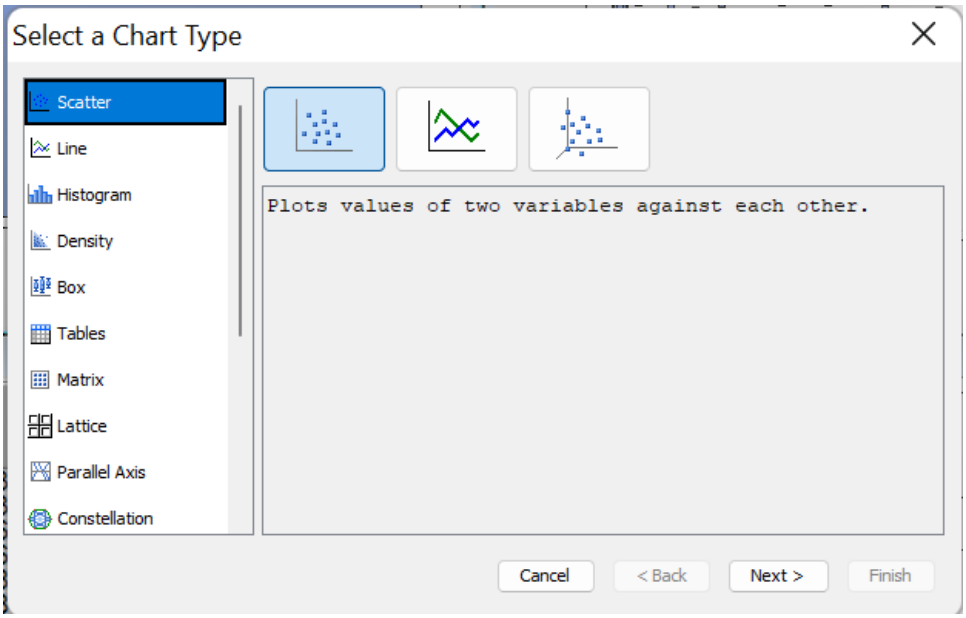


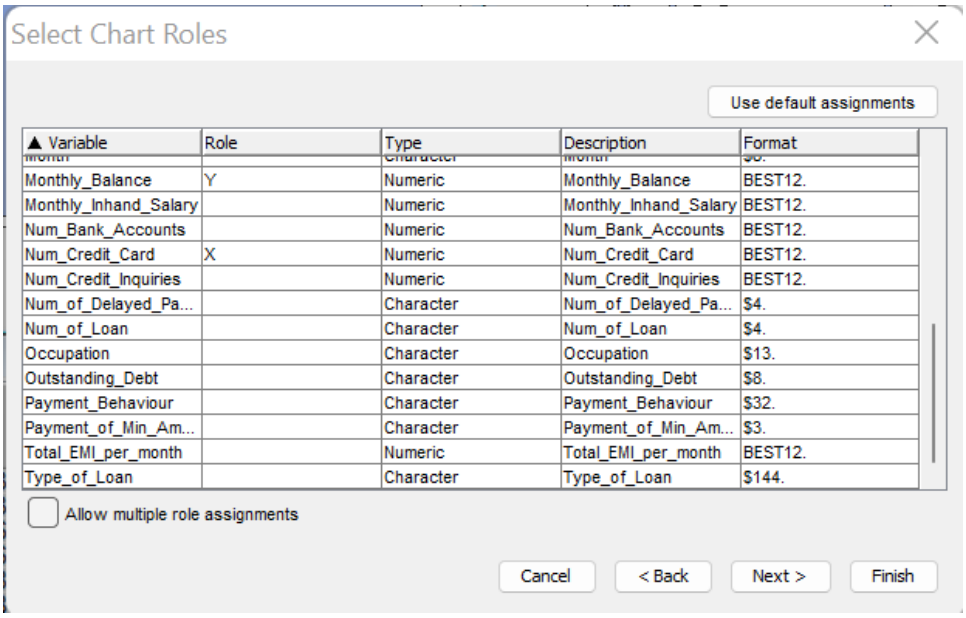
Figures 12.3.5 - 12.3.9: Univariate: Generate graphs (from GraphExplore node results page)





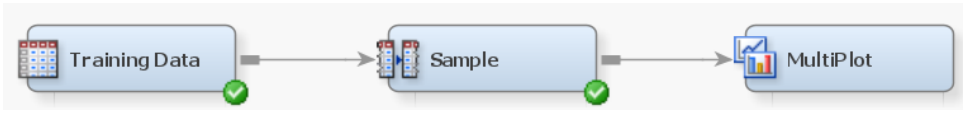
Figures 12.3.10 - 12.3.11: Bivariate (Input Variables): Generate Plot (from GraphExplore node results page)



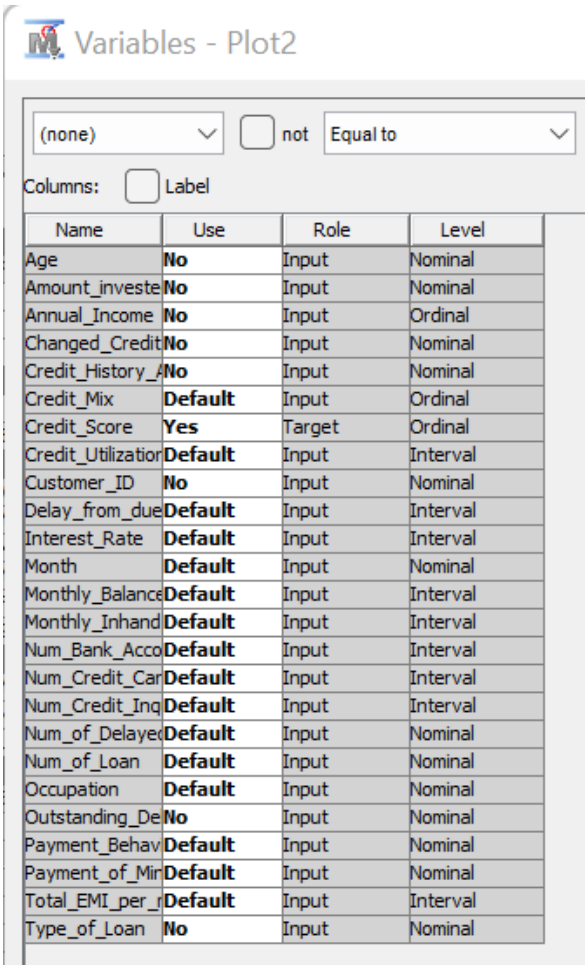


▲ Variable	Role	Type	Description	Format
Monthly_Balance	Y	Numeric	Monthly_Balance	BEST12.
Monthly_Inhand_Salary		Numeric	Monthly_Inhand_Salary	BEST12.
Num_Bank_Accounts		Numeric	Num_Bank_Accounts	BEST12.
Num_Credit_Card	X	Numeric	Num_Credit_Card	BEST12.
Num_Credit_Inquiries		Numeric	Num_Credit_Inquiries	BEST12.
Num_of_Delayed_Pa...		Character	Num_of_Delayed_Pa...	$4.
Num_of_Loan		Character	Num_of_Loan	$4.
Occupation		Character	Occupation	$13.
Outstanding_Debt		Character	Outstanding_Debt	$8.
Payment_Behaviour		Character	Payment_Behaviour	$32.
Payment_of_Min_Am...		Character	Payment_of_Min_Am...	$3.
Total_EMI_per_month		Numeric	Total_EMI_per_month	BEST12.
Type_of_Loan		Character	Type_of_Loan	$144.

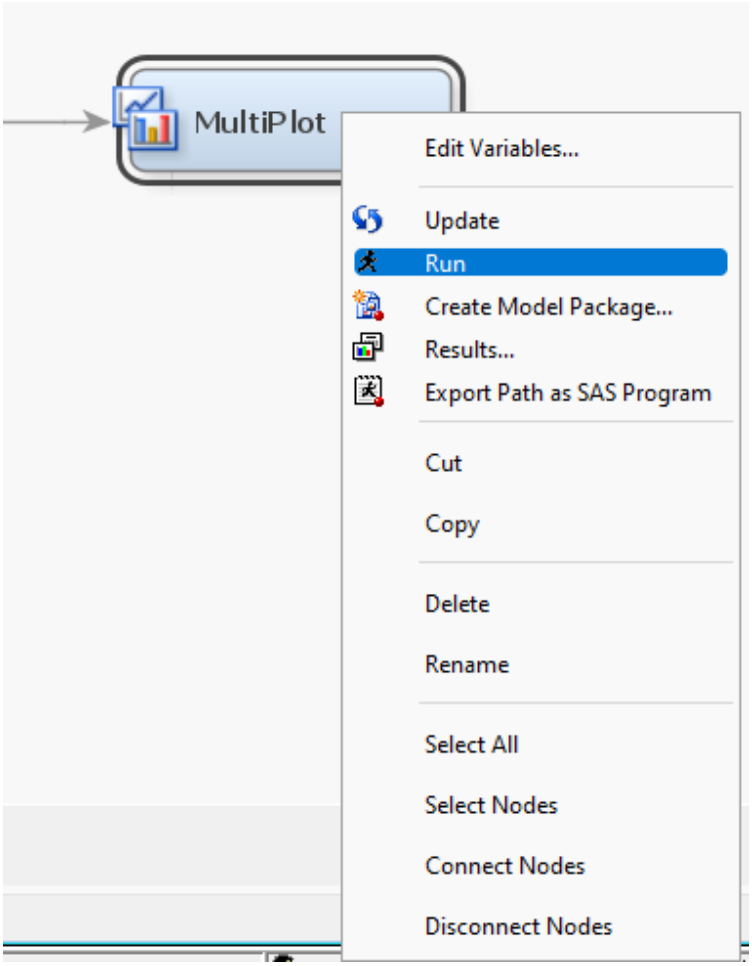
Figures 12.3.12 - 12.3.18: Bivariate (Input and Target Variables): Generate Plot

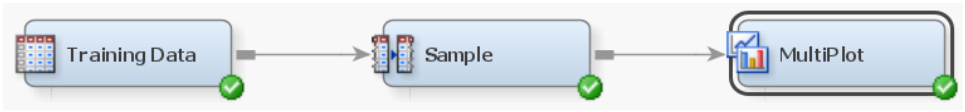


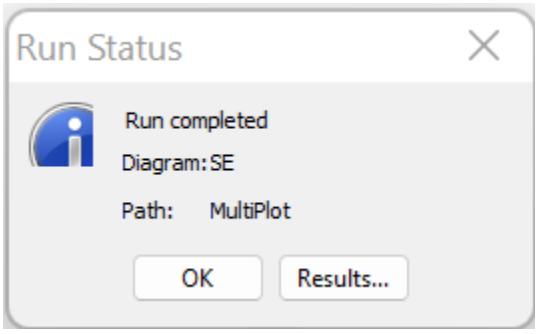
Property	Value
General	
Node ID	Plot2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Type of Charts	Both
Bar Chart Options	
Graph Orientation	Vertical
Include Missing Values	Yes
Interval Target Charts	Mean
Show Values	Yes
Statistic	Freq
Numeric Threshold	20
Scatter Options	
Confidence Interval	Yes
Regression Equation	No
Regression Type	Linear
Status	
Create Time	5/13/23 1:52 AM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

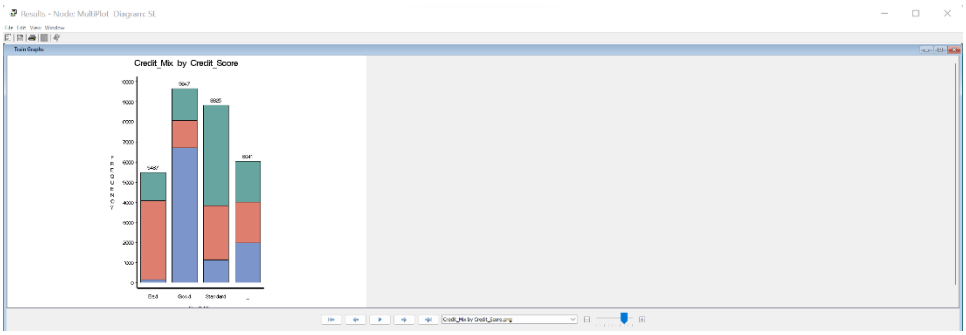


Name	Use	Role	Level
Age	**No**	Input	Nominal
Amount_investe	**No**	Input	Nominal
Annual_Income	**No**	Input	Ordinal
Changed_Credit	**No**	Input	Nominal
Credit_History_A	**No**	Input	Nominal
Credit_Mix	**Default**	Input	Ordinal
Credit_Score	**Yes**	Target	Ordinal
Credit_Utilizatior	**Default**	Input	Interval
Customer_ID	**No**	Input	Nominal
Delay_from_due	**Default**	Input	Interval
Interest_Rate	**Default**	Input	Interval
Month	**Default**	Input	Nominal
Monthly_Balance	**Default**	Input	Interval
Monthly_Inhand	**Default**	Input	Interval
Num_Bank_Acco	**Default**	Input	Interval
Num_Credit_Car	**Default**	Input	Interval
Num_Credit_Inq	**Default**	Input	Interval
Num_of_Delaye	**Default**	Input	Nominal
Num_of_Loan	**Default**	Input	Nominal
Occupation	**Default**	Input	Nominal
Outstanding_De	**No**	Input	Nominal
Payment_Behav	**Default**	Input	Nominal
Payment_of_Mir	**Default**	Input	Nominal
Total_EMI_per_	**Default**	Input	Interval
Type_of_Loan	**No**	Input	Nominal

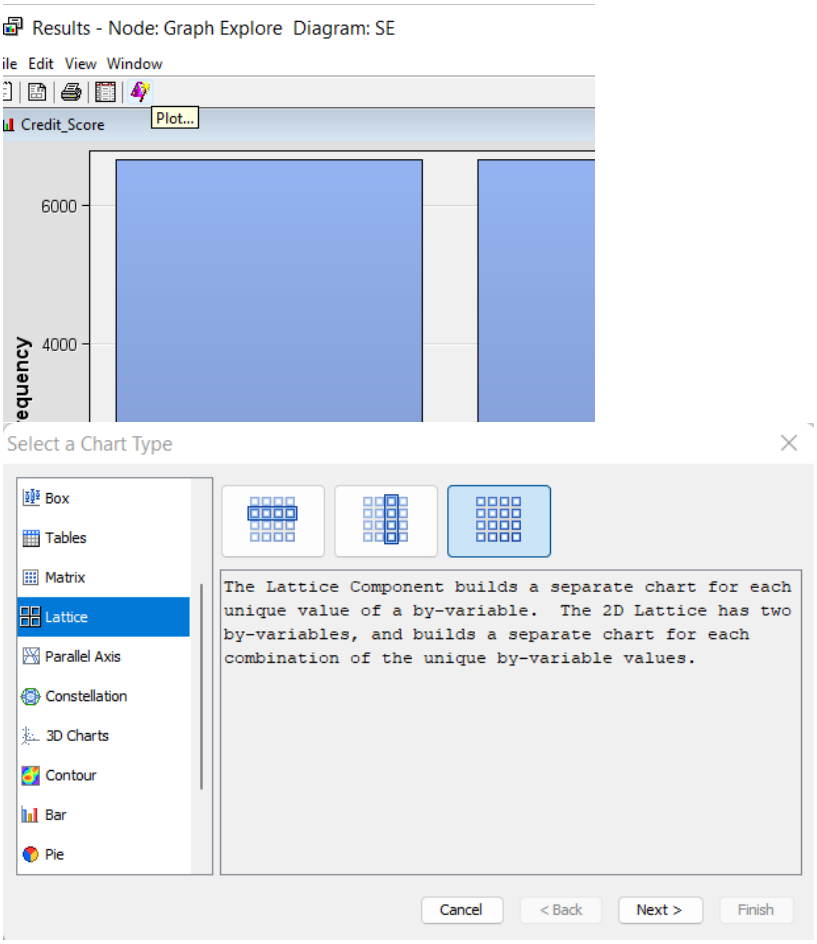

MultiPlot
Edit Variables...
Update
Run
Create Model Package...
Results...
Export Path as SAS Program
Cut
Copy
Delete
Rename
Select All
Select Nodes
Connect Nodes
Disconnect Nodes

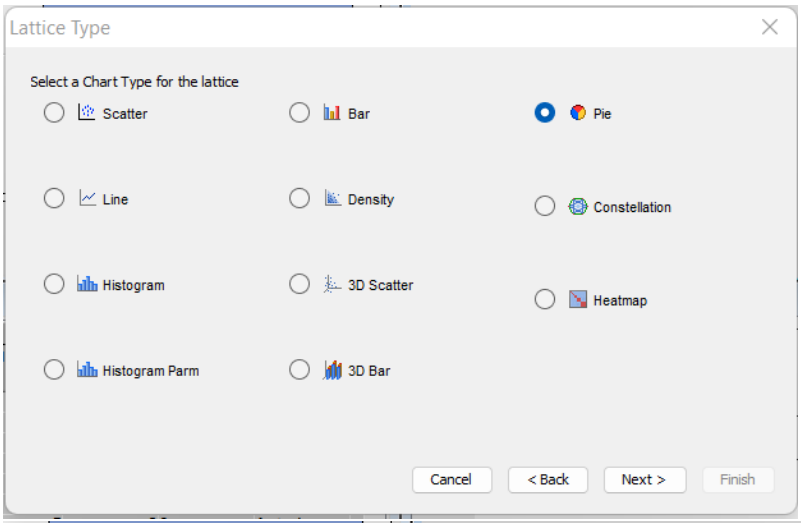

Training Data
Sample
MultiPlot

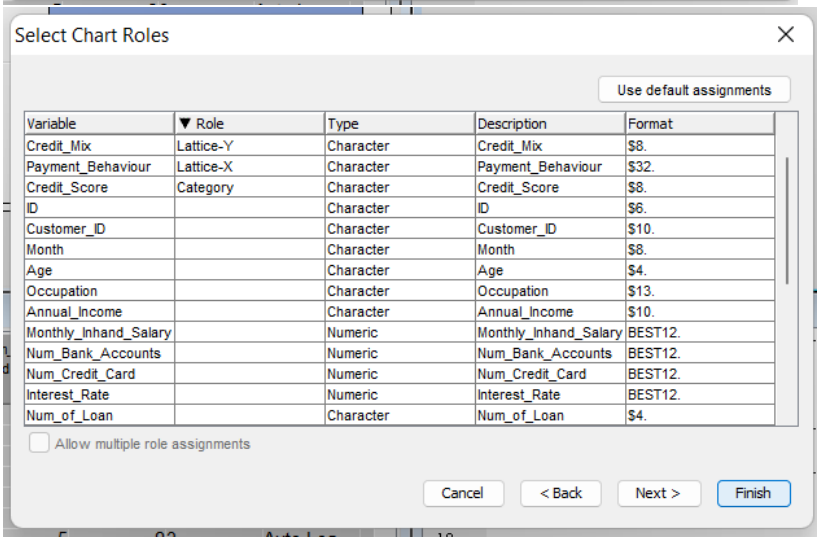

Run Status
Run completed
Diagram: SE
Path: MultiPlot
OK
Results...



Figures 12.3.19 - 12.3.22: Multivariate: Generate a Lattice Plot (from GraphExplore node results page)

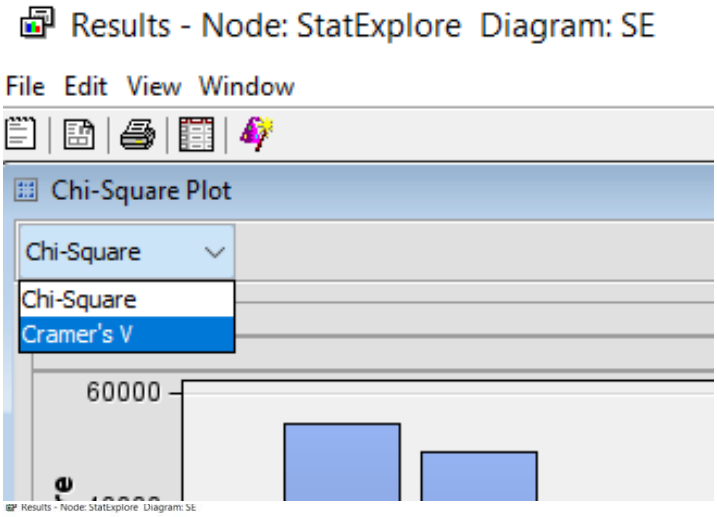


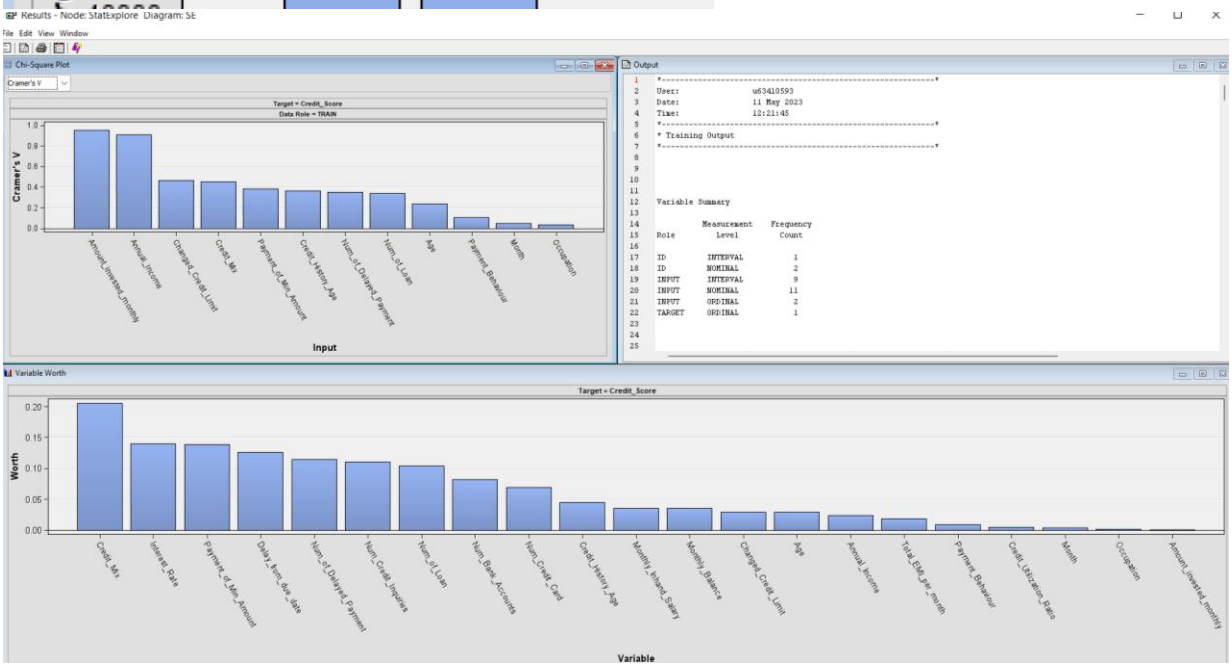




Variable	▼ Role	Type	Description	Format
Credit_Mix	Lattice-Y	Character	Credit_Mix	$8.
Payment_Behaviour	Lattice-X	Character	Payment_Behaviour	$32.
Credit_Score	Category	Character	Credit_Score	$8.
ID		Character	ID	$6.
Customer_ID		Character	Customer_ID	$10.
Month		Character	Month	$8.
Age		Character	Age	$4.
Occupation		Character	Occupation	$13.
Annual_Income		Character	Annual_Income	$10.
Monthly_Inhand_Salary		Numeric	Monthly_Inhand_Salary	BEST12.
Num_Bank_Accounts		Numeric	Num_Bank_Accounts	BEST12.
Num_Credit_Card		Numeric	Num_Credit_Card	BEST12.
Interest_Rate		Numeric	Interest_Rate	BEST12.
Num_of_Loan		Character	Num_of_Loan	$4.

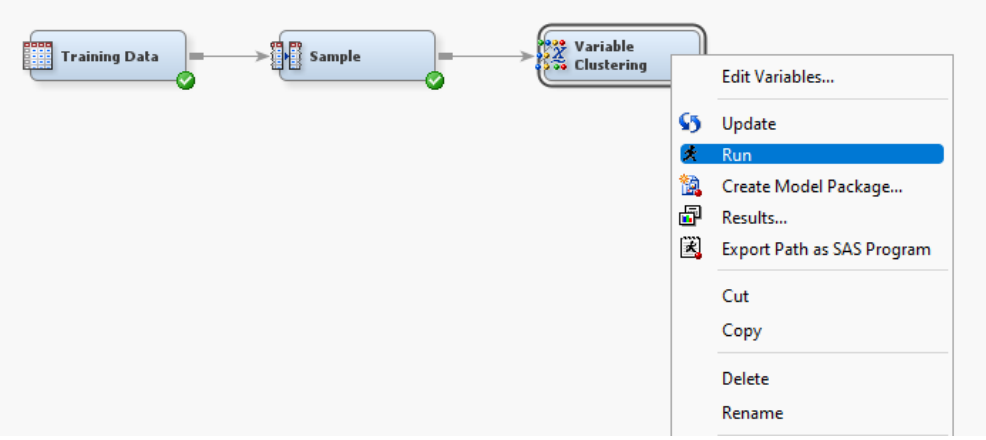
Figures 12.3.23 - 12.3.24: Multivariate: Generate Cramer's V (from StatExplore node results page)

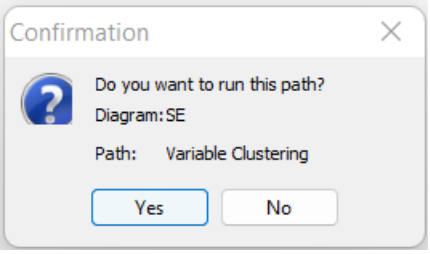


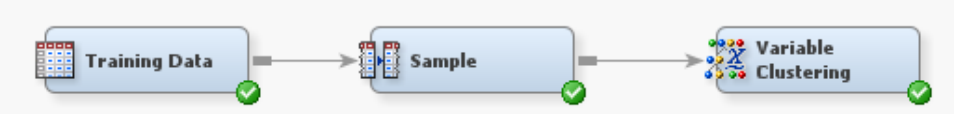


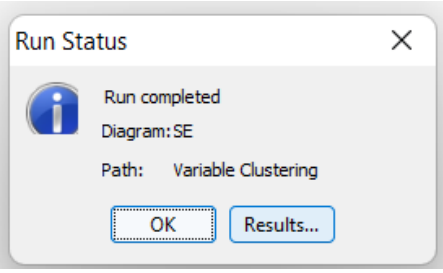
Figures 12.3.25 - 12.3.31: Multivariate: Variable Clustering

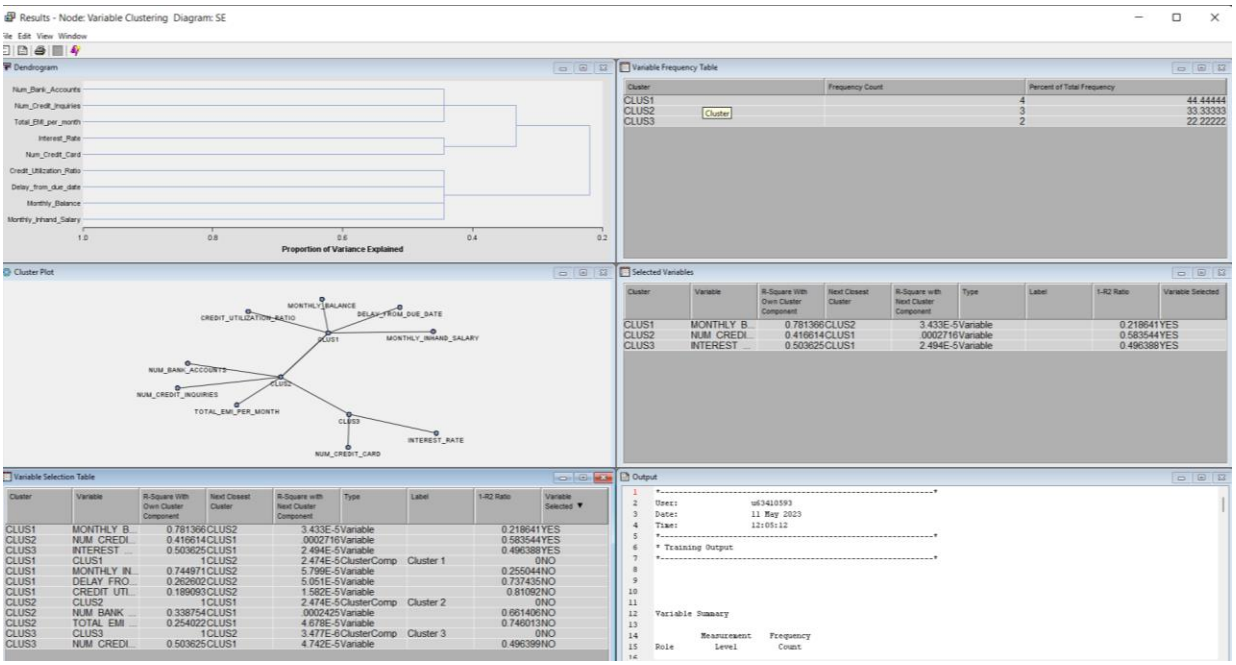
Score	
Variable Selection	Best Variables
Interactive Selection	...
Hides Rejected Variables	Yes


Training Data
Sample
Variable Clustering
Edit Variables...
Update
Run
Create Model Package...
Results...
Export Path as SAS Program
Cut
Copy
Delete
Rename

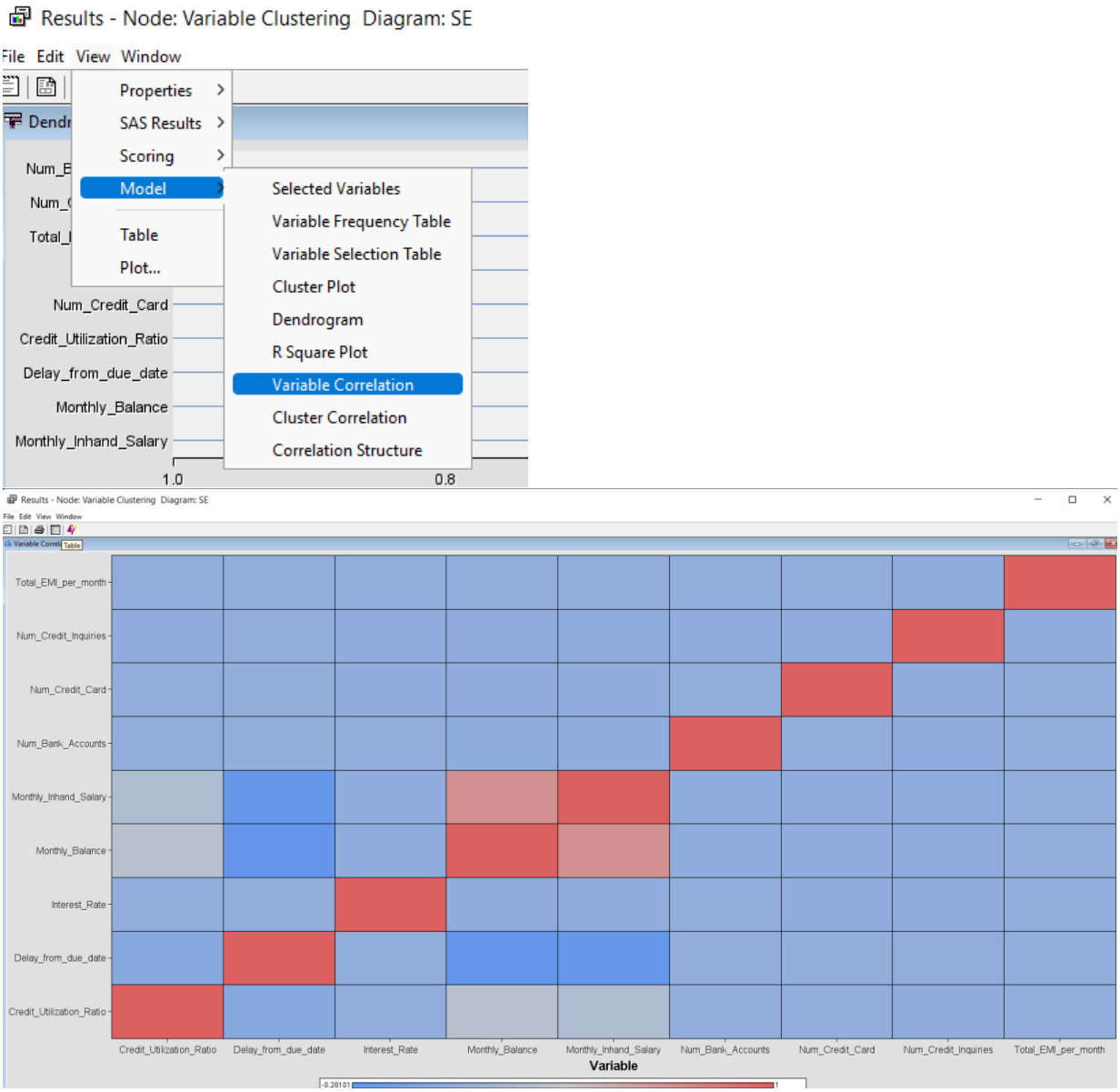

Confirmation
Do you want to run this path?
Diagram: SE
Path: Variable Clustering
Yes
No


Training Data
Sample
Variable Clustering

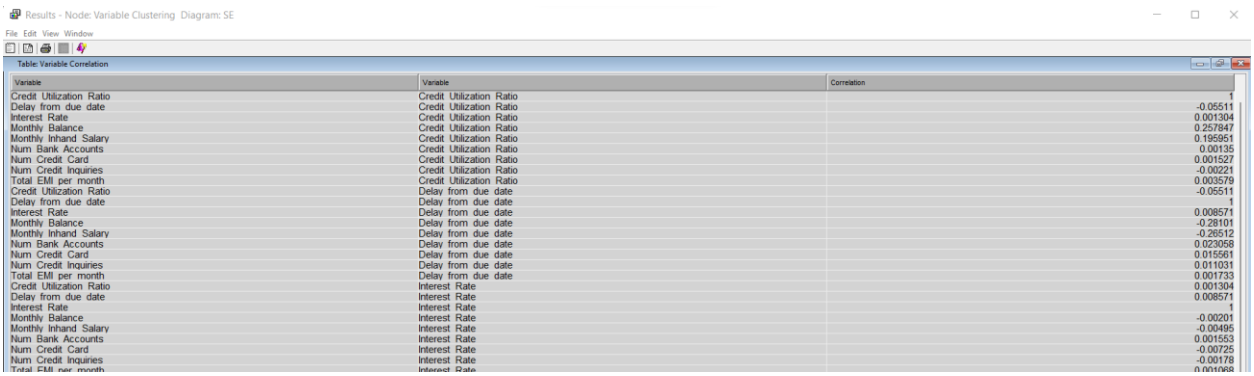

Run Status
Run completed
Diagram: SE
Path: Variable Clustering
OK
Results...



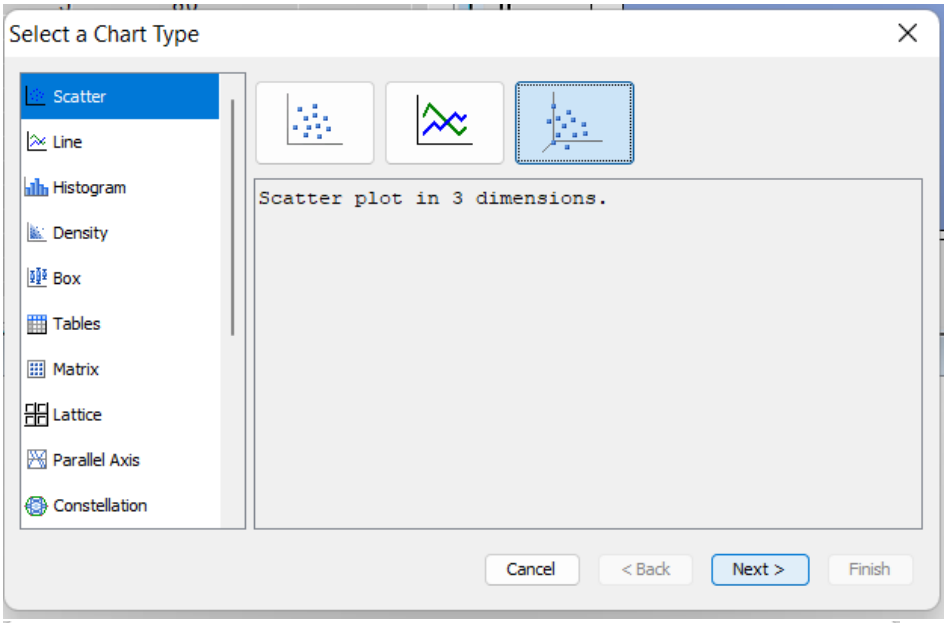
Figures 12.3.32 - 12.3.34: Multivariate: Correlation Matrix (from Variable Clustering node results page)



The correlation table is generated from the correlation matrix page.



Figures 12.3.35 - 12.3.38: Visualizing 3-D Scatter plot (from GraphExplore node plot page)



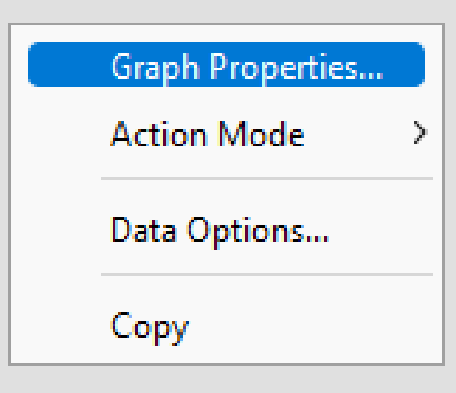
Select Chart Roles

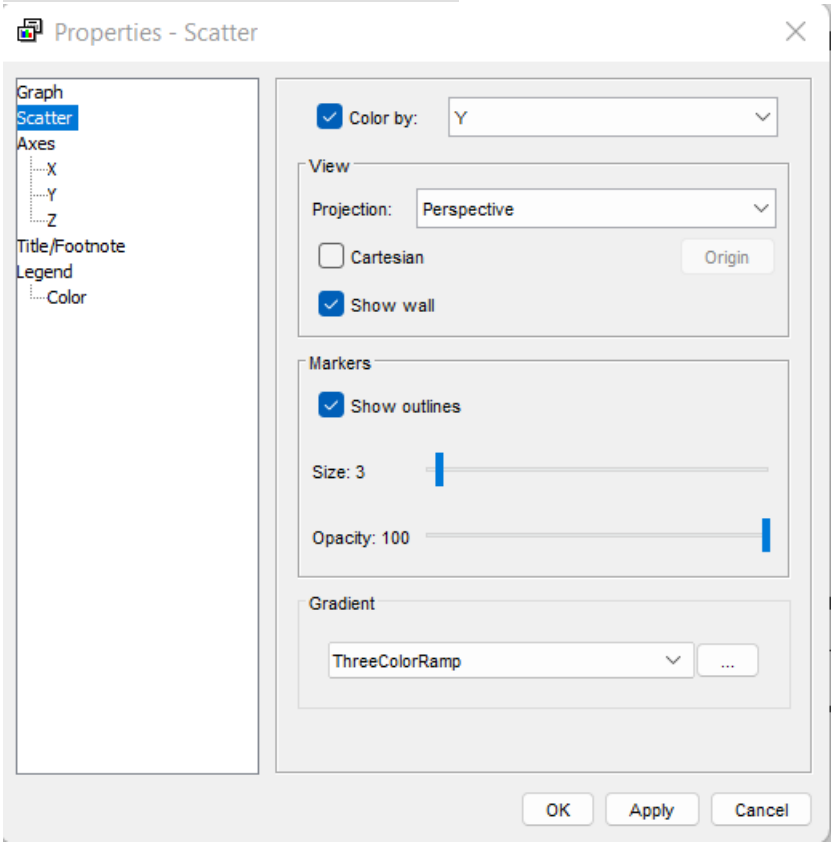
Use default assignments

Variable	▼ Role	Type	Description	Format
Monthly_Balance	Z	Numeric	Monthly_Balance	BEST12.
Delay_from_due_date	Y	Numeric	Delay_from_due_date	BEST12.
Monthly_Inhand_Salary	X	Numeric	Monthly_Inhand_Salary	BEST12.
ID		Character	ID	$6.
Customer_ID		Character	Customer_ID	$10.
Month		Character	Month	$8.
Age		Character	Age	$4.
Occupation		Character	Occupation	$13.
Annual_Income		Character	Annual_Income	$10.
Num_Bank_Accounts		Numeric	Num_Bank_Accounts	BEST12.
Num_Credit_Card		Numeric	Num_Credit_Card	BEST12.
Interest_Rate		Numeric	Interest_Rate	BEST12.
Num_of_Loan		Character	Num_of_Loan	$4.
Type_of_Loan		Character	Type_of_Loan	$144.

Allow multiple role assignments

Cancel < Back Next > Finish





12.4 Modify: Pre-modifying in Base SAS

Figure 12.4.1: Create library "cleaning"

```
/* create library */
options dlcreatedir;
LIBNAME cleaning '/home/u63409714/cleaning';
```

Figures 12.4.2 - 12.4.3: SAS Macro to remove special characters in each character variables

```sql
/* Remove any special characters */
/* Macro */
option mprint;
%macro rm_spec_char(in_tbl, out_tbl);
%let in_tbl = %upcase(&in_tbl);

/* Count of char variables*/
proc sql noprint;
    select count(*)
    into
        :colcnt trimmed
    from dictionary.columns
    where libname = "CLEANING" and memname = "&in_tbl" and type = "char";
quit;

%put &colcnt;

/* Create macro variable for each char variable */
proc sql noprint;
    select name
    into
        :col1-:col&colcnt
    from dictionary.columns
    where libname = "CLEANING" and memname = "&in_tbl" and type = "char";
quit;

/* Loop through each column and remove all special char */
data &out_tbl;
    set CLEANING.&in_tbl.;
        %do i=1 %to &colcnt.;
            "&&col&i"n = compress("&&col&i"n, ",.", "kads");
        %end;
run;

%mend rm spec char;
```

```sql
%rm_spec_char(SMPL_DATA, CLEANING.SMPL_DATA_RC);
```

Figure 12.4.4: Create new unique id

```sql
/* Create new unique id */
data CLEANING.SMPL_DATA_NID;
  set CLEANING.SMPL_DATA_RC;
  id_new = _n_;
run;
```

Figure 12.4.5: Variables Level Reclassification and Formatting

```sql
/* Variables Level Reclassification */
/* Convert character to numeric */
data CLEANING.SMPL_DATA_VLR (drop = Age Annual_Income Num_of_Loan Num_of_Delayed_Payment Changed_Credit_Limit
                                    Outstanding_Debt Amount_invested_monthly);
    set CLEANING.SMPL_DATA_NID;
    Age_2 = input(Age, COMMA3.);
    Annual_Income_2 = round(input(Annual_Income, COMMA20.),.01);
    Num_of_Loan_2 = input(Num_of_Loan, COMMA20.);
    Num_of_Delayed_Payment_2 = input(Num_of_Delayed_Payment, COMMA20.);
    Changed_Credit_Limit_2 = round(input(Changed_Credit_Limit, COMMA20.),.01);
    Outstanding_Debt_2 = round(input(Outstanding_Debt, COMMA20.),.01);
    Amount_invested_monthly_2 = round(input(Amount_invested_monthly, COMMA20.),.01);
    Monthly_Balance_2 = round(Monthly_Balance,.01);
    Total_EMI_per_month_2 = round(Total_EMI_per_month,.01);
    Credit_Utilization_Ratio_2 = round(Credit_Utilization_Ratio,.01);
    Monthly_Inhand_Salary_2 = round(Monthly_Inhand_Salary,.01);
run;
```

Figure 12.4.6: Convert Credit History Age to years format

```
/* Convert  Credit History Age to years format*/
data CLEANING.SMPL_DATA_CHA (drop = Credit_History_Age);
    set CLEANING.SMPL_DATA_VLR;

    if find(TRIM(Credit_History_Age), "Years") ge 1
        then Credit_History_Age_Yr = INPUT(SUBSTR(TRIM(Credit_History_Age), 1, find(TRIM(Credit_History_Age), "Years")-2),
                                    COMMA3.);

    if find(TRIM(Credit_History_Age), "Months") ge 1
        then Credit_History_Age_Mth = INPUT(SUBSTR(TRIM(Credit_History_Age), find(TRIM(Credit_History_Age), "and")+4,
                                        find(TRIM(Credit_History_Age), "Months")-1-(find(TRIM(Credit_History_Age),
                                        "and")+4)), COMMA2.);

    Credit_History_Age_2 = round(Credit_History_Age_Yr + Credit_History_Age_Mth/12, 0.1);
run;
```

Figure 12.4.7: Transpose Type of Loan

```
/* Transpose Type of Loan*/
/* Get all loans in a column */
data Type_of_Loan;
    set CLEANING.SMPL_DATA_CHA (keep = id_new Type_of_Loan);

    length all_Type_of_Loan $1000;

    Type_of_Loan_2 = tranwrd(TRIM(Type_of_Loan),'and','');
    do i=1 to countw(Type_of_Loan_2,',');
        all_Type_of_Loan =TRIM(scan(Type_of_Loan_2,i,","));
        output;
    end;
    drop i;
run;
```

* Note that this dataset named "Type_of_Loan" only contains id_new and Type of Loan variable.

Figure 12.4.8: Get distinct loans and clean dirty data

```
/* Get distinct loans */
proc sort data = Type_of_Loan (keep = all_Type_of_Loan) out = all_Type_of_Loan nodupkey;
    by all_Type_of_Loan;
run;

proc sql;
create table Type_of_Loan_cleaned as
    select id_new,
        case
            when upcase(strip(all_Type_of_Loan)) like "AU%" then "Auto Loan"
            when upcase(strip(all_Type_of_Loan)) like "C%" then "CreditBuilder Loan"
            when upcase(strip(all_Type_of_Loan)) like "DE%" then "Debt Consolidation Loan"
            when upcase(strip(all_Type_of_Loan)) like "HO%" then "Home Equity Loan"
            when upcase(strip(all_Type_of_Loan)) like "M%" then "Mortgage Loan"
            when upcase(strip(all_Type_of_Loan)) like "N%" then "Not Specified"
            when upcase(strip(all_Type_of_Loan)) like "PA%" then "Payday Loan"
            when upcase(strip(all_Type_of_Loan)) like "PE%" then "Personal Loan"
            when upcase(strip(all_Type_of_Loan)) like "S%" then "Student Loan"
        end as loan
    from WORK.Type_of_Loan
    where upcase(trim(all_Type_of_Loan)) not in ("A","AN","P")
    and not missing(trim(all_Type_of_Loan))
    group by id_new, loan;
quit;
```

Figure 12.4.9: Create new variables to tag whether customer hold a loan

```
proc sql;
create table Type_of_Loan_cleaned_2 as
    select id_new,
        SUM(case when loan = "Auto Loan" then 1 else 0 end) as auto_loan,
        SUM(case when loan = "CreditBuilder Loan" then 1 else 0 end) as creditbuilder_loan,
        SUM(case when loan = "Debt Consolidation Loan" then 1 else 0 end) as debt_con_loan,
        SUM(case when loan = "Home Equity Loan" then 1 else 0 end) as home_equi_loan,
        SUM(case when loan = "Mortgage Loan" then 1 else 0 end) as mg_loan,
        SUM(case when loan = "Not Specified" then 1 else 0 end) as not_spec_loan,
        SUM(case when loan = "Payday Loan" then 1 else 0 end) as payday_loan,
        SUM(case when loan = "Personal Loan" then 1 else 0 end) as personal_loan,
        SUM(case when loan = "Student Loan" then 1 else 0 end) as student_loan
    from WORK.Type_of_Loan_cleaned
    group by id_new
    order by id_new;
quit;

data Type_of_Loan_cleaned_3;
    set Type_of_Loan_cleaned_2;
    if auto_loan ge 1 then auto_loan_f = "Y"; else auto_loan_f = "N";
    if creditbuilder_loan ge 1 then creditbuilder_loan_f = "Y"; else creditbuilder_loan_f = "N";
    if debt_con_loan ge 1 then debt_con_loan_f = "Y"; else debt_con_loan_f = "N";
    if home_equi_loan ge 1 then home_equi_loan_f = "Y"; else home_equi_loan_f = "N";
    if mg_loan ge 1 then mg_loan_f = "Y"; else mg_loan_f = "N";
    if not_spec_loan ge 1 then not_spec_loan_f = "Y"; else not_spec_loan_f = "N";
    if payday_loan ge 1 then payday_loan_f = "Y"; else payday_loan_f = "N";
    if personal_loan ge 1 then personal_loan_f = "Y"; else personal_loan_f = "N";
    if student_loan ge 1 then student_loan_f = "Y"; else student_loan_f = "N";
run;
```

Figure 12.4.10: Merge transformed Type of Loan with main dataset

```
/* Merge Type_of_Loan_cleaned_3 and CLEANING.SMPL_DATA_CHA */
proc sort data = CLEANING.SMPL_DATA_CHA;
    by id_new;
quit;

data CLEANING.SMPL_DATA_MERGED (drop = ID Credit_History_Age_Mth Credit_History_Age_Yr Type_of_Loan auto_loan
                                       creditbuilder_loan debt_con_loan home_equi_loan mg_loan not_spec_loan
                                       payday_loan personal_loan student_loan);
merge CLEANING.SMPL_DATA_CHA(in=a) Type_of_Loan_cleaned_3(in=b);
by id_new;
if a;
run;
```

Figure 12.4.11: Convert Month to numeric variable

```
/* Reformat month */
data CLEANING.SMPL_DATA_ReF (drop = Month);
    set CLEANING.SMPL_DATA_MERGED;
    if Month = "January" then Month_2 = 1;
    else if Month = "February" then Month_2 = 2;
    else if Month = "March" then Month_2 = 3;
    else if Month = "April" then Month_2 = 4;
    else if Month = "May" then Month_2 = 5;
    else if Month = "June" then Month_2 = 6;
    else if Month = "July" then Month_2 = 7;
    else if Month = "August" then Month_2 = 8;
    else if Month = "September" then Month_2 = 9;
    else if Month = "October" then Month_2 = 10;
    else if Month = "November" then Month_2 = 11;
    else if Month = "December" then Month_2 = 12;
run;
```

Figure 12.4.12: Rename columns

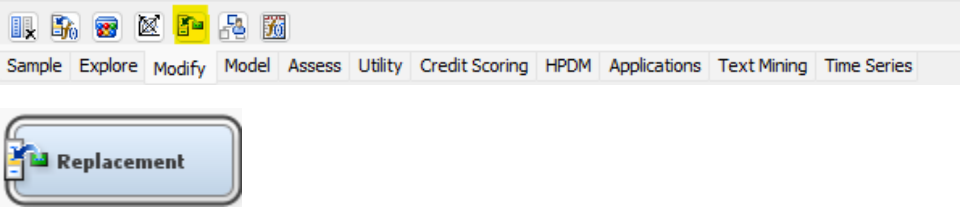
```
/* Rename columns */
data CLEANING.SMPL_DATA_FINAL;
    set CLEANING.SMPL_DATA_ReF (rename=(id_new=id Month_2=Month Age_2=Age
                                        Amount_invested_monthly_2=Amount_invested_monthly
                                        Annual_Income_2=Annual_Income Changed_Credit_Limit_2=Changed_Credit_Limit
                                        Credit_History_Age_2=Credit_History_Age
                                        Num_of_Delayed_Payment_2=Num_of_Delayed_Payment
                                        Num_of_Loan_2=Num_of_Loan Outstanding_Debt_2=Outstanding_Debt
                                        Monthly_Balance_2=Monthly_Balance Total_EMI_per_month_2=Total_EMI_per_month
                                        Credit_Utilization_Ratio_2=Credit_Utilization_Ratio
                                        Monthly_Inhand_Salary_2=Monthly_Inhand_Salary));
run;
```

Figure 12.4.13: Save the modified dataset as "em_save_train_modified"

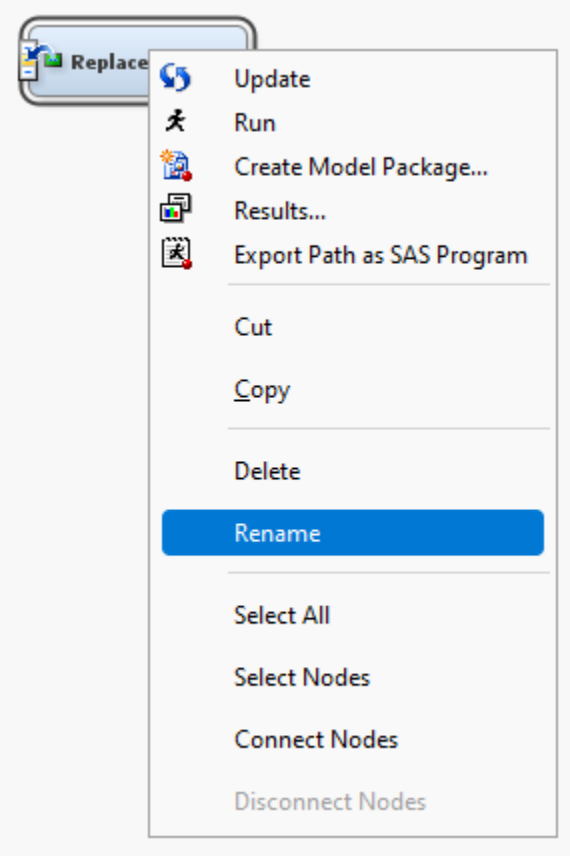
```
228 libname mydata '/home/u63409714/Credit Score Classification/Workspaces/EMWS1/EMSave';
229 data mydata.em_save_train_modified;
230     set CLEANING.SMPL_DATA_FINAL (drop=_dataobs_);
231 run;
```

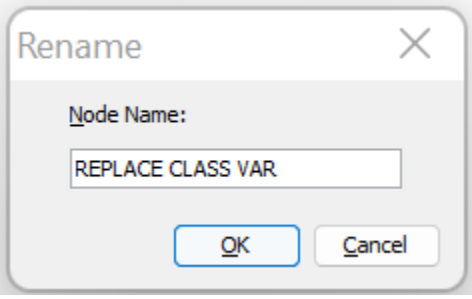
12.5 Modify: Modifying dataset using SAS Enterprise Miner

Figures 12.5.1 - 12.5.2: Replacement node in Modify tab



Figures 12.5.3 - 12.5.5: Rename Replacement node to “REPLACE CLASS VAR”





Figures 12.5.6 - 12.5.8: Replace Class Variables using Replacement Editor under Class Variables

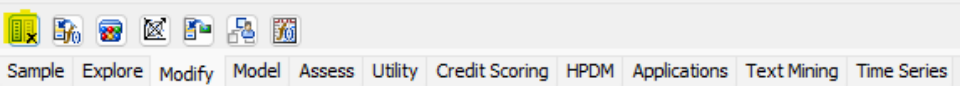
Property	Value
General	
Node ID	Repl
Imported Data	
Exported Data	
Notes	
Train	
Interval Variables	
Replacement Editor	
Default Limits Method	None
Cutoff Values	
Class Variables	
Replacement Editor	
Unknown Levels	Missing Value
Score	
Replacement Values	Missing
Hide	No
Report	
Replacement Report	Yes
Status	
Create Time	5/24/23 10:59 AM
Run ID	23b51a54-54fa-6d4f-ad3f-f98f0dc089a2
Last Error	
Last Status	Complete
Last Run Time	5/29/23 1:55 PM
Run Duration	0 Hr. 0 Min. 8.18 Sec.
Grid Host	
User-Added Node	No

Figure 12.5.9: Rename Replacement node to “REPLACE INT VAR”

Figures 12.5.10 - 12.5.11: Replace Interval Variables using Replacement Editor under Interval Variables

Property	Value
General	
Node ID	Repl2
Imported Data	
Exported Data	
Notes	
Train	
Interval Variables	
Replacement Editor	
Default Limits Method	User-Specified Limits
Cutoff Values	
Class Variables	
Replacement Editor	
Unknown Levels	Ignore
Score	
Replacement Values	Missing
Hide	No
Report	
Replacement Report	Yes
Status	
Create Time	5/24/23 12:28 PM
Run ID	2190dfba-aad2-e045-b411-7bd7895dfa0f
Last Error	
Last Status	Complete
Last Run Time	5/29/23 1:55 PM
Run Duration	0 Hr. 0 Min. 10.24 Sec.
Grid Host	
User-Added Node	No

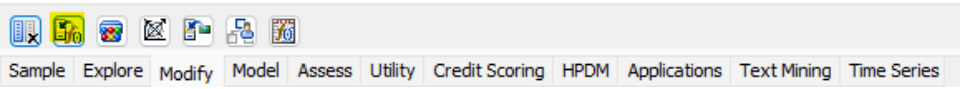
Figures 12.5.12 - 12.5.13: Drop node in Modify tab



Figures 12.5.14 - 12.5.15: Drop unused variables

Property	Value
General	
Node ID	Drop
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Drop Selection Options	
Drop from Tables	No
Assess	No
Classification	No
Frequency	No
Hidden	Yes
Input	No
Predict	No
Rejected	Yes
Residual	No
Target	No
Other	No
Status	
Create Time	5/24/23 2:32 PM
Run ID	16dc5171-6ad0-084b-b8c7-6e45251d5b82
Last Error	
Last Status	Complete
Last Run Time	5/29/23 1:55 PM
Run Duration	0 Hr. 0 Min. 1.97 Sec.
Grid Host	
User-Added Node	No

Figures 12.5.16 - 12.5.17: Impute node in Modify tab

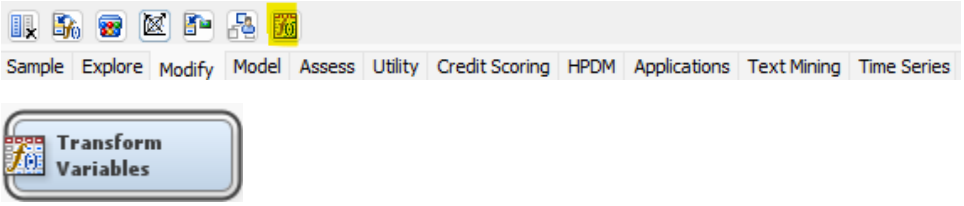


Figures 12.5.18 - 12.5.19: Impute missing variables

Property	Value
General	
Node ID	Impt
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Nonmissing Variables	Yes
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	
Tree Imputation	
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected
Report	
Validation and Test Data	No
Distribution of Missing	No
Status	
Create Time	5/24/23 11:07 AM
Run ID	8420a198-010b-154e-9acf-0ee516c151c2
Last Error	
Last Status	Complete
Last Run Time	5/29/23 1:55 PM
Run Duration	0 Hr. 0 Min. 12.68 Sec.
Grid Host	
User-Added Node	No

Figure 12.5.20: Replace target values using Replacement node

Figures 12.5.21 - 12.5.22: Transform Variables node in Modify tab



Figures 12.5.23 - 12.5.24: Log 10 transformation

Property	Value
General	
Node ID	Trans3
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Formulas	
Interactions	
SAS Code	
Default Methods	
Interval Inputs	None
Interval Targets	None
Class Inputs	None
Class Targets	None
Treat Missing as Level	No
Sample Properties	
Method	Random
Size	Max
Random Seed	12345
Optimal Binning	
Number of Bins	4
Missing Values	Use in Search
Grouping Method	
Cutoff Value	0.1
Group Missing	No
Number of Bins	Variables
Add Minimum Value to Offset Value	Yes
Offset Value	1
Score	
Use Meta Transformation	Yes
Hide	Yes
Reject	Yes
Report	
Summary Statistics	Yes
Status	
Create Time	5/29/23 11:34 AM
Run ID	94055658-5fa0-d844-b2c1-258f147d0c17
Last Error	
Last Status	Complete
Last Run Time	5/29/23 1:56 PM
Run Duration	0 Hr. 0 Min. 44.99 Sec.
Grid Host	
User-Added Node	No

Figures 12.5.25 - 12.5.26: Data Resampling

Property	Value
General	
Node ID	Smpl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Sample Method	Stratify
Random Seed	12345
Size	
Type	Number of Observations
Observations	20000
Percentage	100.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	5/29/23 1:17 PM
Run ID	30c26b36-9974-4747-b8f6-72a4ea9ea760
Last Error	
Last Status	Complete
Last Run Time	5/29/23 1:56 PM
Run Duration	0 Hr. 0 Min. 33.60 Sec.
Grid Host	
User-Added Node	No

Figure 12.5.27 Correlation Matrix Table

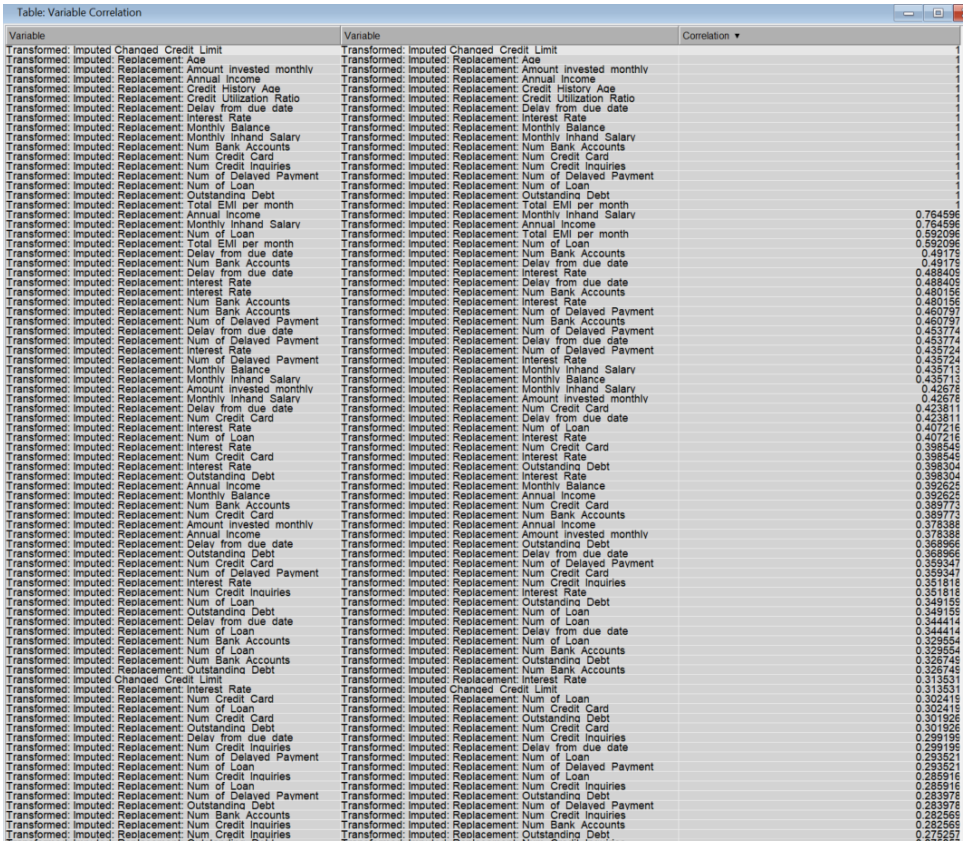


Table: Variable Correlation

Variable	Variable	Correlation ▾
Transformed: Imputed Changed Credit Limit	Transformed: Imputed Changed Credit Limit	1
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Age	1
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Amount invested monthly	1
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Annual Income	1
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Credit History Age	1
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Credit Utilization Ratio	1
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Delay from due date	1
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Interest Rate	1
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Monthly Balance	1
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Monthly Inhand Salary	1
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Num Bank Accounts	1
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Num Credit Card	1
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Num Credit Inquiries	1
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Num of Delayed Payment	1
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Num of Loan	1
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Outstanding Debt	1
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Total EMI per month	1
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Monthly Inhand Salary	0.764596
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Annual Income	0.764596
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Total EMI per month	0.592096
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Num of Loan	0.592096
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Num Bank Accounts	0.49175
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Delay from due date	0.49175
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Interest Rate	0.488409
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Delay from due date	0.488405
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Num Bank Accounts	0.480156
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Interest Rate	0.480156
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Num of Delayed Payment	0.460797
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Num Bank Accounts	0.460797
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Num of Delayed Payment	0.453774
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Delay from due date	0.453774
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Num of Delayed Payment	0.435724
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Interest Rate	0.435724
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Monthly Inhand Salary	0.435713
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Monthly Balance	0.435713
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Monthly Inhand Salary	0.42678
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Amount invested monthly	0.42678
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Num Credit Card	0.423811
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Delay from due date	0.423811
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Num of Loan	0.407216
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Interest Rate	0.407216
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Num Credit Card	0.398549
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Interest Rate	0.398549
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Outstanding Debt	0.398304
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Interest Rate	0.398304
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Monthly Balance	0.392625
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Annual Income	0.392625
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Num Credit Card	0.389773
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Num Bank Accounts	0.389773
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Annual Income	0.378388
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Amount invested monthly	0.378388
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Outstanding Debt	0.368966
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Delay from due date	0.368966
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Num of Delayed Payment	0.359347
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Num Credit Card	0.359347
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Num Credit Inquiries	0.351818
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Interest Rate	0.351818
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Outstanding Debt	0.349155
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Num of Loan	0.349155
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Num of Loan	0.344414
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Delay from due date	0.344414
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Num of Loan	0.329554
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Num Bank Accounts	0.329554
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Outstanding Debt	0.326748
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Num Bank Accounts	0.326748
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Interest Rate	0.313531
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed Changed Credit Limit	0.313531
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Num of Loan	0.302419
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Num Credit Card	0.302419
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Outstanding Debt	0.301926
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Num Credit Card	0.301926
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Num Credit Inquiries	0.299199
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Delay from due date	0.299199
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Num of Loan	0.293521
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Num of Delayed Payment	0.293521
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Num of Loan	0.285916
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Num Credit Inquiries	0.285916
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Outstanding Debt	0.283978
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Num of Delayed Payment	0.283978
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Num Credit Inquiries	0.282565
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Num Bank Accounts	0.282565
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Outstanding Debt	0.275257

Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Num Credit Inquiries	0.275257
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Monthly Balance	0.261265
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Credit History Age	0.261265
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Num of Delayed Payment	0.260366
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Num Credit Inquiries	0.260366
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Num Bank Accounts	0.255933
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed Changed Credit Limit	0.255933
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Delay from due date	0.252808
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed Changed Credit Limit	0.252808
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Num of Loan	0.248869
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed Changed Credit Limit	0.248869
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Num Credit Inquiries	0.247425
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Num Credit Card	0.247425
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Num of Delayed Payment	0.241721
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed Changed Credit Limit	0.241721
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Monthly Inhand Salary	0.241214
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Credit History Age	0.241214
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Outstanding Debt	0.226278
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed Changed Credit Limit	0.226278
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Num Credit Card	0.217587
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed Changed Credit Limit	0.217587
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Monthly Balance	0.216955
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Credit Utilization Ratio	0.216955
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Credit History Age	0.214624
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Annual Income	0.214624
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Num Credit Inquiries	0.200326
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed Changed Credit Limit	0.200326
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Total EMI per month	0.183137
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Monthly Inhand Salary	0.183137
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Total EMI per month	0.164973
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Annual Income	0.164973
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Monthly Inhand Salary	0.14938
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Credit Utilization Ratio	0.14938
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Credit Utilization Ratio	0.142653
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Annual Income	0.142653
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Total EMI per month	0.12956
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Interest Rate	0.12956
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Credit History Age	0.127804
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Age	0.127804
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Credit History Age	0.113719
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Amount invested monthly	0.113719
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Total EMI per month	0.107051
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Outstanding Debt	0.107051
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Total EMI per month	0.102792
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Delay from due date	0.102792
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Total EMI per month	0.101018
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Num Credit Inquiries	0.101018
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Total EMI per month	0.089754
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Amount invested monthly	0.089754
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Total EMI per month	0.089201
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed Changed Credit Limit	0.089201
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Total EMI per month	0.088097
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Num Bank Accounts	0.088097
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Total EMI per month	0.085632
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Num Credit Card	0.085632
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Total EMI per month	0.076839
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Num of Delayed Payment	0.076839
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Credit Utilization Ratio	0.060825
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Credit History Age	0.060825
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Monthly Balance	0.059336
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Age	0.059336
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Monthly Inhand Salary	0.054558
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Age	0.054558
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Annual Income	0.047927
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Age	0.047927
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Amount invested monthly	0.021161
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Age	0.021161
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Credit Utilization Ratio	0.012025
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Age	0.012025
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Credit Utilization Ratio	0.010889
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Amount invested monthly	0.010889
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Total EMI per month	-0.01132
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.01132
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Num Credit Inquiries	-0.04107
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.04107
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.04744
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed Changed Credit Limit	-0.04744
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Monthly Balance	-0.04826
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Amount invested monthly	-0.04826
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Num Credit Card	-0.04965
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.04965
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Outstanding Debt_	-0.05015
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.05015
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Num of Delayed Payment	-0.05191
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.05191
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Delay from due date	-0.05248
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.05248
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Total EMI per month	-0.05776
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Age	-0.05776
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Num Bank Accounts	-0.06605
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.06605
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Interest Rate	-0.06792
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.06792
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Num Credit Inquiries	-0.0712
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Amount invested monthly	-0.0712
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Age	-0.07953
Transformed: Imputed: Replacement: Age	Transformed: Imputed Changed Credit Limit	-0.07953
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Amount invested monthly	-0.07987
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed Changed Credit Limit	-0.07987
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Num of Delayed Payment	-0.08016
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Age	-0.08016
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Num Credit Card	-0.08793
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Age	-0.08793
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Num Credit Inquiries	-0.08931
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Age	-0.08931
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Total EMI per month	-0.08932
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Monthly Balance	-0.08932
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Outstanding Debt	-0.09225
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Amount invested monthly	-0.09225
Transformed: Imputed: Replacement: Credit Utilization Ratio	Transformed: Imputed: Replacement: Num of Loan	-0.09422
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Credit Utilization Ratio	-0.09422
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Num of Loan	-0.09604
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Amount invested monthly	-0.09604
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Delay from due date	-0.09746
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Age	-0.09746
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Outstanding Debt	-0.10064
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Age	-0.10064
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Num Credit Card	-0.10421
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Amount invested monthly	-0.10421
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Delay from due date	-0.10493
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Amount invested monthly	-0.10493
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Num Bank Accounts	-0.11094
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Age	-0.11094
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Num of Delayed Payment	-0.1151
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Amount invested monthly	-0.1151
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Interest Rate	-0.11953
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Age	-0.11953
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Num Bank Accounts	-0.12243
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Amount invested monthly	-0.12243
Transformed: Imputed: Replacement: Age	Transformed: Imputed: Replacement: Num of Loan	-0.12278
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Age	-0.12278
Transformed: Imputed: Replacement: Amount invested monthly	Transformed: Imputed: Replacement: Interest Rate	-0.12693
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Amount invested monthly	-0.12693
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Annual Income	-0.13464
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed Changed Credit Limit	-0.13464
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.15059
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed Changed Credit Limit	-0.15059
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Num Credit Inquiries	-0.15245
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Annual Income	-0.15245
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Monthly Balance	-0.15747
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed Changed Credit Limit	-0.15747
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Num Credit Inquiries	-0.15884
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.15884
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Total EMI per month	-0.16097
Transformed: Imputed: Replacement: Total EMI per month	Transformed: Imputed: Replacement: Credit History Age	-0.16097
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Num Credit Inquiries	-0.17063
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Monthly Balance	-0.17063
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Outstanding Debt	-0.17889
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Annual Income	-0.17889
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Num Credit Card	-0.18119
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Annual Income	-0.18119
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Num of Loan	-0.18563
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Annual Income	-0.18563
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Num Credit Card	-0.18814
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Monthly Balance	-0.18814
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Outstanding Debt	-0.19262
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.19262
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Num of Loan	-0.1982
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.1982
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Num Credit Card	-0.19942
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.19942
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Outstanding Debt	-0.20127
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Monthly Balance	-0.20127

Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Outstanding Debt	-0.20127
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Monthly Balance	-0.20127
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Delay from due date	-0.20312
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Annual Income	-0.20312
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Num of Delayed Payment	-0.20543
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Annual Income	-0.20543
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Num of Delayed Payment	-0.20799
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Monthly Balance	-0.20799
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Num Bank Accounts	-0.21547
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Annual Income	-0.21547
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Monthly Balance	-0.22018
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Delay from due date	-0.22018
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Num Bank Accounts	-0.22377
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Monthly Balance	-0.22377
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Num of Delayed Payment	-0.22438
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.22438
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.22787
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Delay from due date	-0.22787
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Num Bank Accounts	-0.23299
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.23299
Transformed: Imputed: Replacement: Annual Income	Transformed: Imputed: Replacement: Interest Rate	-0.23585
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Annual Income	-0.23585
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Monthly Balance	-0.25445
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Interest Rate	-0.25445
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Monthly Inhand Salary	-0.26053
Transformed: Imputed: Replacement: Monthly Inhand Salary	Transformed: Imputed: Replacement: Interest Rate	-0.26053
Transformed: Imputed Changed Credit Limit	Transformed: Imputed: Replacement: Credit History Age	-0.33372
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed Changed Credit Limit	-0.33372
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Num Credit Inquiries	-0.34456
Transformed: Imputed: Replacement: Num Credit Inquiries	Transformed: Imputed: Replacement: Credit History Age	-0.34456
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Num Credit Card	-0.35372
Transformed: Imputed: Replacement: Num Credit Card	Transformed: Imputed: Replacement: Credit History Age	-0.35372
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Num of Delayed Payment	-0.36224
Transformed: Imputed: Replacement: Num of Delayed Payment	Transformed: Imputed: Replacement: Credit History Age	-0.36224
Transformed: Imputed: Replacement: Monthly Balance	Transformed: Imputed: Replacement: Num of Loan	-0.36962
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Monthly Balance	-0.36962
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Num Bank Accounts	-0.4086
Transformed: Imputed: Replacement: Num Bank Accounts	Transformed: Imputed: Replacement: Credit History Age	-0.4086
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Delay from due date	-0.42548
Transformed: Imputed: Replacement: Delay from due date	Transformed: Imputed: Replacement: Credit History Age	-0.42548
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Outstanding Debt	-0.43916
Transformed: Imputed: Replacement: Outstanding Debt	Transformed: Imputed: Replacement: Credit History Age	-0.43916
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Num of Loan	-0.45794
Transformed: Imputed: Replacement: Num of Loan	Transformed: Imputed: Replacement: Credit History Age	-0.45794
Transformed: Imputed: Replacement: Credit History Age	Transformed: Imputed: Replacement: Interest Rate	-0.48389
Transformed: Imputed: Replacement: Interest Rate	Transformed: Imputed: Replacement: Credit History Age	-0.48389

12.6 Modify: Data Partitioning

Figures 12.6.1 - 12.6.2: Data Partition node in Sample tab

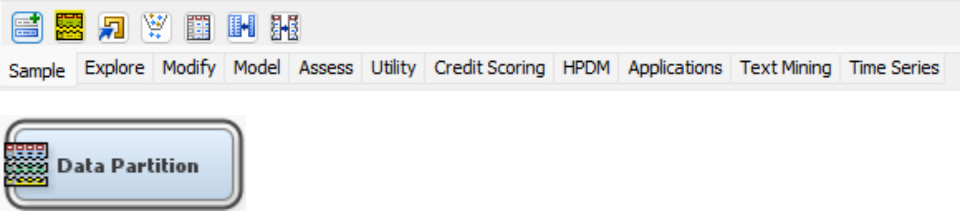
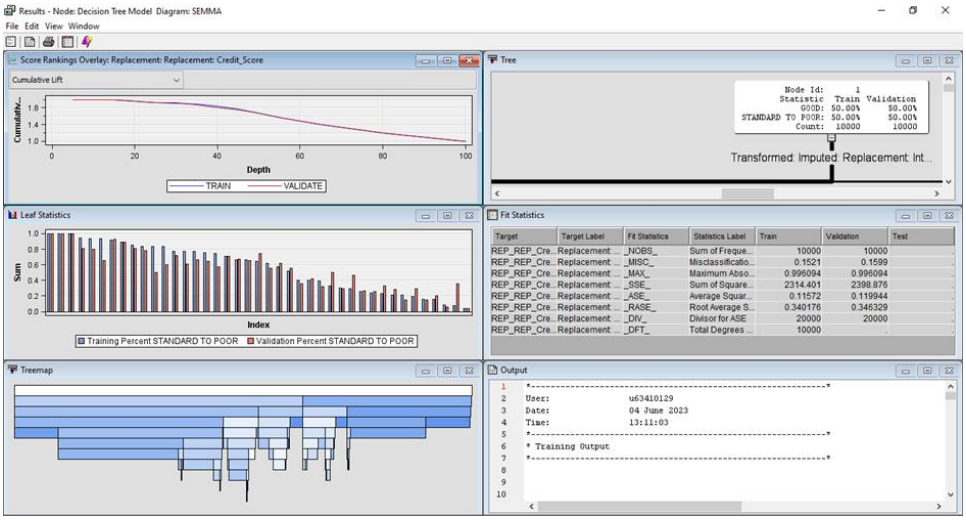
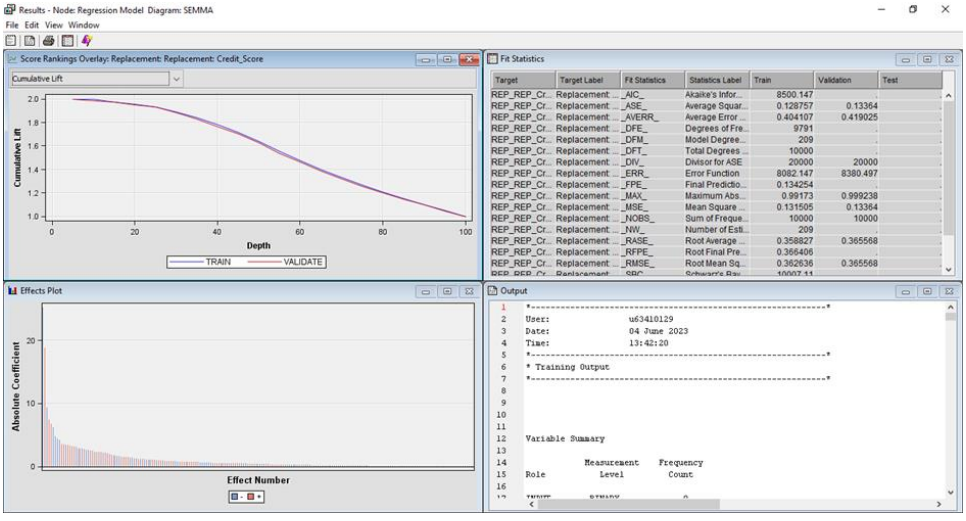


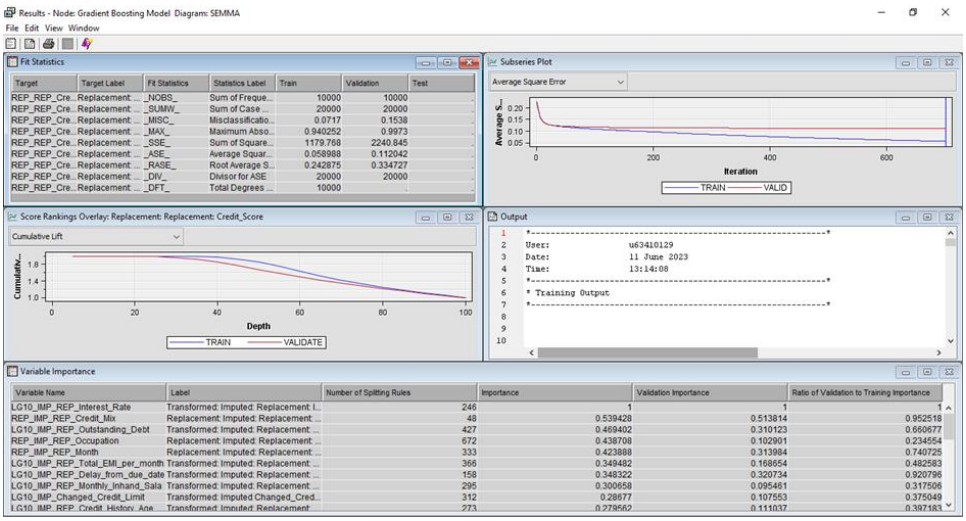
Figure 12.6.3: Partition Data

12.7 Model

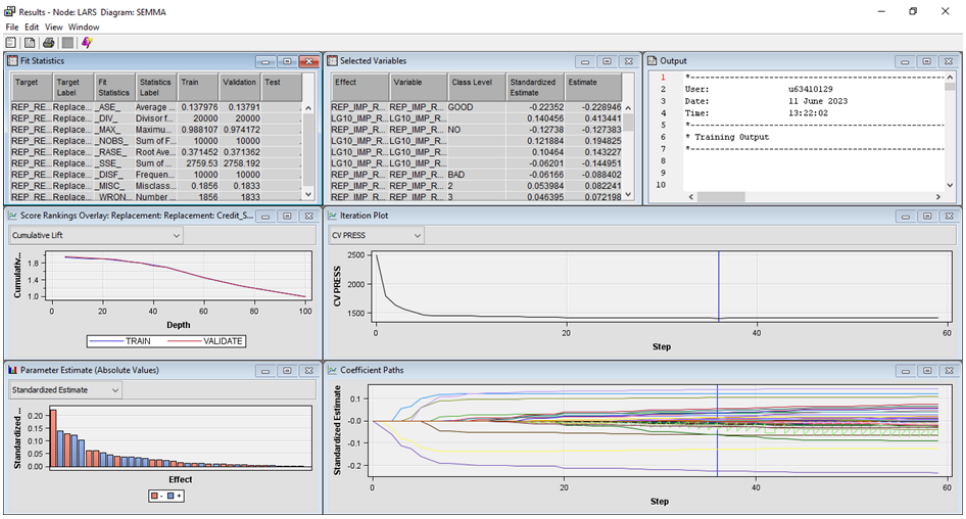
Figures 12.7.1 – 12.7.106Chosen model nodes in Model tab

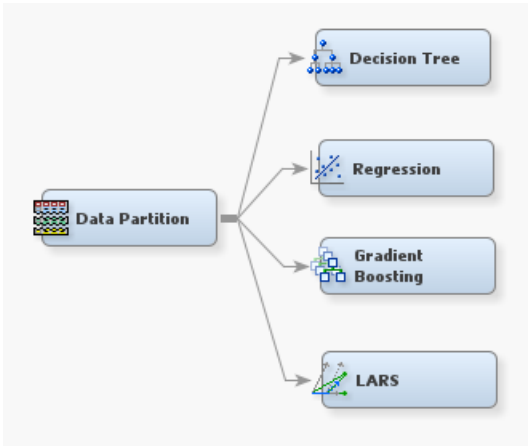





Results - Node: Gradient Boosting Model Diagram: SEMMA
File Edit View Window
Fit Statistics
Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test
REP_REP_Cre... | Replacement: ... | _NOBS_ | Sum of Freque... | 10000 | 10000 |
REP_REP_Cre... | Replacement: ... | _SUMW_ | Sum of Case ... | 20000 | 20000 |
REP_REP_Cre... | Replacement: ... | _MISC_ | Misclassificatio... | 0.0717 | 0.1538 |
REP_REP_Cre... | Replacement: ... | _MAX_ | Maximum Abso... | 0.940252 | 0.9973 |
REP_REP_Cre... | Replacement: ... | _SSE_ | Sum of Square... | 1179.768 | 2240.845 |
REP_REP_Cre... | Replacement: ... | _ASE_ | Average Squar... | 0.058988 | 0.112042 |
REP_REP_Cre... | Replacement: ... | _RASE_ | Root Average S... | 0.242875 | 0.334727 |
REP_REP_Cre... | Replacement: ... | _DIV_ | Divisor for ASE | 20000 | 20000 |
REP_REP_Cre... | Replacement: ... | _DFT_ | Total Degrees ... | 10000 |
Subseries Plot
Average Square Error
Average S...
Iteration
TRAIN VALID
Score Rankings Overlay: Replacement: Replacement: Credit_Score
Cumulative Lift
Cumulativ...
Depth
TRAIN VALIDATE
Output
User: u63410129
Date: 11 June 2023
Time: 13:14:08
* Training Output
Variable Importance
Variable Name | Label | Number of Splitting Rules | Importance | Validation Importance | Ratio of Validation to Training Importance
LG10_IMP_REP_Interest_Rate | Transformed: Imputed: Replacement: I... | 246 | 1 | 1 | 1
REP_IMP_REP_Credit_Mix | Replacement: Imputed: Replacement: ... | 48 | 0.539428 | 0.513814 | 0.952518
LG10_IMP_REP_Outstanding_Debt | Transformed: Imputed: Replacement: ... | 427 | 0.469402 | 0.310123 | 0.660677
REP_IMP_REP_Occupation | Replacement: Imputed: Replacement: ... | 672 | 0.438708 | 0.102901 | 0.234554
REP_IMP_REP_Month | Replacement: Imputed: Replacement: ... | 333 | 0.423888 | 0.313984 | 0.740725
LG10_IMP_REP_Total_EMI_per_month | Transformed: Imputed: Replacement: ... | 366 | 0.349482 | 0.168654 | 0.482583
LG10_IMP_REP_Delay_from_due_date | Transformed: Imputed: Replacement: ... | 158 | 0.348322 | 0.320734 | 0.920796
LG10_IMP_REP_Monthly_Inhand_Sala | Transformed: Imputed: Replacement: ... | 295 | 0.300658 | 0.095461 | 0.317506
LG10_IMP_Changed_Credit_Limit | Transformed: Imputed Changed_Cred... | 312 | 0.28677 | 0.107553 | 0.375049
LG10_IMP_REP_Credit_History_Age | Transformed: Imputed: Replacement: | 273 | 0.279562 | 0.111037 | 0.397183

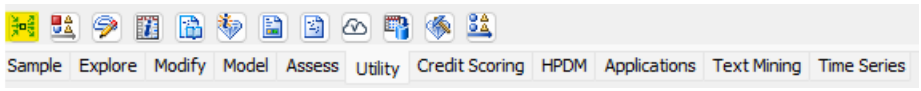

LARS

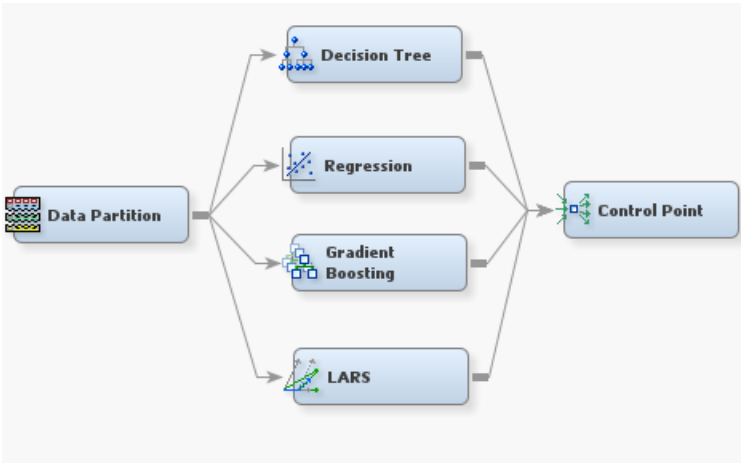

Results - Node: LARS Diagram: SEMMA
File Edit View Window
Fit Statistics
Target | Target Label | Fit Statistics | Statistics Label | Train | Validation | Test
REP_RE... | Replace... | _ASE_ | Average ... | 0.137976 | 0.13791 |
REP_RE... | Replace... | _DIV_ | Divisor f... | 20000 | 20000 |
REP_RE... | Replace... | _MAX_ | Maximu... | 0.988107 | 0.974172 |
REP_RE... | Replace... | _NOBS_ | Sum of F... | 10000 | 10000 |
REP_RE... | Replace... | _RASE_ | Root Ave... | 0.371452 | 0.371382 |
REP_RE... | Replace... | _SSE_ | Sum of ... | 2759.53 | 2758.192 |
REP_RE... | Replace... | _DISF_ | Frequen... | 10000 | 10000 |
REP_RE... | Replace... | _MISC_ | Misclass... | 0.1856 | 0.1833 |
REP_RE... | Replace... | WRON... | Number ... | 1856 | 1833 |
Selected Variables
Effect | Variable | Class Level | Standardized Estimate | Estimate
REP_IMP_R... | REP_IMP_R... | GOOD | -0.22352 | -0.228946
LG10_IMP_R... | LG10_IMP_R... | | 0.140456 | 0.413441
REP_IMP_R... | REP_IMP_R... | NO | -0.12738 | -0.127383
LG10_IMP_R... | LG10_IMP_R... | | 0.121884 | 0.194825
LG10_IMP_R... | LG10_IMP_R... | | 0.10464 | 0.143227
LG10_IMP_R... | LG10_IMP_R... | | -0.06201 | -0.144951
REP_IMP_R... | REP_IMP_R... | BAD | -0.06166 | -0.088402
REP_IMP_R... | REP_IMP_R... | 2 | 0.053984 | 0.082241
REP_IMP_R... | REP_IMP_R... | 3 | 0.046395 | 0.072198
Output
User: u63410129
Date: 11 June 2023
Time: 13:22:02
* Training Output
Score Rankings Overlay: Replacement: Replacement: Credit_S...
Cumulative Lift
Depth
TRAIN VALIDATE
Iteration Plot
CV PRESS
Step
Parameter Estimate (Absolute Values)
Standardized Estimate
Effect
Coefficient Paths
Standardized Estimate
Step


Decision Tree
Regression
Data Partition
Gradient Boosting
LARS

12.8 Assessment

Figures 12.8.1 – 12.8.3: Control Point node in Utility tab





Figures 12.8.4 – 12.8.6: Model Comparison node in Assess tab