



INDIVIDUAL ASSIGNMENT
TECHNOLOGY PARK MALAYSIA

CT045-3-M-ABAV

ADVANCED BUSINESS ANALYTICS VISUALIZATION

APDMF2310DSBA(DE)

STUDENT's TP: TP077702

STUDENT's NAME: MR. MOHAMMAD ADNAN

LECTURER's NAME: DR. PREETHI SUBRAMANIAN

Table of Contents

1	Introduction.....	3
1.1	Problem Statement.....	3
1.2	Objective.....	3
1.3	Scope.....	3
1.4	Methodology.....	4
2	SAS Enterprise Miner Solution	5
2.1	Diagram Flow	5
2.2	Metadata.....	6
2.3	Exploratory Data Analysis	8
2.3.1	Summary Statistics.....	8
2.3.2	Univariate Analysis on Categorical Variables.....	9
2.3.3	Univariate Analysis on Continuous Variable	10
2.3.4	Bivariate Analysis on Variables	12
2.4	Data Preprocessing.....	14
2.4.1	Replacement of Variable	14
2.4.2	Imputing Missing Values	14
2.4.3	Data Transformation	15
2.4.4	Data Sampling.....	16
2.4.5	Data Partitioning	16
2.5	Predictive Modeling.....	17
2.5.1	Tree-Based Model.....	17
2.5.2	Regression Model	20
3	Model Interpretation to Understand Business.....	26
3.1	Tree-based Model Interpretation.....	26
3.2	Regression Model Interpretation.....	29
3.3	Recommendations.....	31
4	Conclusion	32
5	References.....	33

1 Introduction

In the dynamic domain of financial lending, accurately predicting credit risk is paramount for the sustainability and profitability of financial organization. The rapid evolution of market conditions necessitates sophisticated tools that not only process historical data but also adapt to ongoing changes. Traditional methods, though reliable, fall short in addressing the complex, nonlinear relationships that influence borrower behaviour and risk. This project aims to leverage advanced analytical techniques to improve the prediction accuracy of credit risk, thereby enabling better decision-making and risk management.

1.1 Problem Statement

To minimize losses from borrower defaults, financial organization must be able to forecast loan credit risk. Despite the fact that there are many years' worth of historical data at their disposal, determining the odds of default with accuracy still presents a huge challenge. The situation is made worse by the fact that financial markets are always changing, and different economic environments affect the ability of borrowers to pay back. Manually analysing credit risks in the traditional way is both time consuming and labour intensive; additionally, it can lead to errors being made. These methods also tend not take into consideration complex nonlinear relationships among various factors influencing credit worthiness which further complicates matters. As a result, there is an urgent need for more sophisticated predictive models which can handle large datasets efficiently while addressing class imbalances so as to give precise predictions about credit risks; this will enable banks make well thought out lending decisions and manage their loan portfolios appropriately (Kanaparthi, 2023).

1.2 Objective

With respect to credit risk, the objectives of the study can be framed as follows:

1. To develop an appropriate model for predicting credit risk or loan default.
2. To analyse the key factors that influence credit risk or loan default.
3. To provide recommendations that can help improve credit risk management and loan portfolio performance.

1.3 Scope

The scope of the project is to create a predictive model that may be used to estimate credit risk or the chance of loan default. The actions that need to be carried out are data exploration

and transformation, modelling procedures using automated machine learning algorithms, and suitable statistical methods. The accessible information about borrower details, loan features, and relevant factors contained in the dataset will all be used in the study. The study's findings will provide insight into credit risk assessment and management and offer recommendations. The use of the model or its recommendations in a real-world business context is not covered by the scope, yet. Access to the dataset needed for the investigation can be made from the link: <https://www.kaggle.com/datasets/laotse/credit-risk-dataset> .

1.4 Methodology

This study employs the SEMMA process (Sample, Explore, Modify, Model, Assess) approach. By systematically preparing data, identifying patterns, creating models, and assessing their effectiveness, it aids in the extraction of hidden insights. This systematic technique guarantees that data scientists locate important information for improved decision-making and don't overlook any important processes (Omari, 2023).

1. **Sample:** Select a subset of the data that includes various customer demographics, financial behaviours, and historical loan performance to ensure that the sample adequately represents the population.
2. **Explore:** Use statistical methods and visualization tools to explore the dataset. Analyse variables such as income, credit history, and payment behaviour to identify trends and detect any outliers or missing values.
3. **Modify:** Handle missing values, encode categorical variables if required, normalize numerical features, and create new features if necessary. Techniques like sampling can be used to address class imbalances.
4. **Model:** Train multiple machine learning models such as logistic regression, decision trees, and ensemble methods (e.g., Random Forest) on the pre-processed dataset.
5. **Assess:** Validate the models using a separate test dataset to ensure they generalize well to unseen data. Use metrics such as misclassification rate and ROC-AUC to assess the model performance.

2 SAS Enterprise Miner Solution

2.1 Diagram Flow

The provided SAS Enterprise Miner flow diagram outlines a comprehensive SEMMA-based approach to credit risk prediction. The process begins with the "File Import" node, which brings in the dataset for analysis. Initial data exploration is performed using the "StatExplore" and "Graph Explore" nodes to understand categorical and continuous variables, respectively. Data cleaning involves "Impute" and "Replacement" nodes to handle missing values and outliers. "Transform Variables" is then used to modify variables for better model performance. The dataset is then sampled using the "Sample" node, followed by further exploration with "StatExplore #2" and "Graph Explore" nodes to refine understanding. Metadata is defined to ensure proper data handling. The dataset is partitioned into training and validation sets using the "Data Partition" node.

Modeling begins with decision trees, including "Decision Tree," "HP Tree C-C," "HP Tree C4.5," and "HP Forest" nodes. These models are compared using the "Model Comparison #1" node. Logistic regression models are also developed using nodes like "Log Regression Forward," "Log Regression Backward," "Clog Regression Forward," "HP Regression Poly," and "HP Regression Stepwise." These models are compared using the "Model Comparison #2" node to identify the best performing model for predicting credit risk.

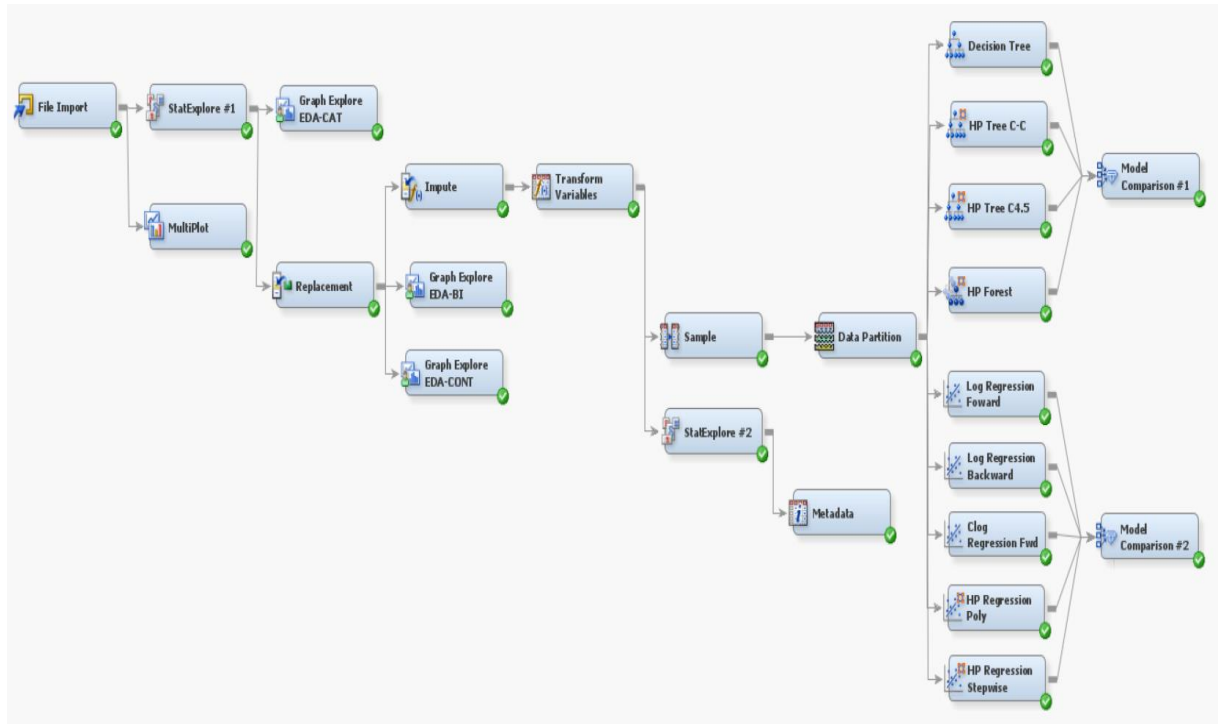


Figure 1 SAS Enterprise Miner Diagram Flow

2.2 Metadata

There are 12 columns and 32582 rows in this dataset, with 7 continuous variables and the remaining 5 categorical variables. The target variable is loan_status which is binary variable.

Variable Name	Description	Data Type
Person_age	Age of the applicant	Numeric
Person_income	Annual income of the applicant	Numeric
Person_home_ownership	Home ownership status of the applicant	Categorical
Person_emp_length	Employment length in years	Numeric
Loan_intent	Purpose of the loan	Categorical
Loan_grade	Grade assigned to the loan	Categorical
Loan_amt	Amount of the loan	Numeric
Loan_int_rate	Interest rate of the loan	Numeric
loan_status	Loan status (0 = non-default, 1 = default)	Binary
loan_percent_income	Loan amount as a percentage of income	Numeric
Cb_person_default_on_file	Historical default status of the applicant	Categorical
Cb_person_cred_hist_length	Credit history length in years	Numeric

The process flow starts with the dataset being imported into a file. The variable description in StatExplore provides metadata that displays the role and interval. As can be seen below, there are 5 category variables and 7 continuous variables in the dataset.

Name	Use	Report	Role	Level
cb_person_cred	Default	No	Input	Interval
cb_person_defa	Default	No	Input	Binary
loan_amnt	Default	No	Input	Interval
loan_grade	Default	No	Input	Nominal
loan_int_rate	Default	No	Input	Interval
loan_intent	Default	No	Input	Nominal
loan_percent_in	Default	No	Input	Interval
loan_status	Default	No	Target	Binary
person_age	Default	No	Input	Interval
person_emp_len	Default	No	Input	Interval
person_home_o	Default	No	Input	Nominal
person_income	Default	No	Input	Interval

Figure 2 Metadata

2.3 Exploratory Data Analysis

EDA is done to understand a dataset's structure, uncover patterns, and find anomalies. This helps prepare data for further analysis and make informed decisions in credit risk.

2.3.1 Summary Statistics

To obtain a quick overview of the dataset and determine whether any data are missing, the data scientist does a summary statistic.

39									
40	Data			Number					
41	Role	Variable Name	Role	Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
42									
43	TRAIN	cb_person_default_on_file	INPUT	2	0	N	82.37	Y	17.63
44	TRAIN	loan_grade	INPUT	7	0	A	33.08	B	32.08
45	TRAIN	loan_intent	INPUT	6	0	EDUCATION	19.81	MEDICAL	18.63
46	TRAIN	person_home_ownership	INPUT	4	0	RENT	50.48	MORTGAGE	41.26
47	TRAIN	loan_status	TARGET	2	0	0	78.18	1	21.82

Figure 3 StatExplore for Categorical Value

We can see that the result above shows the various levels as well as the absence of any missing values from the dataset.

69											
70	Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
71											
72	cb_person_cred_hist_length	INPUT	5.804211	4.055001	32581	0	2	4	30	1.66179	3.716194
73	loan_amt	INPUT	9589.371	6322.087	32581	0	500	8000	35000	1.192477	1.423565
74	loan_int_rate	INPUT	11.01169	3.240459	29465	3116	5.42	10.99	23.22	0.20855	-0.67161
75	loan_percent_income	INPUT	0.170203	0.106782	32581	0	0	0.15	0.83	1.064669	1.223687
76	person_age	INPUT	27.7346	6.348078	32581	0	20	26	144	2.581393	18.56082
77	person_emp_length	INPUT	4.789686	4.14263	31686	895	0	4	123	2.614455	43.72234
78	person income	INPUT	66074.85	61983.12	32581	0	4000	55000	6000000	32.86535	2693.273

Figure 4 StatExplore for Continuous Value

The result above provides detailed descriptive statistics for continuous variables involved in credit risk. The mean values show average trends, while the standard deviations indicate variability. The skewness and kurtosis values suggest data distribution shapes. Missing values are present for the interest rate and person employment length, which has 3116 and 895 missing entries which will be imputed later and the maximum value for Person_age and Person_emp_length is 140 and 123 which needs to be replaced.

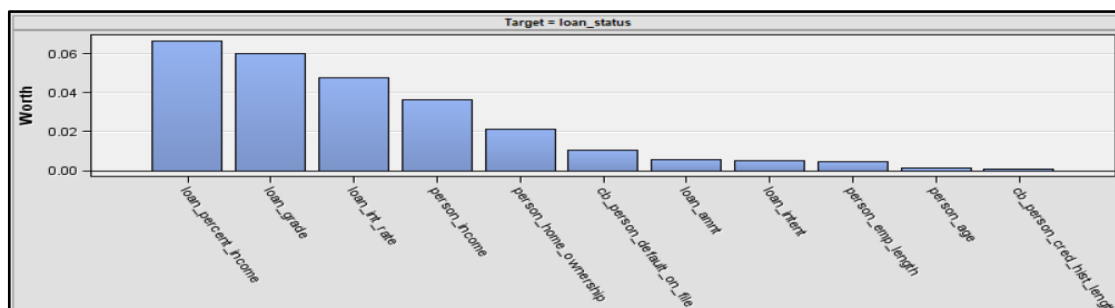
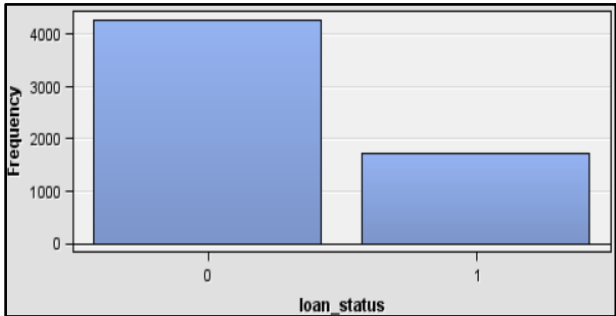
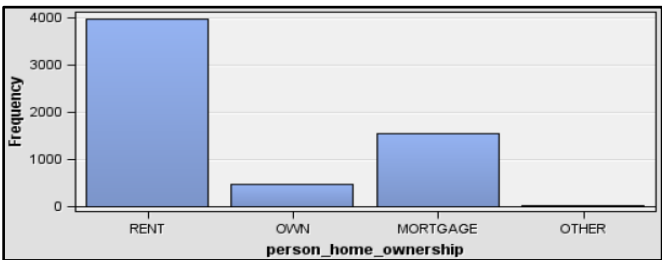
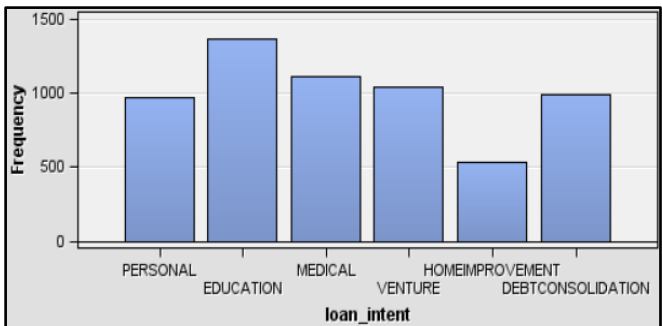
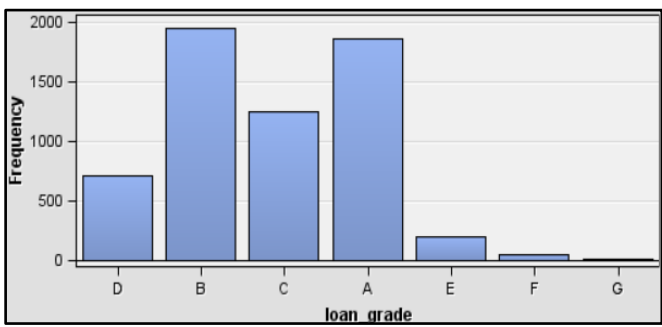
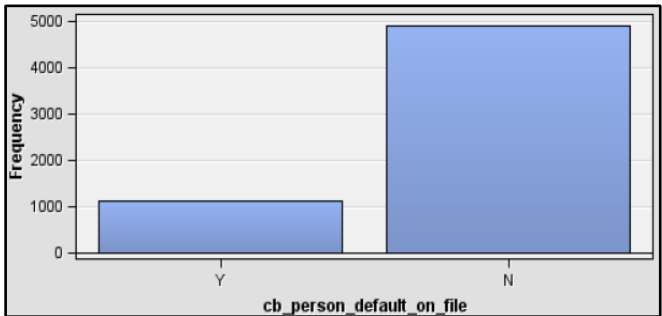


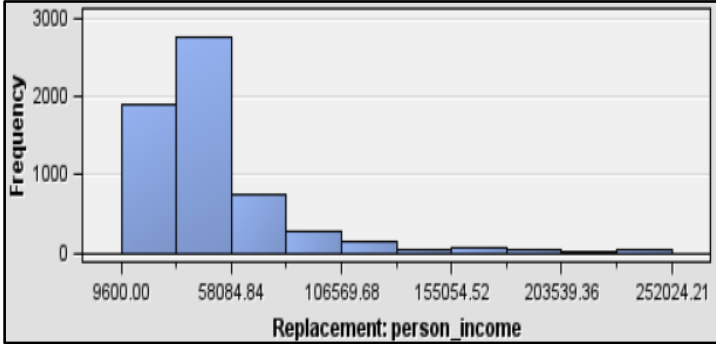
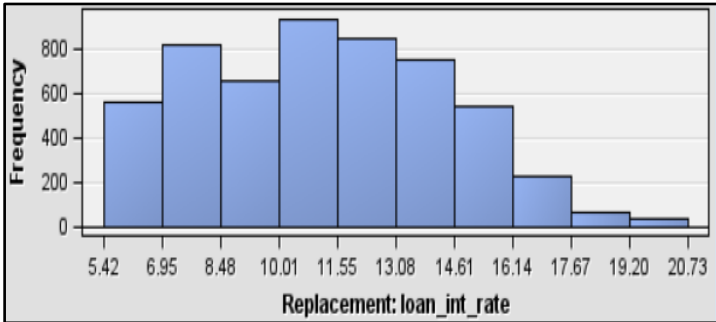
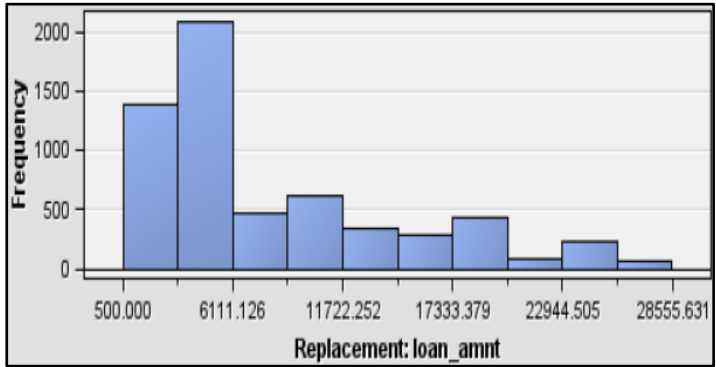
Figure 5 StatExplore Variable Worth

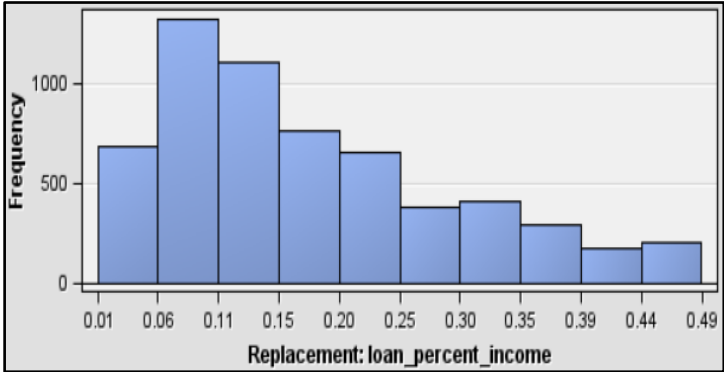
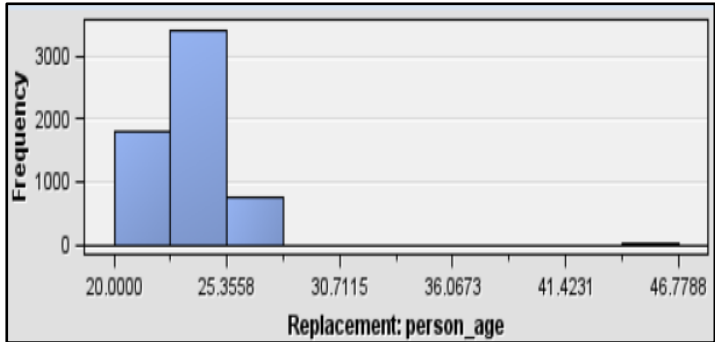
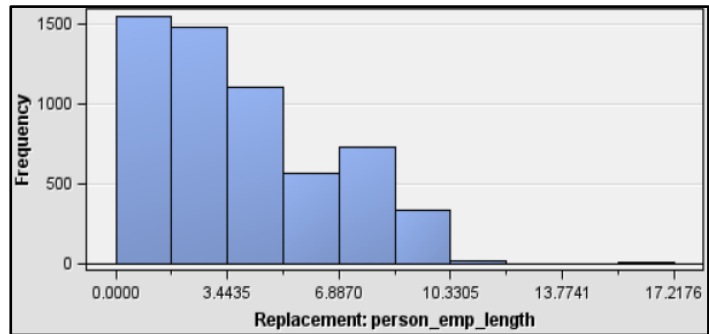
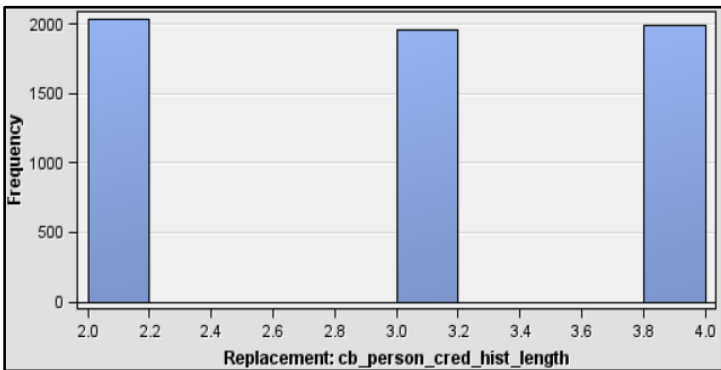
The variable worth is displayed in the above result, and it is clear that loan percent income has the most influence on the target variable, followed by rest, person age, and person credit history length, which have the least influence on the target.

2.3.2 Univariate Analysis on Categorical Variables

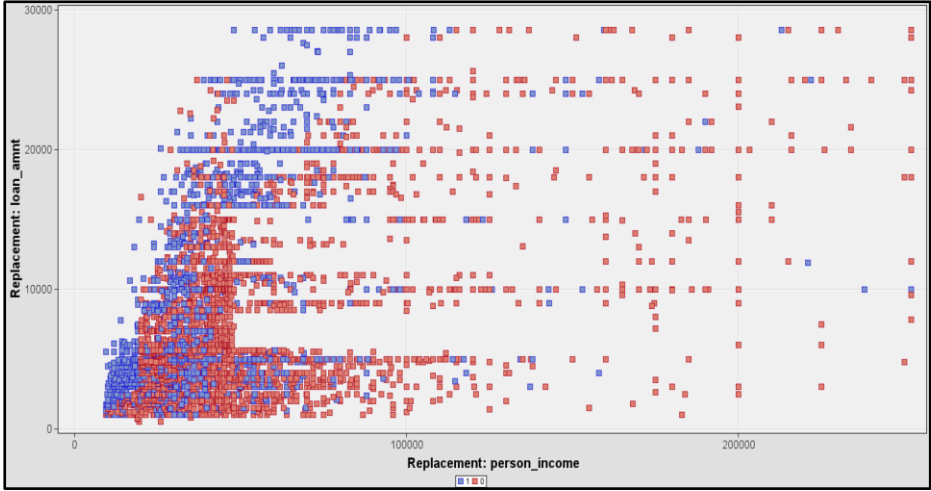
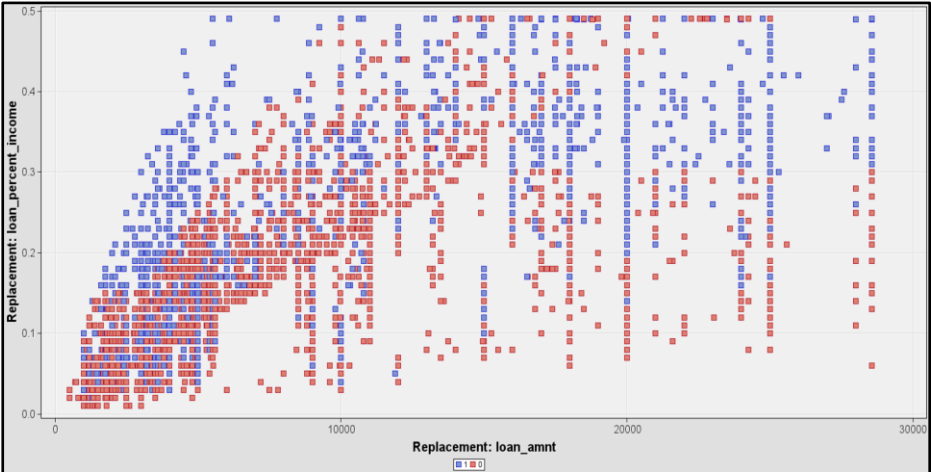
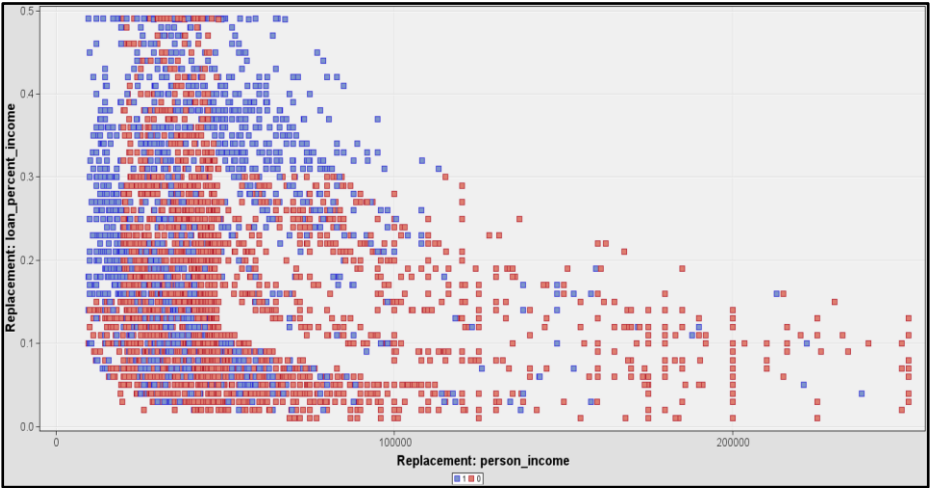
Variable Name	Exploration Result	Interpretation	Preprocessing
Loan_status (Target)		<p>Loan_status is chosen as the target variable for predicting credit risk.</p> <p>Bar chart shows loan status frequencies: '0' (no default) significantly higher than '1' (default), indicating fewer loan defaults.</p>	Sampling is required in order to provide an evenly distributed target variable for training the model.
Person_home_ownership		Bar chart shows home ownership types: most people rent, followed by mortgage holders, fewer own outright, and very few are categorized as 'other'.	Preprocessing is not required
Loan_intent		Bar chart indicates loan purposes: Education and Medical are most common, followed by Personal and Debt Consolidation. Venture loans are also frequent, while Home Improvement loans are notably less common.	Preprocessing is not required
Loan_grade		Bar chart shows loan grades distribution: Grades B, C, and A are most common, indicating higher creditworthiness. Grades D, E, F, and G are less frequent, suggesting fewer loans given to lower credit scores.	Preprocessing is not required
Cb_person_default_on_file		Bar chart shows a smaller percentage of the population in the dataset has a prior default (Y) on file than those who have never defaulted (N), suggesting that creditworthy people make up most of the dataset.	Preprocessing is not required

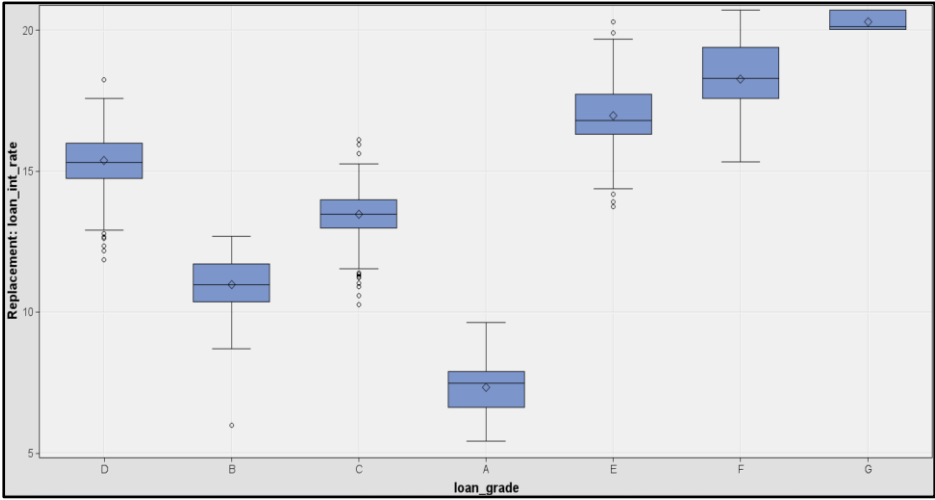
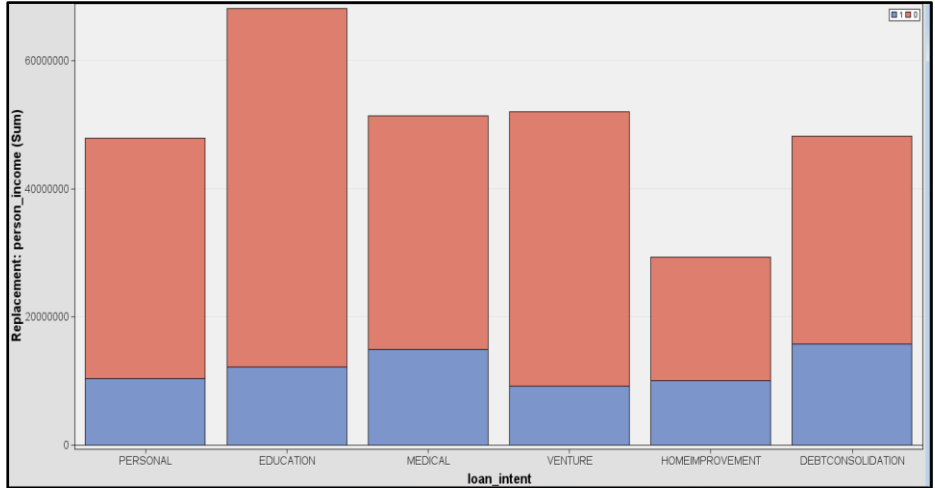
2.3.3 Univariate Analysis on Continuous Variable

Variable Name	Exploration Result	Interpretation	Preprocessing
Person_income		A right-skewed pattern can be seen in the persons income distribution histogram, where most persons earn between \$9,600 and \$106,569.68, while a smaller percentage earn higher earnings, indicating income inequality in the population under study. In this dataset, high-income people (over \$203,539.36) are noticeably uncommon.	We need to do preprocessing and will utilise log transformation because of the right-skew pattern.
Loan_int_rate		The histogram shows the distribution of loan interest rates. Most loans have interest rates between 6.95% and 13.08%, with the frequency gradually decreasing as rates increase. Few loans have very high rates, indicating typical rates are moderately priced, and very high rates are less common.	Preprocessing is not required
Loan_amt		The histogram displays loan amount distribution, showing a right-skewed pattern. Most loans are between \$500 and \$11,722.25, with frequency decreasing as loan amounts increase. Very few loans reach the upper range near \$28,555.63, indicating that higher loan amounts are less common in this dataset.	We need to perform preprocessing and will utilise log transformation because of the right-skew pattern.

Loan_percent_income		This histogram shows the distribution of loans as a percentage of borrowers' income. It indicates most loans constitute between 1% and 15% of income, with frequency decreasing as this percentage increases. Fewer borrowers take loans that make up a larger share of their income, showing caution or income limitations.	We need to perform preprocessing and will utilise log transformation because of the slight right-skew pattern.
Person_age		The histogram of person age shows a right-skewed distribution, with most borrowers aged between 20 and 30 years. The frequency of loans decreases significantly for ages beyond 25, indicating a younger borrower demographic. There is an outlier age.	We need to perform preprocessing and will utilise log transformation because of the slight right-skew pattern.
Person_emp_length		Histogram shows employment length among borrowers, showing a right-skewed distribution. Most borrowers have between 0 and 7 years of employment with significantly fewer individuals having longer employment histories up to 17 years.	We need to perform preprocessing and will utilise log transformation because of the right-skew pattern.
Cb_person_cred_hist_len		It represents credit history length, showing two groups: borrowers with approximately 2 years and 4 years of credit history, with similar frequencies. It suggests these are common durations for credit history among borrowers.	Preprocessing is not required

2.3.4 Bivariate Analysis on Variables

Variables	Exploration Result	Interpretation
Loan_amt vs Person_income		<ul style="list-style-type: none"> The scatter plot correlates loan amounts with person income, highlighting loan defaults (red) versus non-defaults (blue). Defaults are more frequent at lower income levels, suggesting a higher risk associated with these borrowers. Higher incomes show fewer defaults, indicating better loan repayment capabilities among these individuals.
Loan_amt Vs Loan_per_income		<ul style="list-style-type: none"> The scatter plot compares loan amount to loan percent of income, marked by loan status. Higher default rates are concentrated at higher loan percent of income, regardless of loan amount. Non-defaults are spread across various loan percentages and amounts, indicating diversified borrowing behaviour without default.
Person_income Vs Loan_per_income		<ul style="list-style-type: none"> The plot compares person income to loan percent of income, differentiated by loan status. Defaults decrease as income increases, suggesting higher incomes manage loan commitments more effectively. Non-defaults are generally lower in loan percent of income across various incomes, showing careful borrowing behaviour.

<p>Loan_int_rate Vs Loan_grade</p>		<ul style="list-style-type: none"> • The box plot displays loan interest rates by grade (A to G). • Higher grades (A) have lower median interest rates, indicating lower risk. • Lower grades (D to G) show progressively higher median rates, reflecting increased risk. • Grades D and G exhibit more variability and outliers, suggesting less consistency in their rates.
<p>Person_income Vs Loan_intent</p>		<ul style="list-style-type: none"> • The chart compares loan intents by total income, highlighting defaults (red) and non-defaults (blue). • Education and Medical loans show the highest total incomes but also significant defaults. • Venture and Personal loans have moderate total incomes with notable default levels. • Home Improvement and Debt Consolidation show smaller total incomes and lower defaults, indicating potentially lower risks.

A complete examination of data identifies main features of loan activities. These comprise the fact that most people default when they earn less income; additionally, individuals borrowing higher amounts relative to their earnings have an even higher likelihood of defaulting. Conversely, it is worth noting that even though there were very high total incomes reflected for education or medical loans, a lot of them still ended up being bad debts. Transformations such as logarithmic adjustments are recommended for right-skewed distributions of income, loan amount, and person age to improve modelling accuracy. Overall, understanding these patterns helps in predicting risk and tailoring financial products to meet borrower needs effectively. This analysis underscores the importance of targeted financial strategies and robust risk assessment tools in managing credit effectively.

2.4 Data Preprocessing

2.4.1 Replacement of Variable

In the following, we specify the range of lower and upper limit and substitute the values for age and employment length with normal life expectancy of a person and an average employment length.

Name	Use	Limit Method	Replacement Lower Limit	Replacement Upper Limit	Replace Method
cb_person_cred	Default	Default	.	.	Default
loan_amnt	Default	Default	.	.	Default
loan_int_rate	Default	Default	.	.	Default
loan_percent_in	Default	Default	.	.	Default
person_age	Default	Default	0	75	Computed
person_emp_len	Default	Default	0	40	Computed
person_income	Default	Default	.	.	Default

Figure 6 Replacement of Variable Editor

Variable	Replace Variable	Lower limit	Lower Replacement Value	Upper Limit	Upper Replacement Value
cb_person_cred_hist_length	REP_cb_person_cred_hist_length	-6.36	-6.36	17.97	17.97
loan_amnt	REP_loan_amnt	-9376.89	-9376.89	28555.63	28555.63
loan_int_rate	REP_loan_int_rate	1.29	1.29	20.73	20.73
loan_percent income	REP_loan_percent income	-0.15	-0.15	0.49	0.49
person_age	REP_person_age	8.69	8.69	46.78	46.78
person_emp_length	REP_person_emp_length	-7.64	-7.64	17.22	17.22
person_income	REP_person_income	-119874.51	-119874.51	252024.21	252024.21

Figure 7 Output of Replacement

In the above result we can observe that the replacement has been successfully implemented on Person_age and Person_emp_length.

2.4.2 Imputing Missing Values

The variables to be imputed and the imputation method are chosen in the following; in this example, the values of Loan_int_rate and Person_emp_length are imputed using the mean.

Name	Use	Method	Use Tree	Role	Level
REP_cb_person_cred_hist_length	Default	Default	Default	Input	Interval
REP_loan_amnt	Default	Default	Default	Input	Interval
REP_loan_int_rate	Yes	Mean	Default	Input	Interval
REP_loan_percent_income	Default	Default	Default	Input	Interval
REP_person_age	Default	Default	Default	Input	Interval
REP_person_emp_length	Yes	Mean	Default	Input	Interval
REP_person_income	Default	Default	Default	Input	Interval
cb_person_cred_hist_length	Default	Default	Default	Rejected	Interval
cb_person_default_on_file	Default	Default	Default	Input	Binary
loan_amnt	Default	Default	Default	Rejected	Interval
loan_grade	Default	Default	Default	Input	Nominal
loan_int_rate	Default	Default	Default	Rejected	Interval
loan_intent	Default	Default	Default	Input	Nominal
loan_percent_income	Default	Default	Default	Rejected	Interval
loan_status	Default	Default	Default	Target	Binary
person_age	Default	Default	Default	Rejected	Interval
person_emp_length	Default	Default	Default	Rejected	Interval
person_home_ownership	Default	Default	Default	Input	Nominal
person_income	Default	Default	Default	Rejected	Interval

Figure 8 Impute Editor

35	Imputation Summary						
36	Number Of Observations						
37							
38							Number of
39		Impute		Impute	Measurement		Missing
40	Variable Name	Method	Imputed Variable	Value	Role	Level	for TRAIN
41							
42	REP_loan_int_rate	MEAN	IMP_REP_loan_int_rate	11.0109	INPUT	INTERVAL	Replacement: loan_int_rate 3116
43	REP_person_emp_length	MEAN	IMP_REP_person_emp_length	4.7507	INPUT	INTERVAL	Replacement: person_emp_length 895

Figure 9 Impute output

It is evident from the above that the mean value has been successfully imputed for Loan_int_rate and Person_emp_length.

2.4.3 Data Transformation

The method is set to log to the following variables: Person_age, Person_emp_length, Loan_amt, and Person_income. This applies the log transformation and makes the skewed distribution normally distributed.

Name	Method	Number of Bins	Role	Level
IMP_REP_loan_int_rate	Default	4	Input	Interval
IMP_REP_person_emp_length	Log	4	Input	Interval
REP_cb_person	Default	4	Input	Interval
REP_loan_amnt	Log	4	Input	Interval
REP_loan_percent	Default	4	Input	Interval
REP_person_age	Log	4	Input	Interval
REP_person_income	Log	4	Input	Interval
cb_person_cred	Default	4	Rejected	Interval
cb_person_defaul	Default	4	Input	Binary
loan_amnt	Default	4	Rejected	Interval
loan_grade	Default	4	Input	Nominal
loan_int_rate	Default	4	Rejected	Interval
loan_intent	Default	4	Input	Nominal
loan_percent_in	Default	4	Rejected	Interval
loan_status	Default	4	Target	Binary
person_age	Default	4	Rejected	Interval
person_emp_length	Default	4	Rejected	Interval
person_home_ownership	Default	4	Input	Nominal
person_income	Default	4	Rejected	Interval

Figure 10 Transform Variables Editor

28			Input			
29	Input Name	Role	Level	Name	Level	Formula
30						
31	IMP_REP_person_emp_length	INPUT	INTERVAL	LOG_IMP_REP_person_emp_length	INTERVAL	log(IMP_REP_person_emp_length + 1)
32	REP_loan_amnt	INPUT	INTERVAL	LOG_REP_loan_amnt	INTERVAL	log(REP_loan_amnt + 1)
33	REP_person_age	INPUT	INTERVAL	LOG_REP_person_age	INTERVAL	log(REP_person_age + 1)
34	REP_person_income	INPUT	INTERVAL	LOG_REP_person_income	INTERVAL	log(REP_person_income + 1)

Figure 11 Transform Variables Output

It is evident from the above that the log transformation has been successfully implemented for Person_emp_length, Loan_amt, Person_age, Person_income.

To determine whether the preprocessing done has had an impact on the variable in the dataset or not, another StatExplore is used to get a brief overview.

70				Standard	Non						
71	Variable	Role	Mean	Deviation	Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
72											
73	IMP_REP_loan_int_rate	INPUT	11.01085	3.079063	32581	0	5.42	11.01085	20.73307	0.212391	-0.45355
74	LOG_IMP_REP_person_emp_length	INPUT	1.490572	0.776354	32581	0	0	1.609438	2.902387	-0.51247	-0.52048
75	LOG_REP_loan_amnt	INPUT	8.940129	0.708655	32581	0	6.216606	8.987322	10.25964	-0.47127	-0.04725
76	LOG_REP_person_income	INPUT	10.922	0.55473	32581	0	8.2943	10.91511	12.43728	-0.0503	0.368719
77	REP_ch_person_cred_hist_length	INPUT	5.752962	3.843725	32581	0	2	4	17.96921	1.252494	0.943305
78	REP_loan_percent_income	INPUT	0.169594	0.104588	32581	0	0	0.15	0.490549	0.914667	0.413322
79	REP_person_age	INPUT	27.61903	5.774188	32581	0	20	26	46.77884	1.373176	1.581538

Figure 12 StatExplore Output

As can be seen from the above, the dataset has undergone successful preprocessing, and it is now prepared for further model development.

2.4.4 Data Sampling

Under sampling is done to make the target variable more uniformly distributed and to improve the model's prediction because the target variable loan_status was discovered to be imbalanced in the prior EDA.

51		Numeric	Formatted	Frequency		
52	Variable	Value	Value	Count	Percent	Label
53						
54	loan_status	0	0	25473	78.1836	
55	loan_status	1	1	7108	21.8164	
56						
57						
58	Data=	SAMPLE				
59						
60		Numeric	Formatted	Frequency		
61	Variable	Value	Value	Count	Percent	Label
62						
63	loan_status	0	0	7108	50	
64	loan_status	1	1	7108	50	

Figure 13 Sampling Output

It is evident that sampling has been carried out successfully and that the target variable has a balanced distribution.

2.4.5 Data Partitioning

Training the model on one subset of data and assessing its performance on the other require splitting the data into training (70%) and validation (30%) sets.

Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0

Figure 14 Splitting Data into Training and Validation

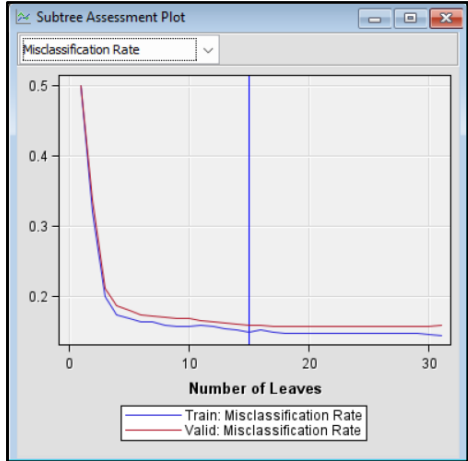
60	Data=	TRAIN				
61						
62	Variable	Numeric	Formatted	Frequency		
63		Value	Value	Count	Percent	Label
64						
65	loan_status	0	0	4975	49.9950	
66	loan_status	1	1	4976	50.0050	
67						
68						
69	Data=	VALIDATE				
70						
71		Numeric	Formatted	Frequency		
72	Variable	Value	Value	Count	Percent	Label
73						
74	loan_status	0	0	2133	50.0117	
75	loan_status	1	1	2132	49.9883	

Figure 15 Train and Validate Data Output

2.5 Predictive Modeling

Both tree-based and regression models are used in the model creation process to meet the first objective. Four variations are carried out with the tree-based model and four variations are carried out with the regression model, as was previously shown in the diagram flow. The one standard Decision Trees, two HP Trees, and one HP Forest node make up the tree-based models. The logistic regression model includes forward, backward and clog-log link function forward logistic regression as well as polynomial and stepwise HP logistic regression.

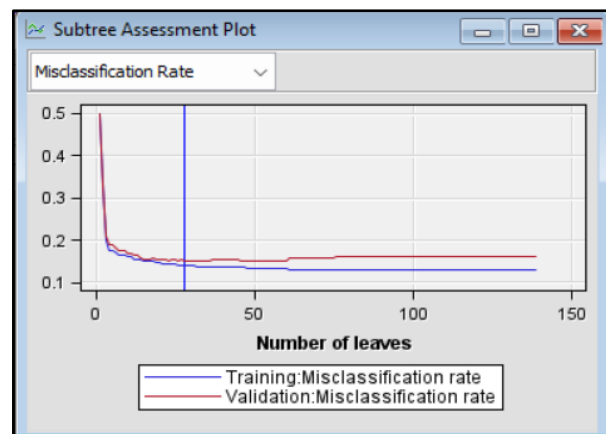
2.5.1 Tree-Based Model

Model Variation	Construction, Optimization Properties	Validation and Result																																																																																																																								
Decision Tree	<p>The maximum branch is set to 2.</p> <table><tr><th colspan="2">Splitting Rule</th></tr><tr><td>Interval Target Crite</td><td>ProbF</td></tr><tr><td>Nominal Target Crite</td><td>ProbChisq</td></tr><tr><td>Ordinal Target Crite</td><td>Entropy</td></tr><tr><td>Significance Level</td><td>0.2</td></tr><tr><td>Missing Values</td><td>Use in search</td></tr><tr><td>Use Input Once</td><td>No</td></tr><tr><td>Maximum Branch</td><td>2</td></tr><tr><td>Maximum Depth</td><td>6</td></tr><tr><td>Minimum Categorical</td><td>5</td></tr></table> <p>The number of leaves in the subtree of a decision tree model is set to 15, as here is where the model is most optimised.</p> <table><tr><th colspan="2">Subtree</th></tr><tr><td>Method</td><td>N</td></tr><tr><td>Number of Leaves</td><td>15</td></tr><tr><td>Assessment Measure</td><td>Misclassification</td></tr><tr><td>Assessment Fraction</td><td>0.25</td></tr></table>	Splitting Rule		Interval Target Crite	ProbF	Nominal Target Crite	ProbChisq	Ordinal Target Crite	Entropy	Significance Level	0.2	Missing Values	Use in search	Use Input Once	No	Maximum Branch	2	Maximum Depth	6	Minimum Categorical	5	Subtree		Method	N	Number of Leaves	15	Assessment Measure	Misclassification	Assessment Fraction	0.25	 <table><tr><th>Target</th><th>Fit Statistics</th><th>Statistics Label</th><th>Train</th><th>Validation</th></tr><tr><td>loan_status</td><td>_NOBS_</td><td>Sum of Frequenc...</td><td>9951</td><td>4265</td></tr><tr><td>loan_status</td><td>_MISC_</td><td>Misclassification ...</td><td>0.149935</td><td>0.159672</td></tr><tr><td>loan_status</td><td>_MAX_</td><td>Maximum Absolu...</td><td>0.99827</td><td>0.978873</td></tr><tr><td>loan_status</td><td>_SSE_</td><td>Sum of Squared ...</td><td>2236.257</td><td>1019.02</td></tr><tr><td>loan_status</td><td>_ASE_</td><td>Average Squared...</td><td>0.112363</td><td>0.119463</td></tr><tr><td>loan_status</td><td>_RASE_</td><td>Root Average Sq...</td><td>0.335207</td><td>0.345634</td></tr><tr><td>loan_status</td><td>_DIV_</td><td>Divisor for ASE</td><td>19902</td><td>8530</td></tr><tr><td>loan_status</td><td>_DFT_</td><td>Total Degrees of ...</td><td>9951</td><td>.</td></tr></table> <p>Model shows lower misclassification in training (14.9%) than validation (15.9%), indicating slight overfitting.</p> <table><tr><th colspan="5">Event Classification Table</th></tr><tr><td colspan="5">Data Role=TRAIN Target=loan_status Target Label=' '</td></tr><tr><td></td><td>False</td><td>True</td><td>False</td><td>True</td></tr><tr><td></td><td>Negative</td><td>Negative</td><td>Positive</td><td>Positive</td></tr><tr><td></td><td>1311</td><td>4794</td><td>181</td><td>3665</td></tr><tr><td colspan="5">Data Role=VALIDATE Target=loan_status Target Label=' '</td></tr><tr><td></td><td>False</td><td>True</td><td>False</td><td>True</td></tr><tr><td></td><td>Negative</td><td>Negative</td><td>Positive</td><td>Positive</td></tr><tr><td></td><td>574</td><td>2026</td><td>107</td><td>1558</td></tr></table> <p>The model displays a lower false positive rate in validation compared to training, indicating improved generalization on unseen data.</p>	Target	Fit Statistics	Statistics Label	Train	Validation	loan_status	_NOBS_	Sum of Frequenc...	9951	4265	loan_status	_MISC_	Misclassification ...	0.149935	0.159672	loan_status	_MAX_	Maximum Absolu...	0.99827	0.978873	loan_status	_SSE_	Sum of Squared ...	2236.257	1019.02	loan_status	_ASE_	Average Squared...	0.112363	0.119463	loan_status	_RASE_	Root Average Sq...	0.335207	0.345634	loan_status	_DIV_	Divisor for ASE	19902	8530	loan_status	_DFT_	Total Degrees of ...	9951	.	Event Classification Table					Data Role=TRAIN Target=loan_status Target Label=' '						False	True	False	True		Negative	Negative	Positive	Positive		1311	4794	181	3665	Data Role=VALIDATE Target=loan_status Target Label=' '						False	True	False	True		Negative	Negative	Positive	Positive		574	2026	107	1558
	Splitting Rule																																																																																																																									
Interval Target Crite	ProbF																																																																																																																									
Nominal Target Crite	ProbChisq																																																																																																																									
Ordinal Target Crite	Entropy																																																																																																																									
Significance Level	0.2																																																																																																																									
Missing Values	Use in search																																																																																																																									
Use Input Once	No																																																																																																																									
Maximum Branch	2																																																																																																																									
Maximum Depth	6																																																																																																																									
Minimum Categorical	5																																																																																																																									
Subtree																																																																																																																										
Method	N																																																																																																																									
Number of Leaves	15																																																																																																																									
Assessment Measure	Misclassification																																																																																																																									
Assessment Fraction	0.25																																																																																																																									
Target	Fit Statistics	Statistics Label	Train	Validation																																																																																																																						
loan_status	_NOBS_	Sum of Frequenc...	9951	4265																																																																																																																						
loan_status	_MISC_	Misclassification ...	0.149935	0.159672																																																																																																																						
loan_status	_MAX_	Maximum Absolu...	0.99827	0.978873																																																																																																																						
loan_status	_SSE_	Sum of Squared ...	2236.257	1019.02																																																																																																																						
loan_status	_ASE_	Average Squared...	0.112363	0.119463																																																																																																																						
loan_status	_RASE_	Root Average Sq...	0.335207	0.345634																																																																																																																						
loan_status	_DIV_	Divisor for ASE	19902	8530																																																																																																																						
loan_status	_DFT_	Total Degrees of ...	9951	.																																																																																																																						
Event Classification Table																																																																																																																										
Data Role=TRAIN Target=loan_status Target Label=' '																																																																																																																										
	False	True	False	True																																																																																																																						
	Negative	Negative	Positive	Positive																																																																																																																						
	1311	4794	181	3665																																																																																																																						
Data Role=VALIDATE Target=loan_status Target Label=' '																																																																																																																										
	False	True	False	True																																																																																																																						
	Negative	Negative	Positive	Positive																																																																																																																						
	574	2026	107	1558																																																																																																																						

HP Tree
C-C

The Subtree method is set to Cost-complexity, and the selection method is automatic because this is where the model is best fit.

Subtree	
Subtree Method	Cost-Complexity
Selection Method	Automatic
Confidence	0.25
Nominal Target Asses	Entropy
Minimum Subtree	No
Assessment Threshold	1.0
Number of Leaves	15
Cross Validation Fold	10
Cross Validation Seed	12345



Target	Fit Statistics	Statistics Label	Train	Validation
loan_status	_ASE_	Average Squared ...	0.103774	0.113144
loan_status	_DIV_	Divisor for ASE	19902	8530
loan_status	_MAX_	Maximum Absolut...	0.997859	1
loan_status	_NOBS_	Sum of Frequenci...	9951	4265
loan_status	_RASE_	Root Average Squ...	0.322139	0.336369
loan_status	_SSE_	Sum of Squared ...	2065.306	965.1201
loan_status	_DISF_	Frequency of Cla...	9951	4265
loan_status	_MISC_	Misclassification ...	0.140589	0.150996
loan_status	_WRONG_	Number of Wrong...	1399	644

Model shows lower misclassification in training (14.05%) than validation (15.10%), indicating slight overfitting.

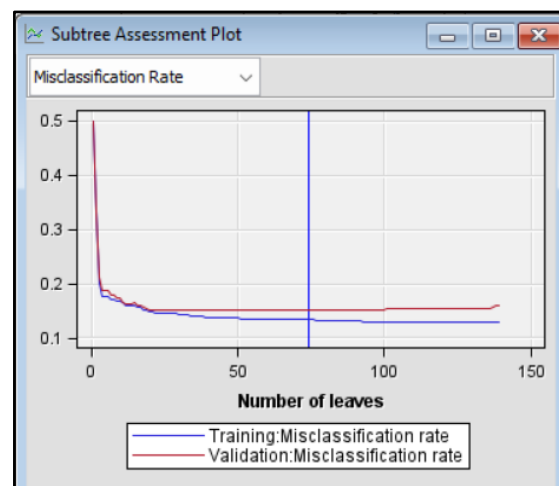
120	Event Classification Table			
121				
122	Data Role=TRAIN Target=loan_status Target Label=' '			
123				
124	False	True	False	True
125	Negative	Negative	Positive	Positive
126				
127	1291	4867	108	3685
128				
129	Data Role=VALIDATE Target=loan_status Target Label=' '			
130				
131				
132	False	True	False	True
133	Negative	Negative	Positive	Positive
134				
135	559	2048	85	1573

Higher true positive rates in training than validation; model might be overfitting. Fewer false positives in validation suggest moderate classification

HP Tree
C4.5

The Subtree method is set to C4.5, and the selection method is automatic as this is where the model is best fit.

Subtree	
Subtree Method	C4.5
Selection Method	Automatic
Confidence	0.25
Nominal Target Asses	Entropy
Minimum Subtree	No
Assessment Threshold	1.0
Number of Leaves	15
Cross Validation Fold	10
Cross Validation Seed	12345



Target	Fit Statistics	Statistics Label	Train	Validation
loan_status	_ASE_	Average Squar...	0.09591	0.111633
loan_status	_DIV_	Divisor for ASE	19902	8530
loan_status	_MAX_	Maximum Abso...	0.997859	1
loan_status	_NOBS_	Sum of Freque...	9951	4265
loan_status	_RASE_	Root Average S...	0.309694	0.334115
loan_status	_SSE_	Sum of Square...	1908.803	952.229
loan_status	_DISF_	Frequency of CI...	9951	4265
loan_status	_MISC_	Misclassification...	0.133956	0.150996
loan_status	_WRONG_	Number of Wro...	1333	644

Model shows lower misclassification in training (13.33%) than validation (15.1%), indicating slight overfitting.

120	Event Classification Table			
121				
122	Data Role=TRAIN Target=loan_status Target Label=' '			
123				
124	False	True	False	True
125	Negative	Negative	Positive	Positive
126				
127	1229	4871	104	3747
128				
129				
130	Data Role=VALIDATE Target=loan_status Target Label=' '			
131				
132	False	True	False	True
133	Negative	Negative	Positive	Positive
134				
135	559	2048	85	1573

Model performs better on training data, yet faces higher false negatives in validation, indicating overfitting risks.

In tree options the maximum no. of tree is set 50.

Tree Options	
Maximum Number of	50
Seed	1234
Type of Sample	Proportion
Proportion of Obs in	0.6
Number of Obs in Ea.	



Target	Fit Statistics	Statistics Label	Train	Validation
loan_status	_ASE_	Average Sq...	0.111533	0.120113
loan_status	_DIV_	Divisor for A...	19902	8530
loan_status	_MAX_	Maximum A...	0.907705	0.928778
loan_status	_NOBS_	Sum of Fre...	9951	4265
loan_status	_RASE_	Root Avera...	0.333966	0.346574
loan_status	_SSE_	Sum of Squa...	2219.733	1024.566
loan_status	_DISF_	Frequency ...	9951	4265
loan_status	_MISC_	Misclassific...	0.150638	0.16694
loan_status	_WRONG_	Number of ...	1499	712

Validation set shows higher misclassification and slightly worse fit statistics compared to the training set, indicating potential overfitting.

HP Forest

369	Event Classification Table			
370				
371	Data Role=TRAIN Target=loan_status Target Label=' '			
372				
373	False	True	False	True
374	Negative	Negative	Positive	Positive
375				
376	1045	4521	454	3931
377				
378	Data Role=VALIDATE Target=loan_status Target Label=' '			
379				
380				
381	False	True	False	True
382	Negative	Negative	Positive	Positive
383				
384	482	1903	230	1650

Training data has fewer false positives and negatives compared to validation data, suggesting possible overfitting in model training.

2.5.2 Regression Model

Model Variation	Construction, Optimization Properties	Validation and Result																																																																																																																									
Log Regression Forward	<p>The link function is set to logit, and the model selection model is forward. The selection criterion is validation misclassification.</p>																																																																																																																										
	<div><div>Class Targets</div><table><tr><td>Regression Type</td><td>Logistic Regression</td></tr><tr><td>Link Function</td><td>Logit</td></tr></table><div>Model Options</div><table><tr><td>Suppress Intercept</td><td>No</td></tr><tr><td>Input Coding</td><td>Deviation</td></tr></table><div>Model Selection</div><table><tr><td>Selection Model</td><td>Forward</td></tr><tr><td>Selection Criterion</td><td>Validation Misclassification</td></tr><tr><td>Use Selection Defaults</td><td>Yes</td></tr><tr><td>Selection Options</td><td></td></tr></table></div>	Regression Type	Logistic Regression	Link Function	Logit	Suppress Intercept	No	Input Coding	Deviation	Selection Model	Forward	Selection Criterion	Validation Misclassification	Use Selection Defaults	Yes	Selection Options		<table><tr><th>Target</th><th>Fit Statistics</th><th>Statistics Label</th><th>Train</th><th>Validation</th></tr><tr><td>loan_status</td><td>_AIC_</td><td>Akaike's Inf...</td><td>8418.519</td><td>.</td></tr><tr><td>loan_status</td><td>_ASE_</td><td>Average Sq...</td><td>0.134141</td><td>0.14172</td></tr><tr><td>loan_status</td><td>_AVERR_</td><td>Average Err...</td><td>0.42129</td><td>0.440747</td></tr><tr><td>loan_status</td><td>_DFE_</td><td>Degrees of ...</td><td>9934</td><td>.</td></tr><tr><td>loan_status</td><td>_DFM_</td><td>Model Degr...</td><td>17</td><td>.</td></tr><tr><td>loan_status</td><td>_DFT_</td><td>Total Degr...</td><td>9951</td><td>.</td></tr><tr><td>loan_status</td><td>_DIV_</td><td>Divisor for A...</td><td>19902</td><td>8530</td></tr><tr><td>loan_status</td><td>_ERR_</td><td>Error Functi...</td><td>8384.519</td><td>3759.573</td></tr><tr><td>loan_status</td><td>_FPE_</td><td>Final Predic...</td><td>0.1346</td><td>.</td></tr><tr><td>loan_status</td><td>_MAX_</td><td>Maximum A...</td><td>0.992547</td><td>0.993443</td></tr><tr><td>loan_status</td><td>_MSE_</td><td>Mean Squa...</td><td>0.13437</td><td>0.14172</td></tr><tr><td>loan_status</td><td>_NOBS_</td><td>Sum of Fre...</td><td>9951</td><td>4265</td></tr><tr><td>loan_status</td><td>_NW_</td><td>Number of ...</td><td>17</td><td>.</td></tr><tr><td>loan_status</td><td>_RASE_</td><td>Root Avera...</td><td>0.366252</td><td>0.376457</td></tr><tr><td>loan_status</td><td>_RFPE_</td><td>Root Final ...</td><td>0.366878</td><td>.</td></tr><tr><td>loan_status</td><td>_RMSE_</td><td>Root Mean ...</td><td>0.366565</td><td>0.376457</td></tr><tr><td>loan_status</td><td>_SBC_</td><td>Schwarz's ...</td><td>8541.012</td><td>.</td></tr><tr><td>loan_status</td><td>_SSE_</td><td>Sum of Squ...</td><td>2669.667</td><td>1208.872</td></tr><tr><td>loan_status</td><td>_SUMW_</td><td>Sum of Cas...</td><td>19902</td><td>8530</td></tr><tr><td>loan_status</td><td>_MISC_</td><td>Misclassific...</td><td>0.185308</td><td>0.196483</td></tr></table>	Target	Fit Statistics	Statistics Label	Train	Validation	loan_status	_AIC_	Akaike's Inf...	8418.519	.	loan_status	_ASE_	Average Sq...	0.134141	0.14172	loan_status	_AVERR_	Average Err...	0.42129	0.440747	loan_status	_DFE_	Degrees of ...	9934	.	loan_status	_DFM_	Model Degr...	17	.	loan_status	_DFT_	Total Degr...	9951	.	loan_status	_DIV_	Divisor for A...	19902	8530	loan_status	_ERR_	Error Functi...	8384.519	3759.573	loan_status	_FPE_	Final Predic...	0.1346	.	loan_status	_MAX_	Maximum A...	0.992547	0.993443	loan_status	_MSE_	Mean Squa...	0.13437	0.14172	loan_status	_NOBS_	Sum of Fre...	9951	4265	loan_status	_NW_	Number of ...	17	.	loan_status	_RASE_	Root Avera...	0.366252	0.376457	loan_status	_RFPE_	Root Final ...	0.366878	.	loan_status	_RMSE_	Root Mean ...	0.366565	0.376457	loan_status	_SBC_	Schwarz's ...	8541.012	.	loan_status	_SSE_	Sum of Squ...	2669.667	1208.872	loan_status	_SUMW_	Sum of Cas...	19902	8530	loan_status	_MISC_	Misclassific...	0.185308	0.196483
	Regression Type	Logistic Regression																																																																																																																									
	Link Function	Logit																																																																																																																									
	Suppress Intercept	No																																																																																																																									
	Input Coding	Deviation																																																																																																																									
	Selection Model	Forward																																																																																																																									
	Selection Criterion	Validation Misclassification																																																																																																																									
	Use Selection Defaults	Yes																																																																																																																									
	Selection Options																																																																																																																										
Target	Fit Statistics	Statistics Label	Train	Validation																																																																																																																							
loan_status	_AIC_	Akaike's Inf...	8418.519	.																																																																																																																							
loan_status	_ASE_	Average Sq...	0.134141	0.14172																																																																																																																							
loan_status	_AVERR_	Average Err...	0.42129	0.440747																																																																																																																							
loan_status	_DFE_	Degrees of ...	9934	.																																																																																																																							
loan_status	_DFM_	Model Degr...	17	.																																																																																																																							
loan_status	_DFT_	Total Degr...	9951	.																																																																																																																							
loan_status	_DIV_	Divisor for A...	19902	8530																																																																																																																							
loan_status	_ERR_	Error Functi...	8384.519	3759.573																																																																																																																							
loan_status	_FPE_	Final Predic...	0.1346	.																																																																																																																							
loan_status	_MAX_	Maximum A...	0.992547	0.993443																																																																																																																							
loan_status	_MSE_	Mean Squa...	0.13437	0.14172																																																																																																																							
loan_status	_NOBS_	Sum of Fre...	9951	4265																																																																																																																							
loan_status	_NW_	Number of ...	17	.																																																																																																																							
loan_status	_RASE_	Root Avera...	0.366252	0.376457																																																																																																																							
loan_status	_RFPE_	Root Final ...	0.366878	.																																																																																																																							
loan_status	_RMSE_	Root Mean ...	0.366565	0.376457																																																																																																																							
loan_status	_SBC_	Schwarz's ...	8541.012	.																																																																																																																							
loan_status	_SSE_	Sum of Squ...	2669.667	1208.872																																																																																																																							
loan_status	_SUMW_	Sum of Cas...	19902	8530																																																																																																																							
loan_status	_MISC_	Misclassific...	0.185308	0.196483																																																																																																																							
		<p>Model demonstrates higher error and misclassification rates in validation compared to training, suggesting potential overfitting and limited generalization.</p>																																																																																																																									
		<div>Event Classification Table</div> <div>Data Role=TRAIN Target=loan_status Target Label=' '</div> <table><tr><td>False</td><td>True</td><td>False</td><td>True</td></tr><tr><td>Negative</td><td>Negative</td><td>Positive</td><td>Positive</td></tr><tr><td>1012</td><td>4143</td><td>832</td><td>3964</td></tr></table> <div>Data Role=VALIDATE Target=loan_status Target Label=' '</div> <table><tr><td>False</td><td>True</td><td>False</td><td>True</td></tr><tr><td>Negative</td><td>Negative</td><td>Positive</td><td>Positive</td></tr><tr><td>472</td><td>1767</td><td>366</td><td>1660</td></tr></table>	False	True	False	True	Negative	Negative	Positive	Positive	1012	4143	832	3964	False	True	False	True	Negative	Negative	Positive	Positive	472	1767	366	1660																																																																																																	
False	True	False	True																																																																																																																								
Negative	Negative	Positive	Positive																																																																																																																								
1012	4143	832	3964																																																																																																																								
False	True	False	True																																																																																																																								
Negative	Negative	Positive	Positive																																																																																																																								
472	1767	366	1660																																																																																																																								
		<p>Training data shows higher true positives and negatives than validation data, suggesting possible overfitting or model instability on new data.</p>																																																																																																																									

</

Parameter	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	157.45	63.5984	Infty	2.48	0.0133
cb_person_default_on_file N	0.02665	0.08522	Infty	0.31	0.7545
cb_person_default_on_file Y	0
loan_grade A	-13.8515	52.9979	Infty	-0.26	0.7938
loan_grade B	-13.5790	52.9978	Infty	-0.26	0.7978
loan_grade C	-13.1926	52.9977	Infty	-0.25	0.8034
loan_grade D	-11.0043	52.9975	Infty	-0.21	0.8355
loan_grade E	-10.7838	52.9973	Infty	-0.20	0.8388
loan_grade F	-10.6204	52.9977	Infty	-0.20	0.8412
loan_grade G	0
loan_intent DEBTCONSOLIDATION	1.2089	0.1028	Infty	11.76	<.0001
loan_intent EDUCATION	0.4920	0.1021	Infty	4.82	<.0001
loan_intent HOMEIMPROVEMENT	1.3691	0.1146	Infty	11.95	<.0001
loan_intent MEDICAL	0.8739	0.1007	Infty	8.68	<.0001
loan_intent PERSONAL	0.5933	0.1037	Infty	5.72	<.0001
loan_intent VENTURE	0
person_home_ownership MORTGAGE	-0.6127	0.06530	Infty	-9.38	<.0001
person_home_ownership OTHER	-0.3642	0.4475	Infty	-0.81	0.4157
person_home_ownership OWN	-3.1540	0.1633	Infty	-19.31	<.0001
person_home_ownership RENT	0
IMP_REP_loan_int_rate	-0.3983	0.4122	Infty	-0.97	0.3339
LOG_IMP_REP_person_emp_length	4.3966	1.5513	Infty	2.83	0.0046
LOG_REP_loan_amnt	5.6640	4.8655	Infty	1.16	0.2444
LOG_REP_person_age	-2.6734	17.4119	Infty	-0.15	0.8780
LOG_REP_person_income	-27.7188	5.5909	Infty	-4.96	<.0001
REP_cb_person_cred_hist_length	0.2362	0.7894	Infty	0.30	0.7648
REP_loan_percent_income	-20.3593	33.6037	Infty	-0.61	0.5446
IMP_REP_loan_int_rate*IMP_REP_loan_int_rate	-0.01053	0.003972	Infty	-2.65	0.0080
IMP_REP_loan_int_rate*LOG_IMP_REP_person_emp_length	-0.09159	0.01353	Infty	-6.77	<.0001
IMP_REP_loan_int_rate*LOG_REP_loan_amnt	0.09802	0.04135	Infty	2.37	0.0178
IMP_REP_loan_int_rate*LOG_REP_person_age	0.1289	0.1109	Infty	1.16	0.2449
IMP_REP_loan_int_rate*LOG_REP_person_income	-0.03072	0.04393	Infty	-0.70	0.4844
IMP_REP_loan_int_rate*REP_cb_person_cred_hist_length	0.002807	0.005227	Infty	0.54	0.5913
IMP_REP_loan_int_rate*REP_loan_percent_income	-0.8610	0.2805	Infty	-3.07	0.0021
LOG_IMP_REP_person_emp_length*LOG_IMP_REP_person_emp_length	0.1122	0.04710	Infty	2.38	0.0172
LOG_IMP_REP_person_emp_length*LOG_REP_loan_amnt	0.2672	0.1603	Infty	1.67	0.0956
LOG_IMP_REP_person_emp_length*LOG_REP_person_age	-0.2435	0.4069	Infty	-0.60	0.5495
LOG_IMP_REP_person_emp_length*LOG_REP_person_income	-0.4261	0.1712	Infty	-2.49	0.0128
LOG_IMP_REP_person_emp_length*REP_cb_person_cred_hist_length	0.003218	0.01899	Infty	0.17	0.8654
LOG_IMP_REP_person_emp_length*REP_loan_percent_income	-3.1537	1.0849	Infty	-2.91	0.0036
LOG_REP_loan_amnt*LOG_REP_loan_amnt	-0.05551	0.3009	Infty	-0.18	0.8536
LOG_REP_loan_amnt*LOG_REP_person_age	-0.3239	1.3453	Infty	-0.24	0.8097
LOG_REP_loan_amnt*LOG_REP_person_income	-0.2947	0.5740	Infty	-0.51	0.6076
LOG_REP_loan_amnt*REP_cb_person_cred_hist_length	-0.02453	0.06103	Infty	-0.40	0.6878
LOG_REP_loan_amnt*REP_loan_percent_income	26.4836	4.9656	Infty	5.33	<.0001
LOG_REP_person_age*LOG_REP_person_age	-1.1731	2.5505	Infty	-0.46	0.6456
LOG_REP_person_age*LOG_REP_person_income	1.0092	1.4199	Infty	0.71	0.4772
LOG_REP_person_age*REP_cb_person_cred_hist_length	-0.00041	0.2296	Infty	-0.00	0.9986
LOG_REP_person_age*REP_loan_percent_income	4.3728	8.8757	Infty	0.49	0.6222
LOG_REP_person_income*LOG_REP_person_income	1.1211	0.3276	Infty	3.42	0.0006
LOG_REP_person_income*REP_cb_person_cred_hist_length	-0.00708	0.06447	Infty	-0.11	0.9126
LOG_REP_person_income*REP_loan_percent_income	-23.3747	4.6497	Infty	-5.03	<.0001
REP_cb_person_cred_hist_length*REP_cb_person_cred_hist_length	0.002033	0.005974	Infty	0.34	0.7337
REP_cb_person_cred_hist_length*REP_loan_percent_income	0.05747	0.4090	Infty	0.14	0.8883
REP_loan_percent_income*REP_loan_percent_income	3.0047	15.5408	Infty	0.19	0.8467

Parameter Estimates

Target	Fit Statistics	Statistics Label	Train	Validation
loan_status	_ASE_	Average Squ...	0.134141	0.14172
loan_status	_DIV_	Divisor for ASE	19902	8530
loan_status	_MAX_	Maximum Ab...	0.992547	0.993443
loan_status	_NOBS_	Sum of Freq...	9951	4265
loan_status	_RASE_	Root Average...	0.366252	0.376457
loan_status	_SSE_	Sum of Squa...	2669.667	1208.872
loan_status	_DISF_	Frequency of ...	9951	4265
loan_status	_MISC_	Misclassifica...	0.185308	0.196483
loan_status	_WRONG_	Number of W...	1844	838

Model demonstrates higher error and misclassification rates in validation compared to training, suggesting slight overfitting.

HP
Regression
Stepwise

Event Classification Table

Data Role=TRAIN Target=loan_status Target Label=' '

False Negative	True Negative	False Positive	True Positive
1012	4143	832	3964

Data Role=VALIDATE Target=loan_status Target Label=' '

False Negative	True Negative	False Positive	True Positive
472	1767	366	1660

Model has lower false positive and higher true negative rates in validation, suggesting improved accuracy and generalization from training.

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	21.8494	63.5573	Infty	0.34	0.7310
loan_grade A	-14.4056	63.5556	Infty	-0.23	0.8207
loan_grade B	-13.8781	63.5556	Infty	-0.22	0.8271
loan_grade C	-13.4585	63.5556	Infty	-0.21	0.8323
loan_grade D	-11.1177	63.5556	Infty	-0.17	0.8611
loan_grade E	-10.8157	63.5557	Infty	-0.17	0.8649
loan_grade F	-10.4714	63.5562	Infty	-0.16	0.8691
loan_grade G	0
loan_intent DEBTCONSOLIDATION	1.1435	0.09909	Infty	11.54	<.0001
loan_intent EDUCATION	0.4743	0.09713	Infty	4.88	<.0001
loan_intent HOMEIMPROVEMENT	1.2879	0.1096	Infty	11.75	<.0001
loan_intent MEDICAL	0.9517	0.09542	Infty	9.97	<.0001
loan_intent PERSONAL	0.5627	0.09915	Infty	5.68	<.0001
loan_intent VENTURE	0
person_home_ownership MORTGAGE	-0.5822	0.06120	Infty	-9.51	<.0001
person_home_ownership OTHER	-0.2702	0.4273	Infty	-0.63	0.5273
person_home_ownership OWN	-2.5788	0.1420	Infty	-18.15	<.0001
person_home_ownership RENT	0
LOG_REP_loan_amnt	-1.3294	0.05362	Infty	-24.79	<.0001
REP_loan_percent_income	15.4701	0.3934	Infty	39.32	<.0001

3 Model Interpretation to Understand Business

3.1 Tree-based Model Interpretation

The best tree-based model that was previously constructed is compared using a model comparison node below to determine which model performs the best at predicting credit risk.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Misclassification Rate	Valid: Misclassification Rate	Train: Roc Index	Valid: Roc Index
Y	HPTree2	HPTree2	HP Tree C4.5	loan_status	0.133956	0.150996	0.926	0.902
	HPTree	HPTree	HP Tree C-C	loan_status	0.140589	0.150996	0.91	0.896
	Tree	Tree	Decision Tr...	loan_status	0.149935	0.159672	0.885	0.876
	HPDMForest	HPDMForest	HP Forest	loan_status	0.150638	0.16694	0.924	0.911

Figure 16 Fit Statistics of Model Comparison

In determining the loan status, a comparison is made among the various predictive models using different metrics such as misclassification rates and ROC indices for both training and validation datasets. Based on the result of the assessment, HP Tree with C4.5 subtree method emerged as the best performer having the lowest misclassification rate during training with 13% and high ROC indices of 0.926 (training) and 0.902 (validation) implying that it predicts well on seen but generalizes better on unseen data. The HP Tree with Cost Complexity subtree method also exhibited good results but had comparatively higher misclassification rates as well as lower ROCs than HP Tree C4.5 thus indicating strong class distinction capabilities. On the other hand, Decision Tree recorded least desirable performance with highest misclassification rates both in the lowest ROC indices across all models tested further confirming its inefficiency in prediction and generalization. Despite achieving highest training ROC index i.e. 0.924, HP Forest experienced relatively high validation misclassification rate which could be associated with overfitting problems. This comparison demonstrates why it's critical to use models, such as HP Tree C4.5, that balance predictive power and complexity to achieve accuracy without overfitting and effectively separate classes.

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
HPDMForest	HP Forest	TRAIN	loan_status		1045	4521	454	3931
HPDMForest	HP Forest	VALIDATE	loan_status		482	1903	230	1650
Tree	Decision Tree	TRAIN	loan_status		1311	4794	181	3665
Tree	Decision Tree	VALIDATE	loan_status		574	2026	107	1558
HPTree	HP Tree C-C	TRAIN	loan_status		1291	4867	108	3685
HPTree	HP Tree C-C	VALIDATE	loan_status		559	2048	85	1573
HPTree2	HP Tree C4.5	TRAIN	loan_status		1229	4871	104	3747
HPTree2	HP Tree C4.5	VALIDATE	loan_status		559	2048	85	1573

Figure 17 Event Classification Table

Event classification table helps explain how HP Tree C4.5 model excels in accurately identifying loan defaults, minimizing incorrect default predictions (false positives) and adeptly recognizing genuine defaults (true positives). This accuracy means fewer customers are mistakenly classified as risky, fostering better customer relationships and trust. Simultaneously, the model's ability to correctly identify defaults protects the business from potential losses due to unpaid loans. Its balanced performance ensures that financial resources are used effectively, enhancing decision-making for loan approvals and risk management. Overall, adopting the HP Tree C4.5 model contributes to more efficient operational processes, reduces financial risks, and promotes a reliable lending environment.

HP Tree C4.5 Node Interpretation

1. Node 5

```
*-----*
NODE = 5
*-----*
MISSING(person_home_ownership) OR (person_home_ownership IS ONE OF OTHER, RENT)
AND ('Replacement: loan_percent_income'n >= 0.30414023)
  PREDICTED VALUE IS 1
  PREDICTED 1 = 1( 1641/1641)
  PREDICTED 0 = 0( 0/1641)
```

Conditions:

- Person_home_ownership is either "Other" or "Rent".
- Loan percent income is greater than or equal to 0.30414023.

Interpretation:

This node predicts a value of 1 with 100% probability, meaning it predicts a default every time under these conditions. It suggests that individuals who rent or have unspecified home ownership status, combined with a higher loan percent relative to their income, are very likely to default on their loans. This might indicate financial strain or instability among renters or those with unlisted home ownership statuses when taking on loans that constitute a significant portion of their income.

2. Node 18

```
*-----*
NODE = 18
*-----*
MISSING('Replacement: loan_percent_income'n) OR ('Replacement: loan_percent_income'n >= 0.15207012)
AND ('Transformed: Replacement: person_income'n < 9.8686338)
AND MISSING(loan_grade) OR (loan_grade IS ONE OF A, B, C)
AND MISSING('Replacement: loan_percent_income'n) OR ('Replacement: loan_percent_income'n < 0.30414023)
  PREDICTED VALUE IS 1
  PREDICTED 1 = 1( 252/252)
  PREDICTED 0 = 0( 0/252)
```

Conditions:

- Loan percent income is greater than or equal to 0.15207012.
- Transformed Person_income is less than 9.8686338.
- Loan_grade is one of A, B, or C.
- Loan_percent_income is less than 0.30414023.

Interpretation:

This node forecasts a 100% default rate, indicating an extremely high risk of default. Despite having relatively acceptable loan grades, the conditions indicate that people with lower income levels who take out loans that represent a sizable but manageable amount of their income are more likely to fail, probably because they lack the financial reserves necessary to handle the debt responsibly.

3. Node 45

```

*-----*
NODE = 45
*-----*
MISSING('Imputed: Replacement: loan_int_rate'n) OR ('Imputed: Replacement: loan_int_rate'n < 16.445413)
AND ('Transformed: Replacement: person_income'n < 10.697231)
AND MISSING(person_home_ownership) OR (person_home_ownership IS ONE OF MORTGAGE, OTHER, RENT)
AND (loan_intent IS ONE OF DEBTCONSOLIDATION, MEDICAL)
AND (loan_grade IS ONE OF D, E, F, G)
AND MISSING('Replacement: loan_percent_income'n) OR ('Replacement: loan_percent_income'n < 0.30414023)
PREDICTED VALUE IS 1
PREDICTED 1 = 0.9957( 232/233)
PREDICTED 0 = 0.004292( 1/233)

```

Conditions:

- Low loan interest rate.
- Transformed Person_income is less than 10.697231.
- Person_home_ownership is either "Mortgage," "Other," or "Rent."
- Loan intent is either "Debt Consolidation" or "Medical."
- Loan grade is one of D, E, F, or G.
- Loan percent income less than 0.30414023.

Interpretation:

This node forecasts a default at almost 100% certainty. The scenarios represent high risk situations in which people earning less money with below average credit (loan grades D-G) are borrowing money for either consolidating their debts or hospital bills. These types of credits have very high rates for not being paid back because they might be taken out of desperation rather than responsibility, this calls for a serious rethinking about how we approve loans and evaluate risks when dealing with such categories to ensure effective management is maintained.

3.2 Regression Model Interpretation

To determine which regression model is the best model for predicting credit risk, previous regression models are compared using model comparison node.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Misclassification Rate	Valid: Misclassification Rate	Train: Roc Index	Valid: Roc Index
Y	HPReg2	HPReg2	HP Regression Poly	loan_status	0.17747	0.183822	0.9	0.888
	Reg	Reg	Clog Regression Fwd	loan_status	0.179178	0.187339	0.892	0.881
	HPReg3	HPReg3	HP Regression Stepwise	loan_status	0.185308	0.196483	0.886	0.875
	Reg2	Reg2	Log Regression Forward	loan_status	0.185308	0.196483	0.886	0.875
	Reg4	Reg4	Log Regression Backw...	loan_status	0.185308	0.196483	0.886	0.875

Figure 18 Fit Statistics for Model Comparison Regression

The above table presents performance metrics for various regression models used to predict loan status. The metrics include training and validation misclassification rates, along with the ROC Index for both training and validation phases. The HP Regression Poly model shows the best overall performance with the lowest misclassification rates of 17% and 18% in both training and validation phases. It also maintains high ROC indices, scoring 0.9 in training and 0.888 in validation, indicating a strong ability to distinguish between loan statuses. The Clog Regression Fwd and two Log Regression models (Forward and Backward) display slightly higher misclassification rates in both training and validation phases, all hovering around 0.185308 to 0.196483 for validation, and similar ROC Index values, ranging from 0.875 to 0.892. These regression models are evidently varied in their approach to the problem, with polynomial, clog-log, and logistic regression techniques applied in forward and backward stepwise manners. This variety reflects different strategies in handling the complexity of the data and the decision boundaries between the different loan statuses. The ROC Index values suggest all models are reasonably good at predicting outcomes, but the Polynomial Regression with degree 2 stands out for its superior balance of error rate and ability to discriminate between classes.

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	Log Regression Forward	TRAIN	loan_status		1012	4143	832	3964	
Reg2	Log Regression Forward	VALIDATE	loan_status		472	1767	366	1660	
Reg4	Log Regression Backward	TRAIN	loan_status		1012	4143	832	3964	
Reg4	Log Regression Backward	VALIDATE	loan_status		472	1767	366	1660	
HPReg2	HP Regression Poly	TRAIN	loan_status		983	4192	783	3993	
HPReg2	HP Regression Poly	VALIDATE	loan_status		452	1801	332	1680	
HPReg3	HP Regression Stepwise	TRAIN	loan_status		1012	4143	832	3964	
HPReg3	HP Regression Stepwise	VALIDATE	loan_status		472	1767	366	1660	
Reg	Regression	TRAIN	loan_status		1091	4283	692	3885	
Reg	Regression	VALIDATE	loan_status		504	1838	295	1628	

Figure 19 Event Classification of Model Comparison

The HP Regression Polynomial model demonstrates superior predictive accuracy compared to other models when analysing loan status. In a business context, this means the model effectively distinguishes between clients who will likely repay loans (True Positives) and those who might default (True Negatives). During the validation phase, the model had fewer incorrect predictions (both False Positives and False Negatives) than other models, resulting in a misclassification rate of 784 instances out of total validations. This accuracy is critical because it helps minimize the risk of loan defaults while ensuring that loans are approved for clients who are financially capable, thereby optimizing financial performance and customer satisfaction.

Parameter Estimates					
Parameter	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	157.45	63.5984	Infty	2.48	0.0133
cb_person_default_on_file N	0.02665	0.08522	Infty	0.31	0.7545
cb_person_default_on_file Y	0
loan_grade A	-13.8515	52.9979	Infty	-0.26	0.7938
loan_grade B	-13.5790	52.9978	Infty	-0.26	0.7978
loan_grade C	-13.1926	52.9977	Infty	-0.25	0.8034
loan_grade D	-11.0043	52.9975	Infty	-0.21	0.8355
loan_grade E	-10.7838	52.9973	Infty	-0.20	0.8388
loan_grade F	-10.6204	52.9977	Infty	-0.20	0.8412
loan_grade G	0
loan_intent DEBTCONSOLIDATION	1.2089	0.1028	Infty	11.76	<.0001
loan_intent EDUCATION	0.4920	0.1021	Infty	4.82	<.0001
loan_intent HOMEIMPROVEMENT	1.3691	0.1146	Infty	11.95	<.0001
loan_intent MEDICAL	0.8739	0.1007	Infty	8.68	<.0001
loan_intent PERSONAL	0.5933	0.1037	Infty	5.72	<.0001
loan_intent VENTURE	0
person_home_ownership MORTGAGE	-0.6127	0.06530	Infty	-9.38	<.0001
person_home_ownership OTHER	-0.3642	0.4475	Infty	-0.81	0.4157
person_home_ownership OWN	-3.1540	0.1633	Infty	-19.31	<.0001
person_home_ownership RENT	0
IMP_REP_loan_int_rate	-0.3983	0.4122	Infty	-0.97	0.3339
LOG_IMP_REP_person_emp_length	4.3966	1.5513	Infty	2.83	0.0046
LOG_REP_loan_amnt	5.6640	4.8655	Infty	1.16	0.2444
LOG_REP_person_age	-2.6734	17.4119	Infty	-0.15	0.8780
LOG_REP_person_income	-27.7188	5.5909	Infty	-4.96	<.0001
REP_cb_person_cred_hist_length	0.2362	0.7894	Infty	0.30	0.7648
REP_loan_percent_income	-20.3593	33.6037	Infty	-0.61	0.5446
IMP_REP_loan_int_rate*IMP_REP_loan_int_rate	-0.01053	0.003972	Infty	-2.65	0.0080
IMP_REP_loan_int_rate*LOG_IMP_REP_person_emp_length	-0.09159	0.01353	Infty	-6.77	<.0001
IMP_REP_loan_int_rate*LOG_REP_loan_amnt	0.09802	0.04135	Infty	2.37	0.0178
IMP_REP_loan_int_rate*LOG_REP_person_age	0.1289	0.1109	Infty	1.16	0.2449
IMP_REP_loan_int_rate*LOG_REP_person_income	-0.03072	0.04393	Infty	-0.70	0.4844
IMP_REP_loan_int_rate*REP_cb_person_cred_hist_length	0.002807	0.005227	Infty	0.54	0.5913
IMP_REP_loan_int_rate*REP_loan_percent_income	-0.8610	0.2805	Infty	-3.07	0.0021
LOG_IMP_REP_person_emp_length*LOG_IMP_REP_person_emp_length	0.1122	0.04710	Infty	2.38	0.0172
LOG_IMP_REP_person_emp_length*LOG_REP_loan_amnt	0.2672	0.1603	Infty	1.67	0.0956
LOG_IMP_REP_person_emp_length*LOG_REP_person_age	-0.2435	0.4069	Infty	-0.60	0.5495
LOG_IMP_REP_person_emp_length*LOG_REP_person_income	-0.4261	0.1712	Infty	-2.49	0.0128
LOG_IMP_REP_person_emp_length*REP_cb_person_cred_hist_length	0.003218	0.01899	Infty	0.17	0.8654
LOG_IMP_REP_person_emp_length*REP_loan_percent_income	-3.1537	1.0849	Infty	-2.91	0.0036
LOG_REP_loan_amnt*LOG_REP_loan_amnt	-0.05551	0.3009	Infty	-0.18	0.8536
LOG_REP_loan_amnt*LOG_REP_person_age	-0.3239	1.3453	Infty	-0.24	0.8097
LOG_REP_loan_amnt*LOG_REP_person_income	-0.2947	0.5740	Infty	-0.51	0.6076
LOG_REP_loan_amnt*REP_cb_person_cred_hist_length	-0.02453	0.06103	Infty	-0.40	0.6878
LOG_REP_loan_amnt*REP_loan_percent_income	26.4836	4.9656	Infty	5.33	<.0001
LOG_REP_person_age*LOG_REP_person_age	-1.1731	2.5505	Infty	-0.46	0.6456
LOG_REP_person_age*LOG_REP_person_income	1.0092	1.4199	Infty	0.71	0.4772
LOG_REP_person_age*REP_cb_person_cred_hist_length	-0.00041	0.2296	Infty	-0.00	0.9986
LOG_REP_person_age*REP_loan_percent_income	4.3728	8.8757	Infty	0.49	0.6222
LOG_REP_person_income*LOG_REP_person_income	1.1211	0.3276	Infty	3.42	0.0006
LOG_REP_person_income*REP_cb_person_cred_hist_length	-0.00708	0.06447	Infty	-0.11	0.9126
LOG_REP_person_income*REP_loan_percent_income	-23.3747	4.6497	Infty	-5.03	<.0001
REP_cb_person_cred_hist_length*REP_cb_person_cred_hist_length	0.002033	0.005974	Infty	0.34	0.7337
REP_cb_person_cred_hist_length*REP_loan_percent_income	0.05747	0.4090	Infty	0.14	0.8883
REP_loan_percent_income*REP_loan_percent_income	3.0047	15.5408	Infty	0.19	0.8467

Figure 20 Parameter Estimate of HP Regression Poly

The HP Regression Polynomial model utilizes various predictors to estimate loan status, evident from the parameter estimates provided. The intercept and coefficients for loan intent, Person_home_ownership, and other transformed or interaction variables have significant effects on the model. Key highlights from the estimates include the strong influence of loan intents like debt consolidation, home improvement, medical, personal, and educational purposes on loan status, each positively linked with outcomes, indicating a higher likelihood of loan approval or favourable terms for these intents. Home ownership status, particularly owning a home, shows a substantial negative coefficient, significantly impacting loan status, suggesting those who own homes may have lower risk profiles or better financial stability. Another critical observation is the interaction terms, especially those involving the transformed loan interest rate and employment length, which also show significant effects. This indicates that the interaction between these variables is crucial in determining loan status, where higher interest rates combined with certain employment lengths or other financial indicators could increase the risk of loan default.

3.3 Recommendations

Below recommendations aim to leverage the predictive power of the regression model to optimize lending practices, enhance financial stability, and reduce the incidence of loan defaults, ultimately improving both customer satisfaction and financial outcomes for the financial organization the third objective is met.

- **Enhance Loan Approval Criteria:** Focus on refining loan approval criteria by placing significant weight on loan intents such as debt consolidation, home improvement, medical, personal, and educational purposes. These variables have shown strong positive coefficients, suggesting that loans for these purposes have a higher likelihood of being repaid. Tailoring loan products and approval criteria to these intents can improve loan performance.
- **Target Homeowners for Promotions:** The model indicates that owning a home is associated with a substantial negative coefficient in loan status prediction, suggesting lower risk profiles. Financial organization should consider targeted marketing and promotion strategies for homeowners, offering them favourable loan terms due to their demonstrated financial stability.
- **Adjust Interest Rates Based on Risk:** Variables involving the loan interest rate, particularly its interaction with employment length, show significant effects on loan

default risk. Adjust interest rates dynamically based on employment length and other financial indicators to better manage risk. This could involve higher rates for riskier loans or more competitive rates for borrowers with stable, long-term employment.

- **Develop Financial Products for Renters and Those with Unspecified Home Ownership:** Given that the model predicts high default rates for individuals who rent or have unspecified home ownership statuses when combined with high loan percent incomes, develop specialized financial products or advisory services aimed at helping these groups manage their debt more effectively.
- **Implement Robust Risk Assessment Tools:** Integrate the insights from the regression model into existing risk assessment frameworks to enhance the prediction of loan defaults. This integration should include attention to the identified significant predictors and their interactions, such as loan amount relative to income and person income transformations, to refine risk assessment processes and decision-making tools in real-time lending environments.

4 Conclusion

The comprehensive analysis of various predictive models, specifically focusing on the HP Tree with C4.5 and HP Regression Polynomial models, demonstrates their effectiveness in accurately predicting loan defaults. These models not only help in identifying potential defaulters with high accuracy but also ensure that loans are extended to credit-worthy applicants, thereby minimizing financial risks. The detailed node interpretations provide actionable insights into specific circumstances that increase the likelihood of defaults, such as high loan percentages of income among renters and specific loan intents like debt consolidation. By implementing the recommended strategies, financial organization can refine their lending practices, target the right customer segments, and adjust their risk assessment protocols. This will lead to more robust financial health for the organization, reduced incidence of defaults, and enhanced customer satisfaction. Overall, adopting these advanced predictive models and incorporating the suggested recommendations will significantly strengthen the decision-making process in loan approvals and risk management, aligning operational strategies with business objectives for better financial outcomes.

5 References

- Kanaparthi, V. (2023, April). Credit Risk Prediction using Ensemble Machine Learning Algorithms. *International Conference on Inventive Computation Technologies (ICICT)*, 41-47. doi:<https://doi.org/10.1109/ICICT57646.2023.10134486>
- Omari, F. S. (2023, January). A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach. *World Journal of Advanced Engineering Technology and Science* , 9-14. doi:10.30574/wjaets.2023.8.1.0147