SINGAPORE MANAGEMENT UNIVERSITY

ISSS608– Visual Analytics and Application

Shiny App User Guide

Loan Default Prediction Challenge of Nigeria loans

Professor:

Dr. KAM Tin Seong

Group 5

HOANG HUY, HOU TAO, LI ZIYI

LOAN DEFAULT PREDICTION SHINY APP

Operation Manual

V 1.0

Published Date: 31 Mar 2023

Table of Content

1. General Settings

In this shiny application, A common parameter - Loan Type is available on all functionalities; it has two options:

- **New loan** data is designed for customer loan default analysis for new customers based on variables identified more relative to new customers.
- **Repeat Loan** data is designed for customer loan default analysis for existing customers with multiple bank loan histories; the variables identified are more relative to existing customers.

For more details about the variables of each loan type, please refer to the introduction page on the main **website.**

2. Univariate

Step1: Select which datasets to be used (New Loans datasets or Repeated Loans dataset)

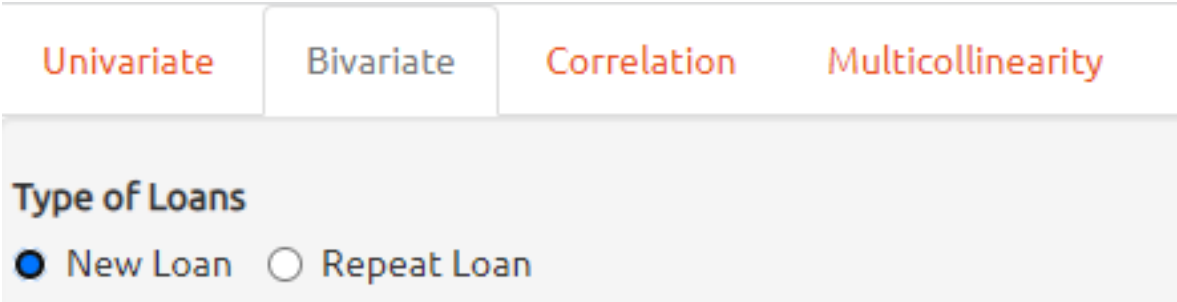
Step 2: Next users must choose which variables to be analyze under Univariate Analysis

The result bar chart will show you the distributions of the loan data as per the selected variable.

3. Bivariate

Step 1: Select which datasets to be used (New Loans datasets or Repeated Loans dataset).

Different type of loans would generate different variables to be studied for bivariate analysis later, as can been below.



Step 2: Choose two variables X and Y to analyze if there is any concurrent relation between two variables. Elements from variable X will be put on the horizontal axis and elements from variable Y will be located on the vertical axis.

Continuous Variable vs. Continuous Variable

Univariate | Bivariate | Correlation | Multicollinearity

Type of Loans

- (●) New Loan
- ( ) Repeat Loan

Select variable X

- (●) Age at Loan
- ( ) Age at Loan 25th Pctile
- ( ) Approval Duration Category
- ( ) Bank Account Type
- ( ) Bank Account Type Recode
- ( ) Bank Name
- ( ) Credit Rating
- ( ) Education Level Risk Category
- ( ) Employment Status Risk Category
- ( ) Term Days
- ( ) Referral

Select variable Y

- ( ) Age at Loan
- (●) Age at Loan 25th Pctile
- ( ) Approval Duration Category
- ( ) Bank Account Type
- ( ) Bank Account Type Recode
- ( ) Bank Name
- ( ) Credit Rating
- ( ) Education Level Risk Category
- ( ) Employment Status Risk Category
- ( ) Term Days
- ( ) Referral

Categorical Variable vs. Categorical Variable

Univariate | Bivariate | Correlation | Multicollinearity

Type of Loans

- ( ) New Loan
- (●) Repeat Loan

Select variable X

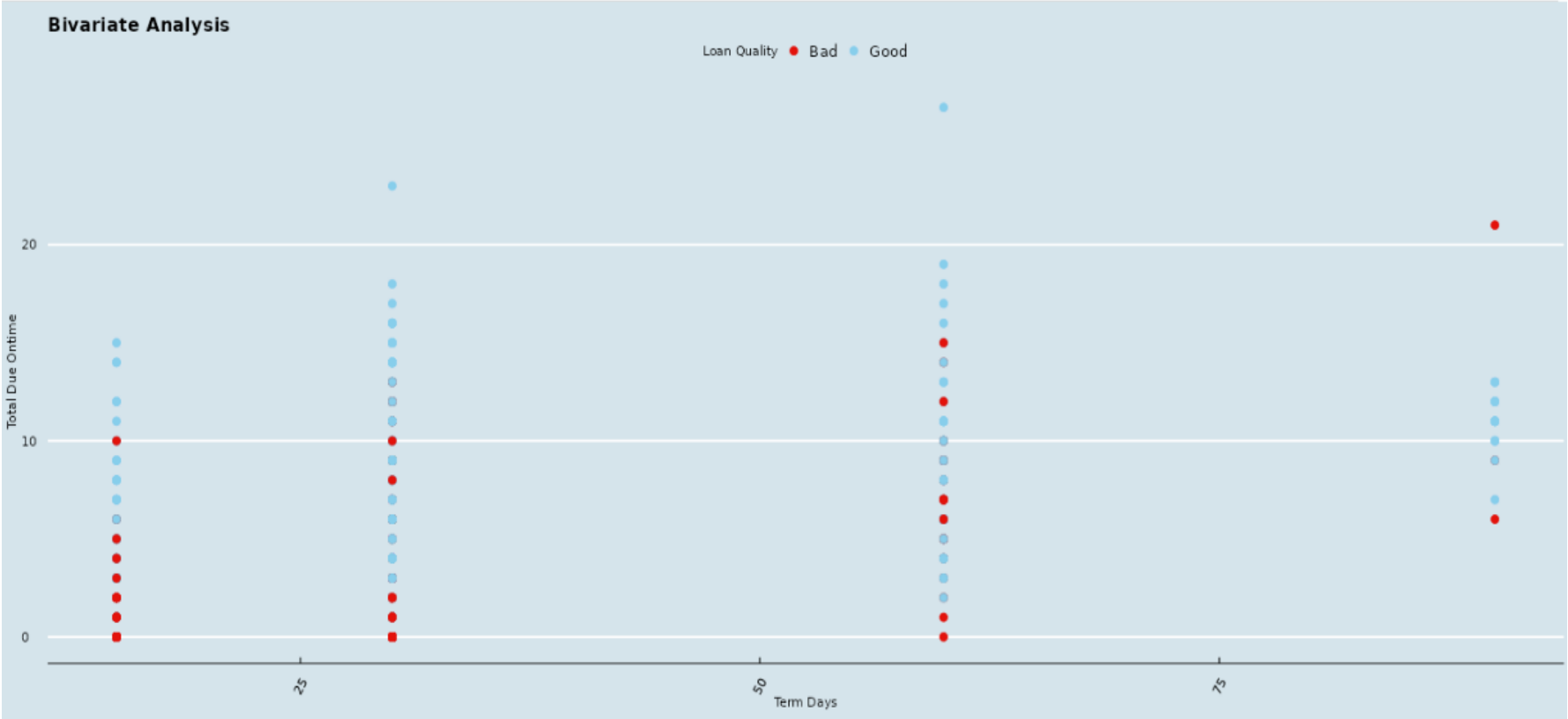
- ( ) Avg Age at Loan
- ( ) Bank Account Type
- ( ) Bank Name
- (●) Due Ontime Pctile
- ( ) Employment Status
- ( ) Term Days
- ( ) Total Due Ontime
- ( ) Max Active Loans
- ( ) Max Age at Loan
- ( ) Total no. of Loans
- ( ) Max approval Duration

Select variable Y

- ( ) Avg Age at Loan
- ( ) Bank Account Type
- ( ) Bank Name
- ( ) Due Ontime Pctile
- (●) Employment Status
- ( ) Term Days
- ( ) Total Due Ontime
- ( ) Max Active Loans
- ( ) Max Age at Loan
- ( ) Total no. of Loans
- ( ) Max approval Duration

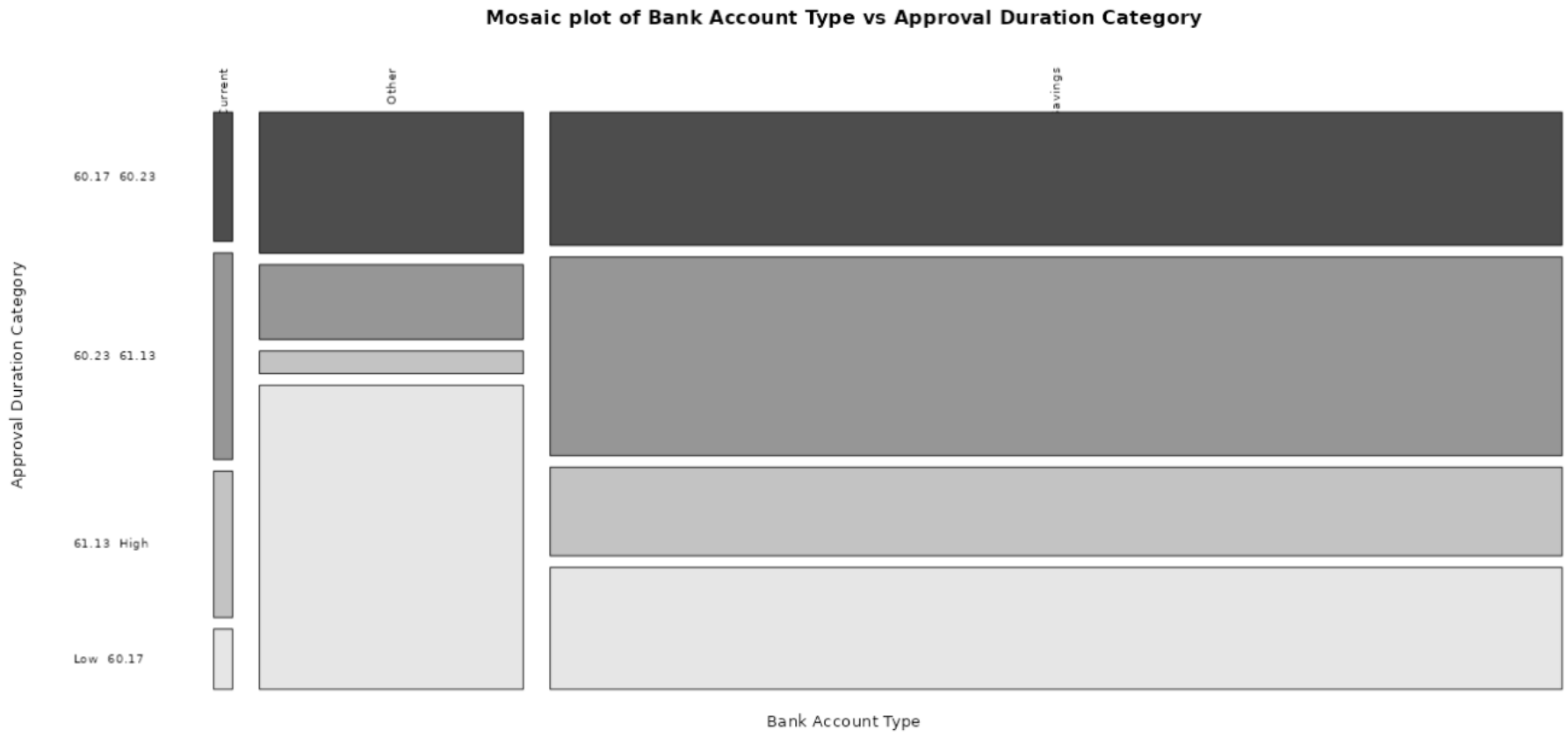
A bivariate chart would automatically pop out on the right panel. The plot could be either a box-plot, scatter plot or Mosaic plot, depending on whether the chosen variables are continuous or categorical. Because both chosen variables are continuous variables, a scatter plot would produce like above.

Continuous Variable vs. Continuous Variable



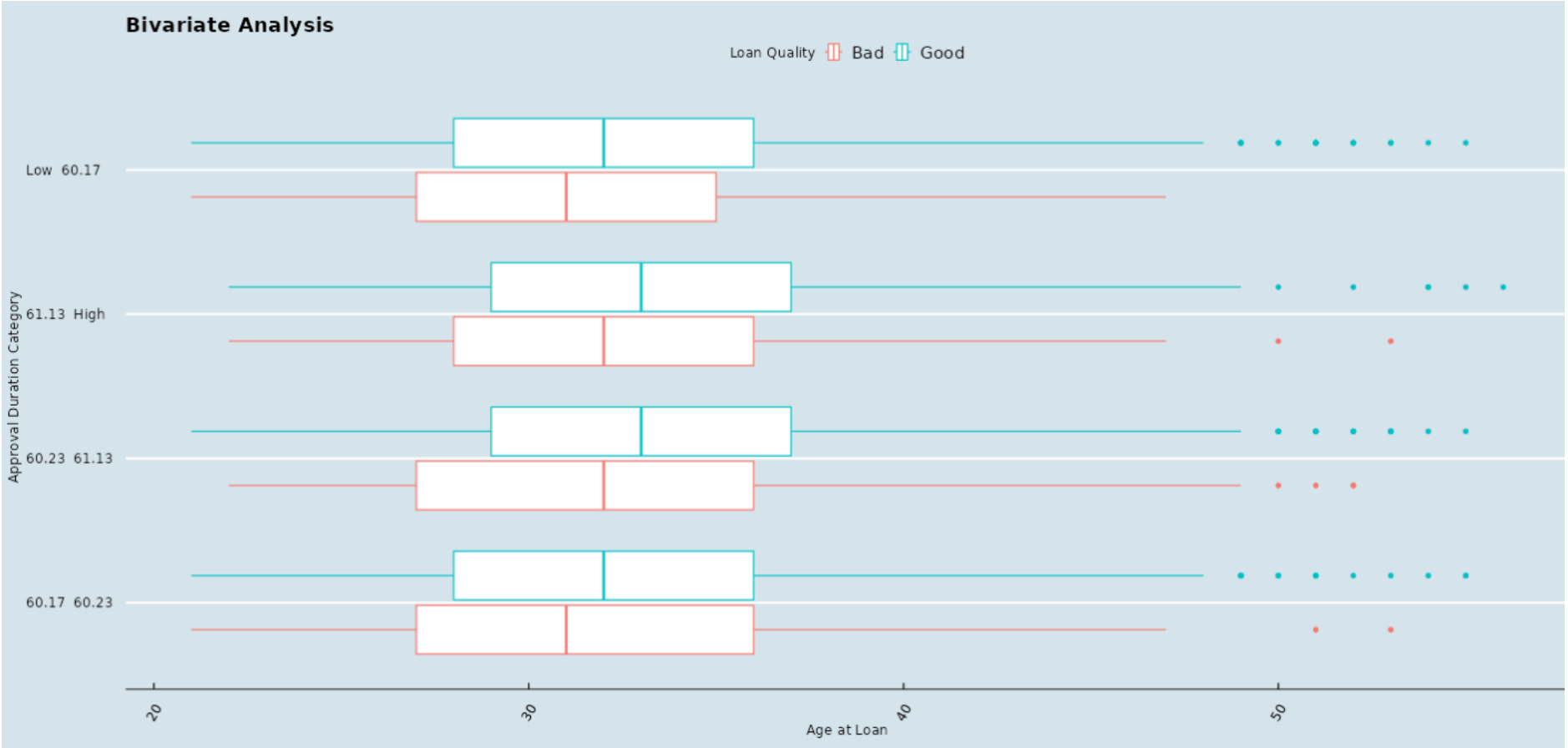
If two categorical variables are chosen, a mosaic plot that shows a representation view of certain group within the segment would appear as below.

Categorical Variable vs. Categorical Variable



Lastly, if one categorical variable and one continuous variable are chosen, a box plot with the continuous variable on the horizontal axis and the categorical variable on the vertical axis would appear as below.

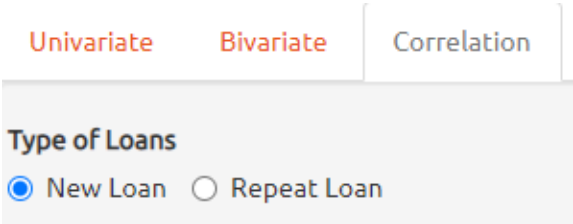
Categorical Variable vs. Continuous Variable



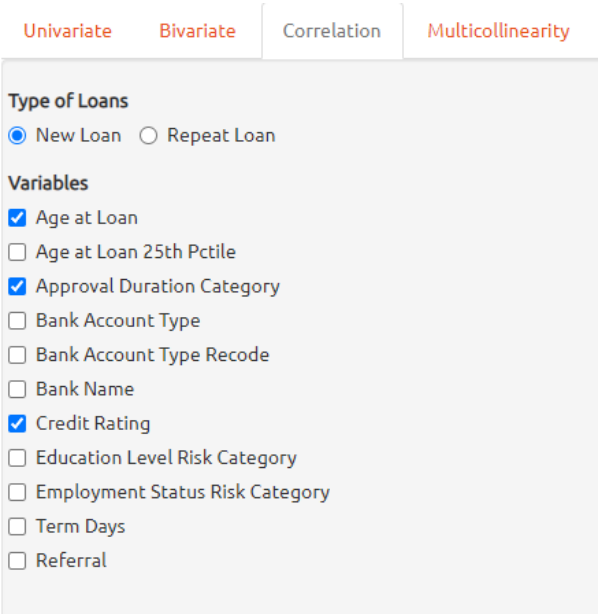
4. Correlation

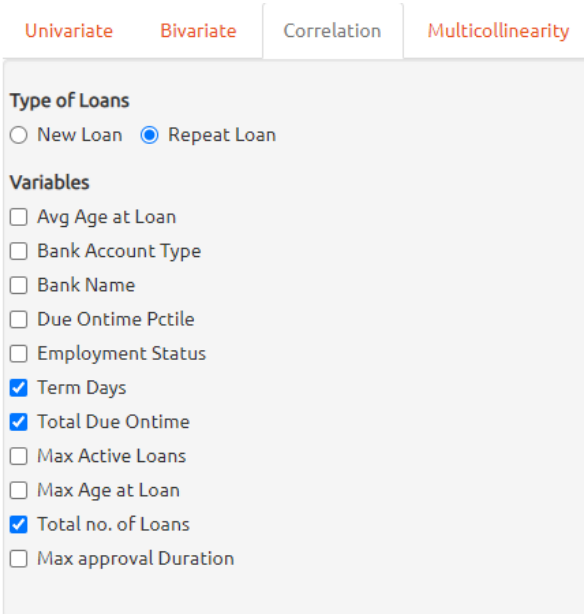
Step 1: Select which datasets to be used (New Loans datasets or Repeated Loans dataset).

Different type of loans would generate different variables to be studied for bivariate analysis later, as can been below.



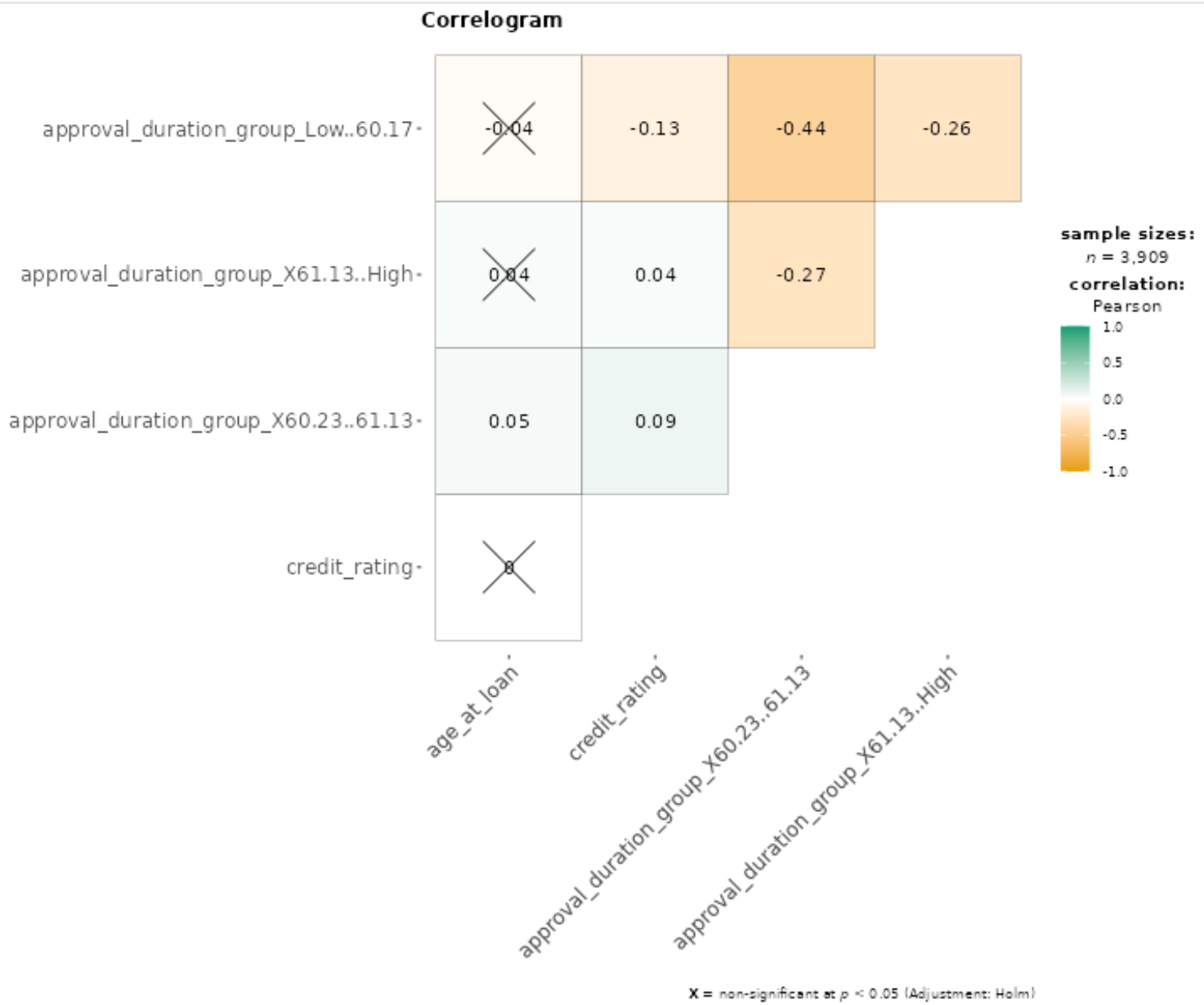
Step 2: Choose variables that you would like to analyze their correlations in between. Take note that at least two variables have to be chosen in order to display a proper correlation pairwise plot.





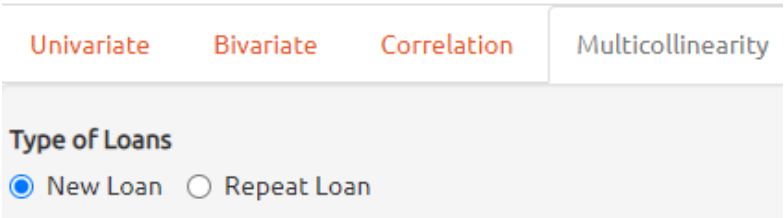
A color legend would show up on the right. Green color signifies positive correlation and orange color represents negative correlation while the brightness translates the degree of correlation into visual representation.

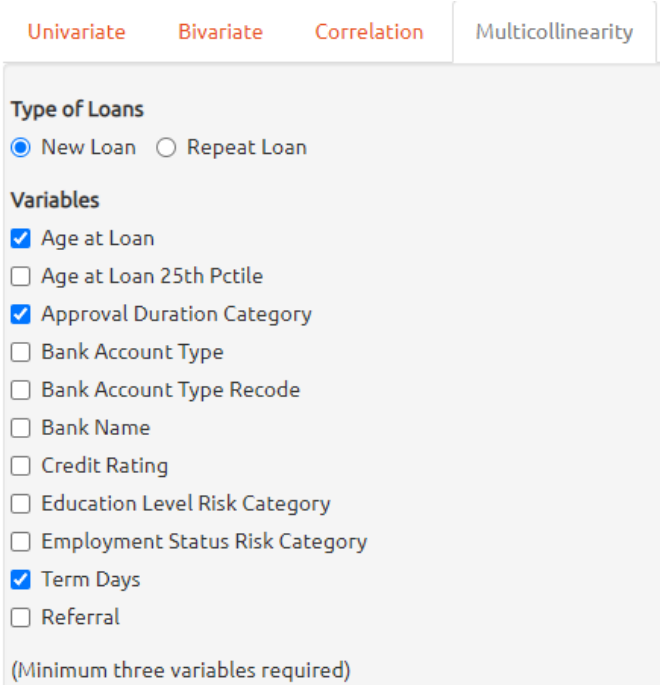
A significance test is also performed on all correlation between pairs of variables. A cross-mark would be displayed if the two paired variables produce a non-significant result.


Correlogram
approval_duration_group_Low..60.17
approval_duration_group_X61.13..High
approval_duration_group_X60.23..61.13
credit_rating
-0.04
-0.13
-0.44
-0.26
0.04
0.04
-0.27
0.05
0.09
0
age_at_loan
credit_rating
approval_duration_group_X60.23..61.13
approval_duration_group_X61.13..High
sample sizes:
n = 3,909
correlation:
Pearson
1.0
0.5
0.0
-0.5
-1.0
X = non-significant at p < 0.05 (Adjustment: Holm)

5. Multi Collinearity tab

Step 1: Select which datasets to be used (New Loans datasets or Repeated Loans dataset).

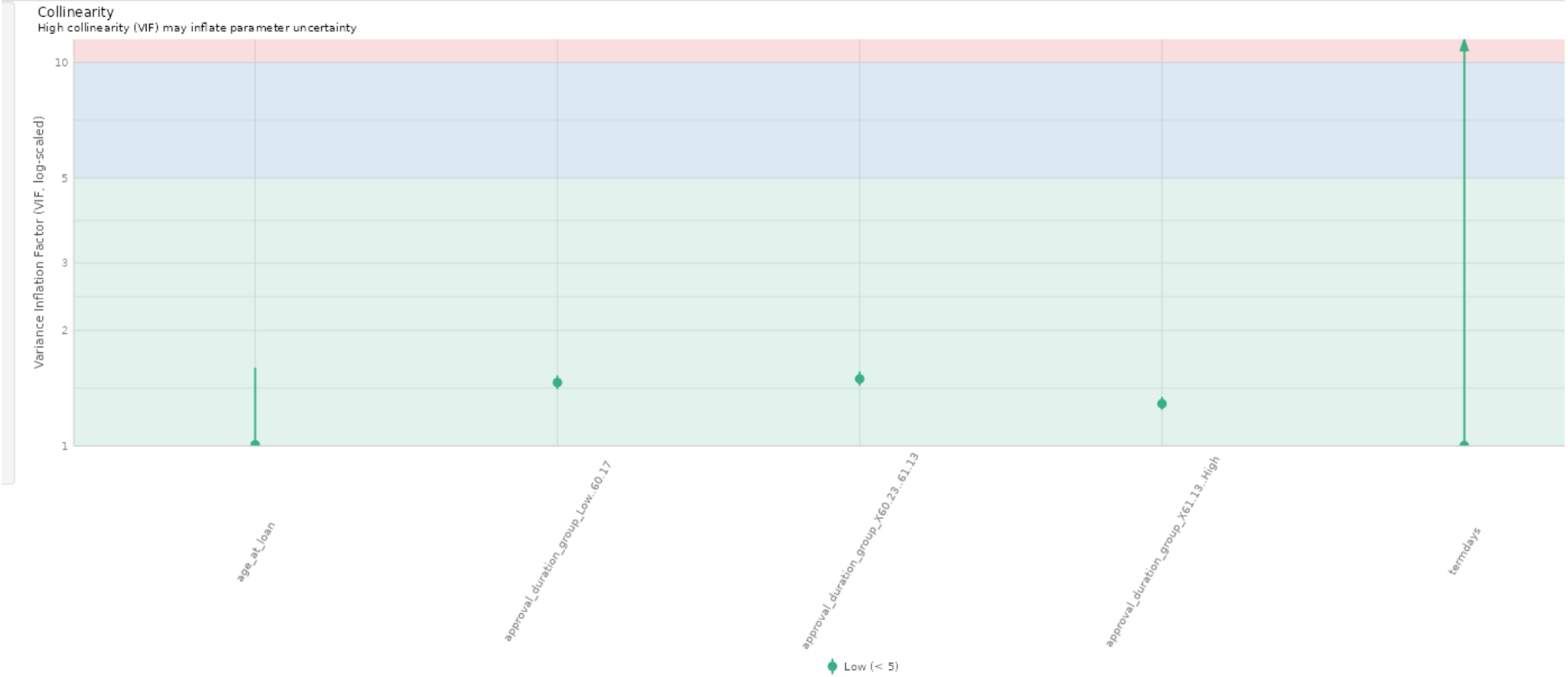




Step 2: Choose variables with which you would like to conduct a multicollinearity study. At least three variables must be chosen to display a proper multicollinearity study.

The multilinearity plot would automatically pop out inside the right panel. A table that comes along with multicollinearity would be produced at the bottom with values Variance Inflation Factor (VIF) as well as some of its associated statistical values like upper and lower confidence interval.

Quasi-Complete Separation Loan Default Prediction



Term	VIF	VIF_CI_low	VIF_CI_high	SE_factor	Tolerance	Tolerance_CI_low	Tolerance_CI_high
age_at_loan	1.007171	1.000086	1.598701	1.003579	0.9928803	6.255078e-01	0.9999141
termdays	1.001874	1.000000	35354.026076	1.000937	0.9981293	2.828532e-05	1.0000000
approval_duration_group_X60.23..61.13	1.495475	1.436953	1.561836	1.222896	0.6686837	6.402721e-01	0.6959170
approval_duration_group_X61.13..High	1.287996	1.242687	1.341764	1.134899	0.7764001	7.452877e-01	0.8047081
approval_duration_group_Low..60.17	1.463320	1.406809	1.527681	1.209678	0.6833776	6.545870e-01	0.7108285

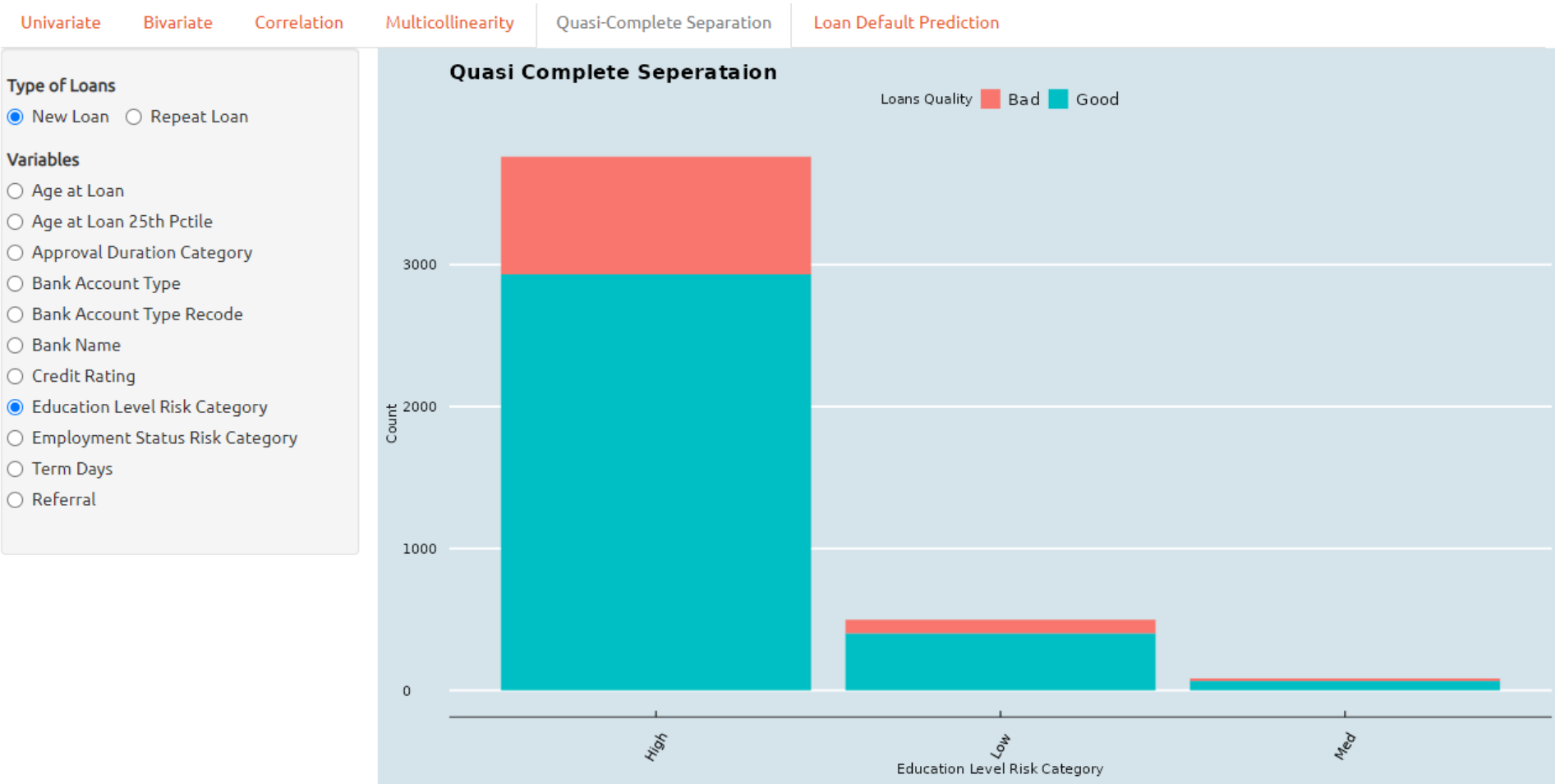
A VIF value equal to 1 represents that variables chosen are not correlated. For VIF values fall within 1 to 5 (green area), it means variables chosen are moderately correlated. For VIF values fall within 5 to 10 (blue area), it means variables chosen are highly correlated.

6. Quasi-Complete Separation Tab

Step 1: Select which datasets to be used (New Loans datasets or Repeated Loans datasets)

Step 2: Next, choose which variables to analyze to determine whether it violates the quasi-complete separation.

The result will show whether this variable would violate the quasi-complete separation issue. If it is, you will see one column/bar complete, including the majority (or total number) of one type of loan quality.



For example, in the above situation, the Variable "level of education risk" for the high level of education risk (for borrowers with lowest education levels), the majority of bad loans concentrated under this group, hence if the user includes this variable in the modeling, it may lead to overfitting and unreliable coefficient estimates.

7. Loan Default Prediction Tab

The loan default prediction function supports the following features for loan default classification.

1. Data Sampling with different sampling methods and parameter adjustments
2. Loan Prediction with different prediction algorithms and parameter adjustments

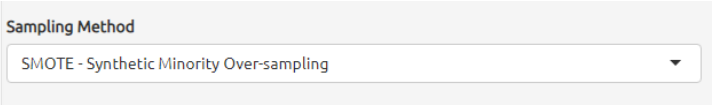
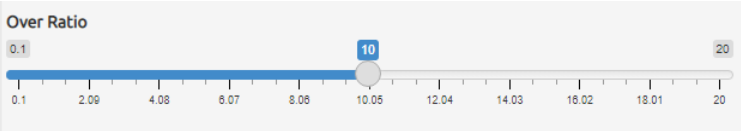
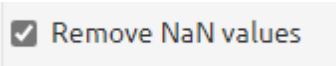
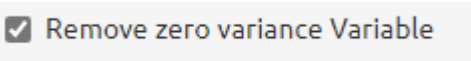
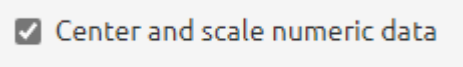
7.1. Data Sampling

7.1.1. Data Sampling Parameters

This function provides three different sampling methods to eliminate imbalanced data problems, allowing users to select the ratio of the new data size to be generated.

The table below details parameters that can be adjusted to tune the data sampling as part of the loan default prediction.

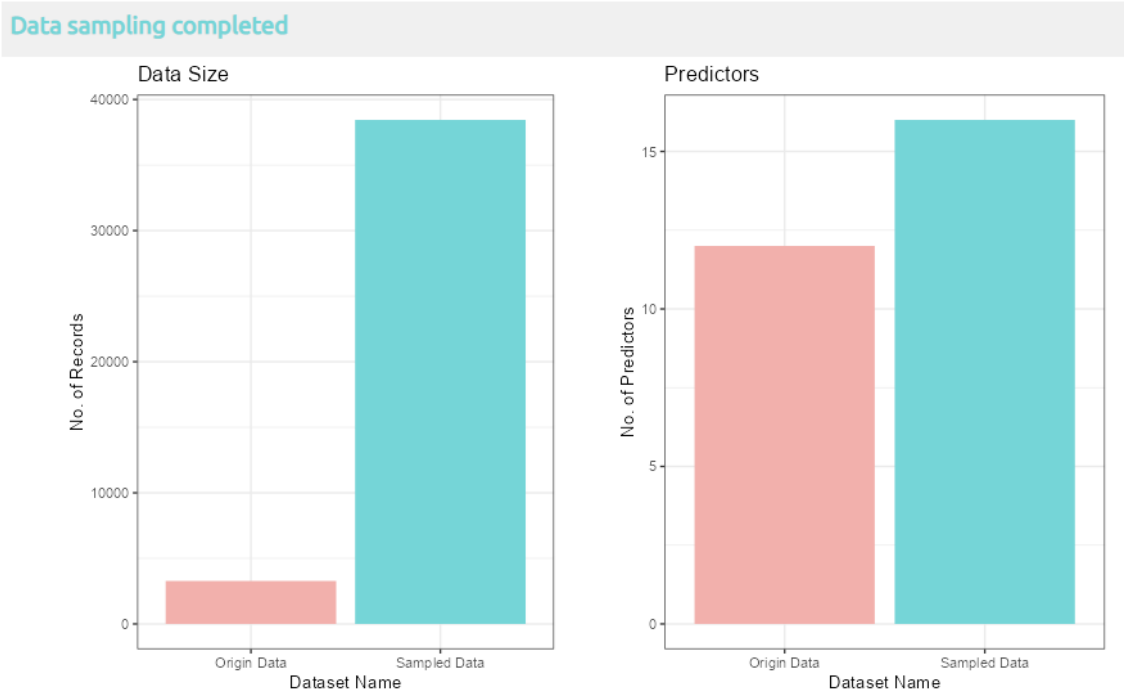
Data Sampling Parameters	Parameter Description
Type of Loans ○ New Loan ◉ Repeat Loan	Type of Loan customer

	Three sampling methods: - UP Sampling - Synthetic Minority Over-Sampling (SMOTE) - Random Over Sampling (ROSE)
	A numeric value for the ratio of the majority-to-minority frequencies. A value of 10 would mean that the minority levels will have (at most) (approximately) 10 times as many rows than the majority level.
	Data pre-processing of removing the NaN values before the oversampling.
	Data pre-processing of removing the zero variances variables before the over-sampling.
	Data pre-processing of normalizing (center and scale) numeric data before the over-sampling.

7.1.2. Performing Data Sampling

To perform the data sampling, click the "Start Sampling" button.

1. A successful message, "**Data sampling completed,**" is shown on the screen to indicate that the sampling process has been completed.
2. Two bar charts are rendered on the page.
   - 1st bar chart is plotted to compare the original data size vs. oversampled data size.
   - 2nd bar chart is plotted to compare the original predictors vs. predictors after data processing and sampling.



3. Two tables are also rendered on the page to provide more details for the predictors “before” and “after” sampling.

Variable
good_bad_flag
pct_ontime
total_ontime
max_active_of_loans
max_approval_duration
bank_name_clients
max_age_at_loan
avg_age_at_loan
employment_status
bank_account_type

Previous Next

Variable
pct_ontime
total_ontime
max_approval_duration
avg_age_at_loan
total_num_of_loans
termdays
bank_name_clients_Diamond.Bank
bank_name_clients_First.Bank
bank_name_clients_GT.Bank
bank_name_clients_UBA

Previous Next

Variable
total_num_of_loans
termdays

Previous Next

Variable
bank_name_clients_Zenith.Bank
employment_status_Permanent
employment_status_Self.Employed
employment_status_Unknown
bank_account_type_Savings
good_bad_flag

Previous Next

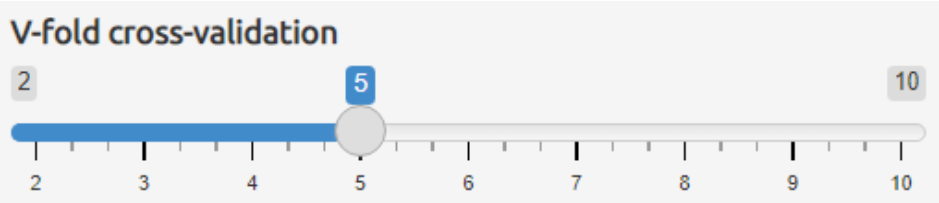
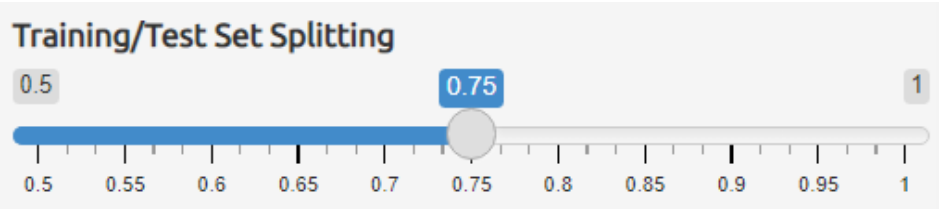
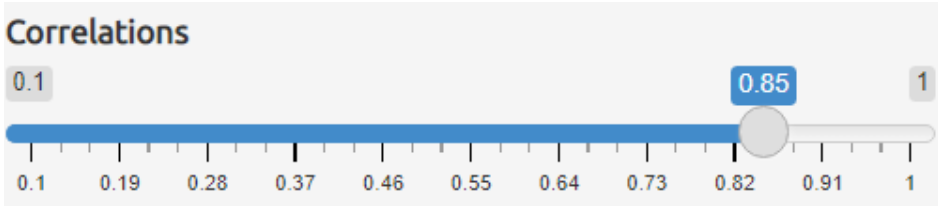
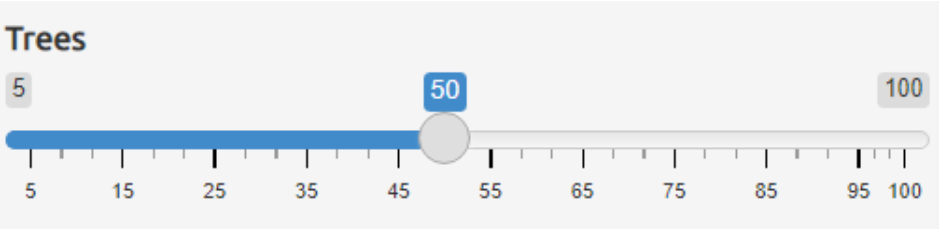
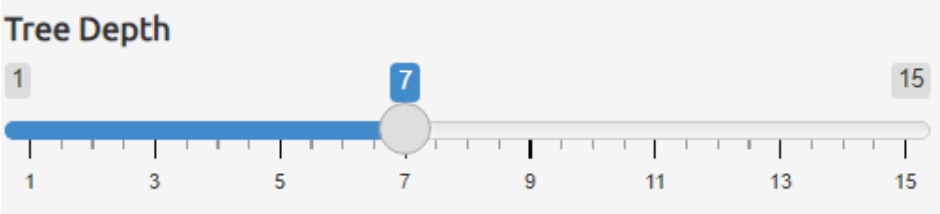
7.1.3. Loan Default Prediction

7.1.3.1. Prediction Parameters

This function provides three different R implementations of machine learning algorithms for loan default prediction.

This project selects V-Fold Cross-Validation (a.k.a k-fold cross-validation) from tidy models. It randomly splits the data into V groups of roughly equal size (called "folds"), and the cross-validation dataset is applied to resamples in model training.

Data Sampling Parameters	Parameter Description
Predicting Algorithm ☑ Boosted Tree ☐ Random Forest ☐ Logistic Regression	Three different loan prediction algorithms are provided in this project and allow multi-selection. - Boosted Tree - Random Forest - Logistic Regression
Variables Avg Age at Loan Due Ontime Pctile Max approval Duration	The details of all variables can be found from the introduction page on the main **website**.

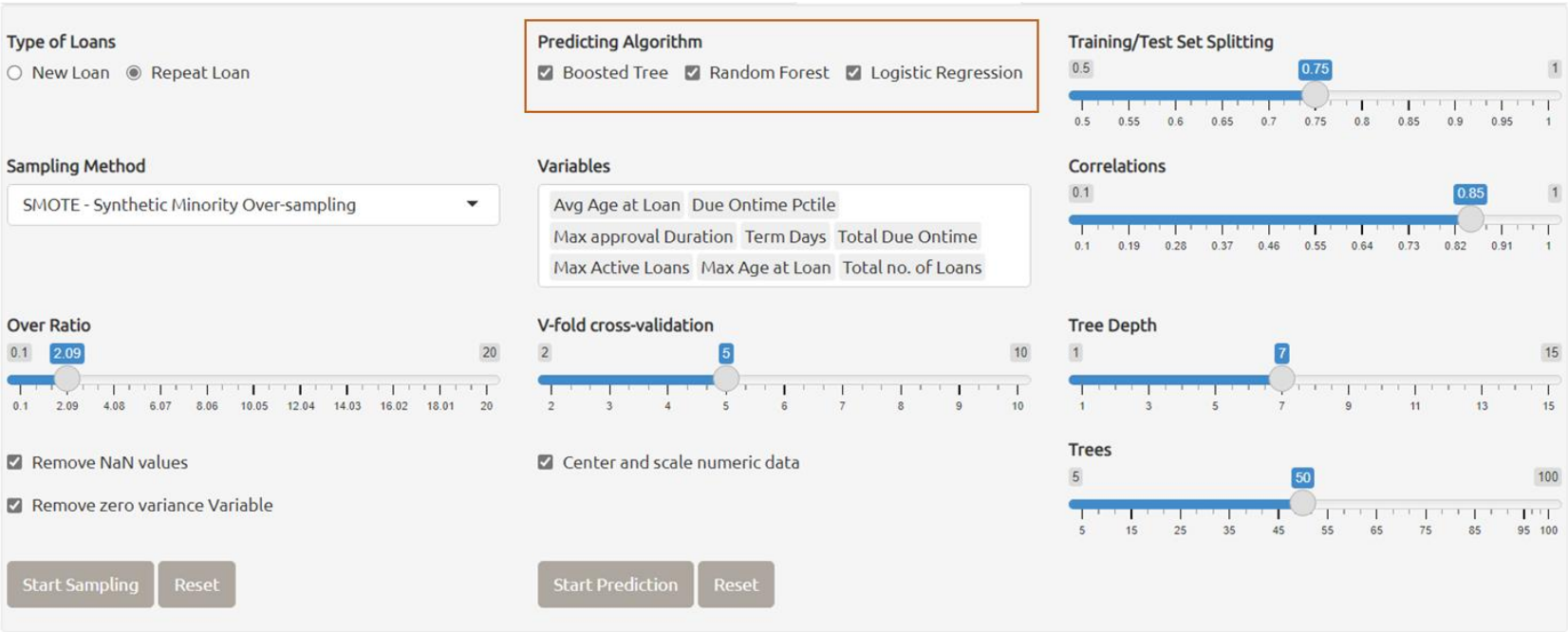
	Randomly splits the data into V groups of roughly equal size (called "folds")
	The size ratio of training dataset vs. testing dataset in the initial_split of rsample.
	Variables are to be removed by the recipe when the variable correlates larger than the screen value.
	Number of trees contained for rand_forest and Boosted Tree
	An integer for the maximum depth of the tree for Boosted trees

7.1.3.2. Performing Prediction

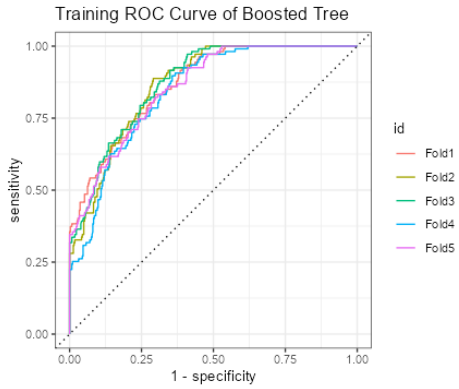
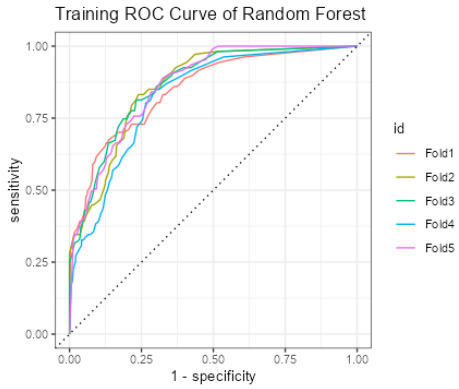
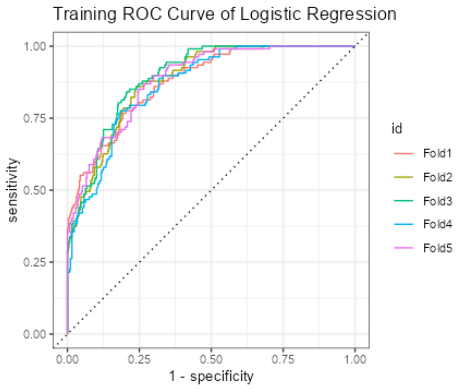
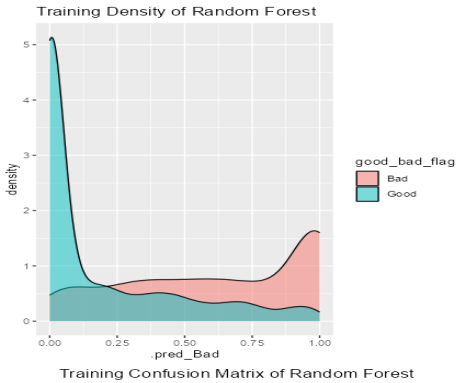
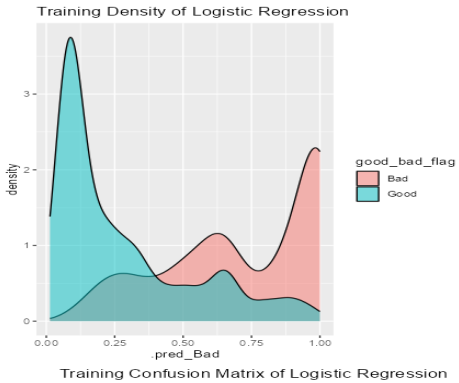
To perform loan default prediction, click “Start Prediction” after all parameters are identified and set properly on the UI.

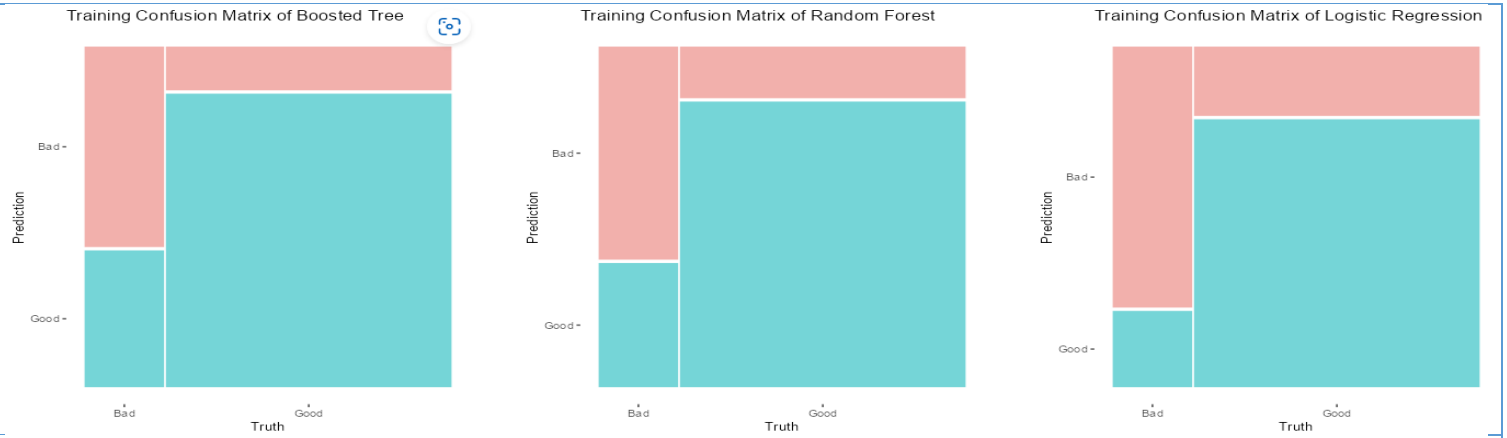
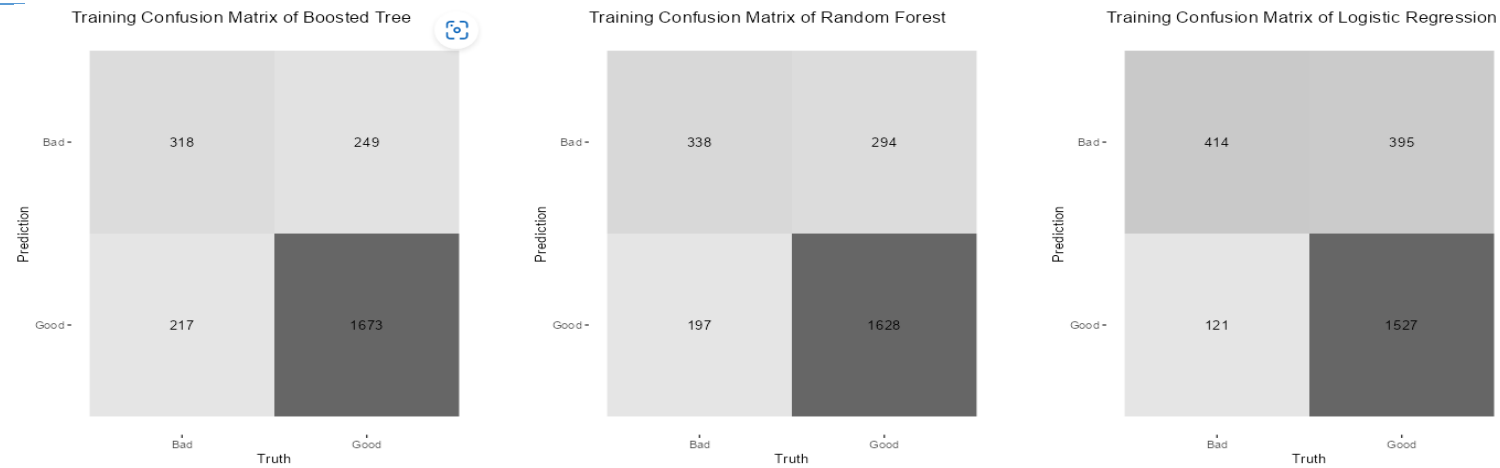
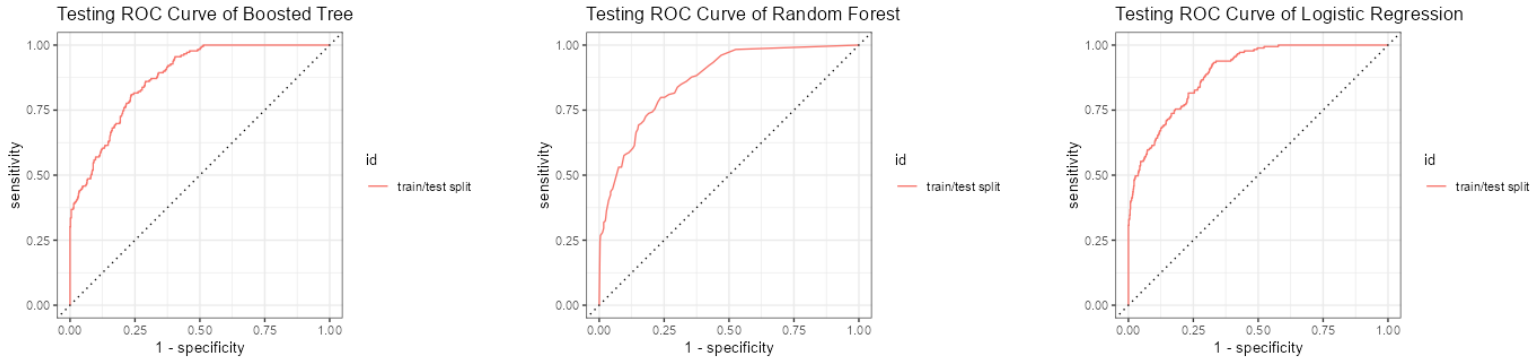
When the prediction is completed, a successful message Loan default prediction completed will be shown on the page.

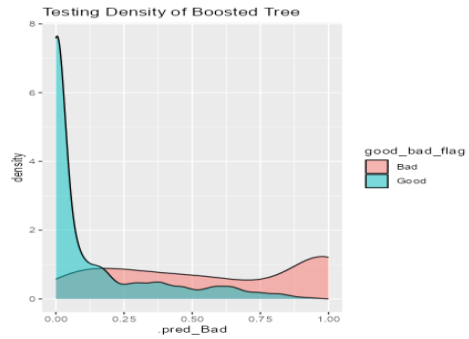
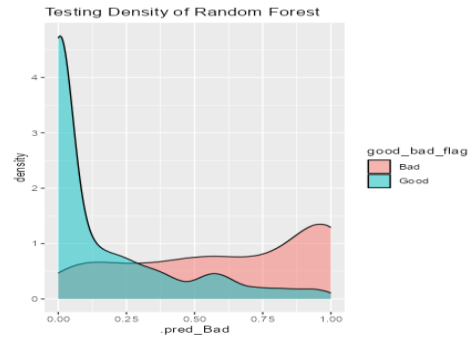
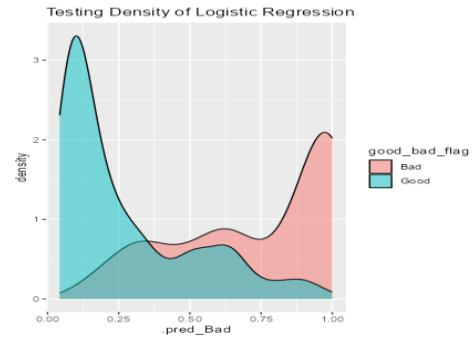
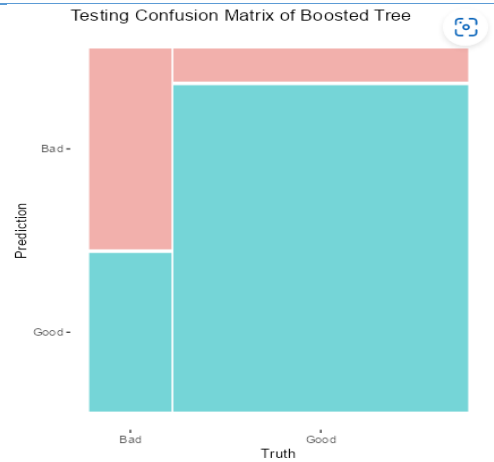
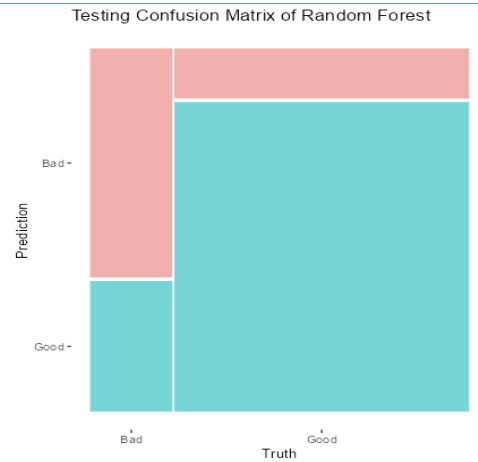
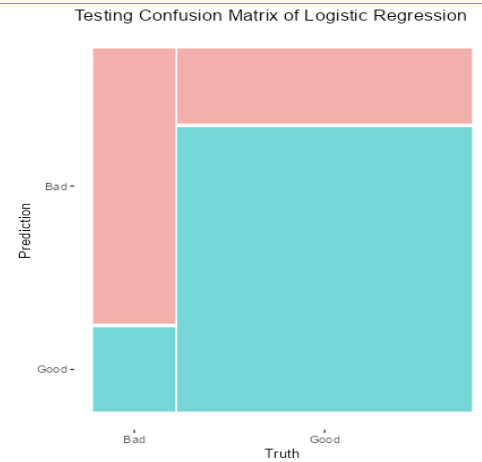
In this user guide, all three algorithms are selected for loan default prediction; the output results will show on three different columns for each prediction algorithm.



The outcome of the prediction is shown in the table below:

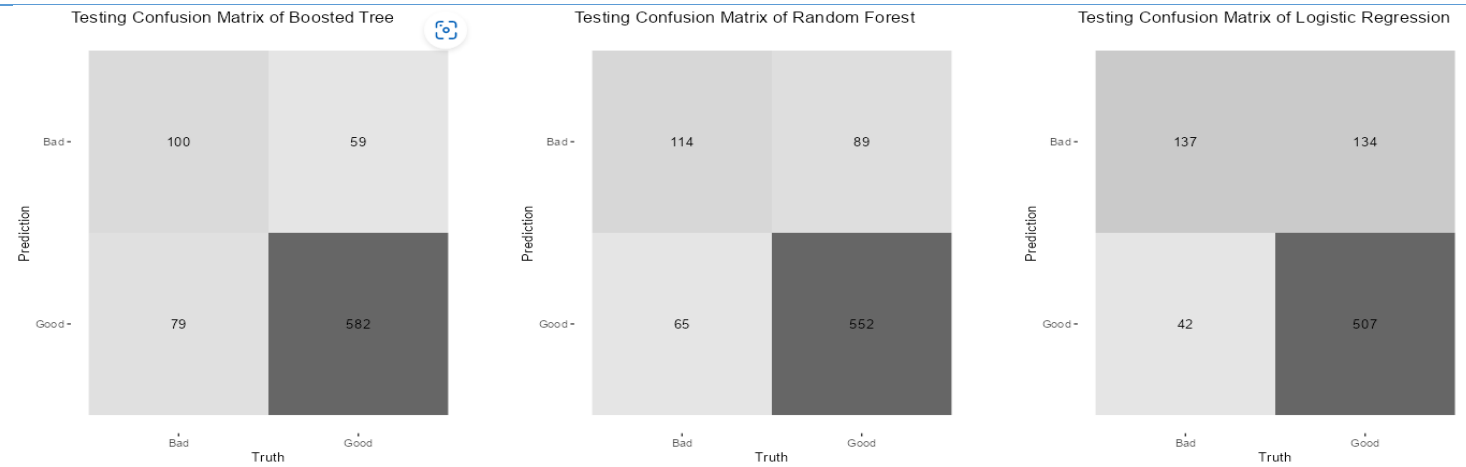
Visualization Name	Visualization Plot
ROC Curve of training data	Loan default prediction completed Boosted Tree Prediction  Random Forest Prediction  Logistic Regression Prediction 
The density of Good vs. Bad of training data (Logistic Regression has the best separation of bad loans from the dataset)	 Training Confusion Matrix of Boosted Tree  Training Confusion Matrix of Random Forest  Training Confusion Matrix of Logistic Regression

Mosaic Plot of Confusion Matrix of training data	
Heatmap Plot of Confusion Matrix of training data (Prediction accuracy parameters output are visualizable from the plot)	
ROC Curve of testing data	

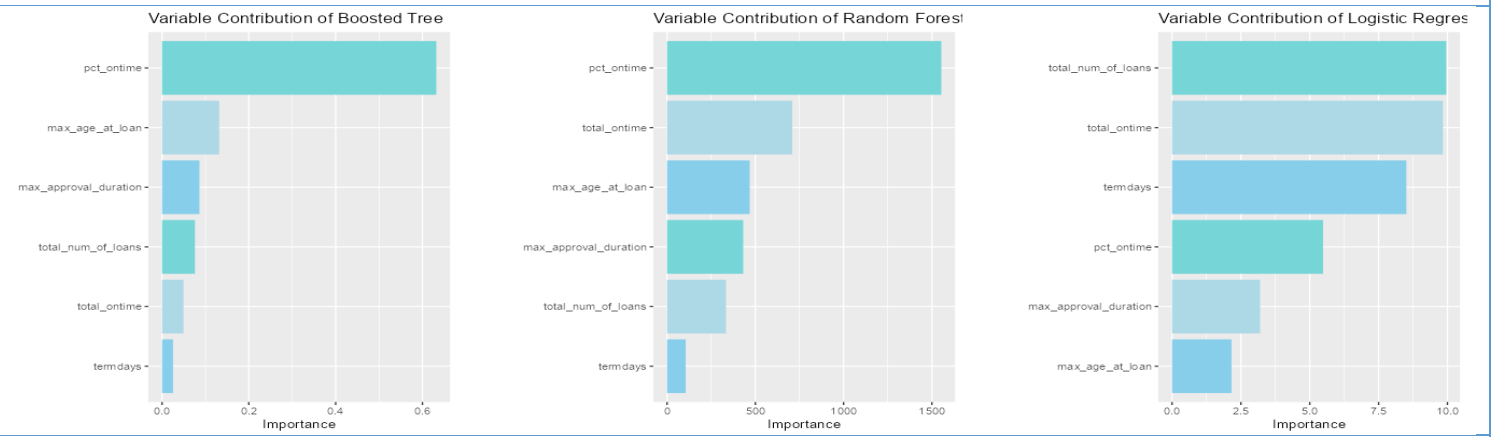
The density of Good vs. Bad of testing data (Logistic Regression has the best separation of bad loans from the dataset)	 Testing Confusion Matrix of Boosted Tree  Testing Confusion Matrix of Random Forest  Testing Confusion Matrix of Logistic Regression
Mosaic Plot of Confusion Matrix of testing data	 Testing Confusion Matrix of Boosted Tree  Testing Confusion Matrix of Random Forest  Testing Confusion Matrix of Logistic Regression

Heatmap Plot of Confusion Matrix of testing data

(Prediction accuracy parameters output are visualizable from the plot)



Predictor Contribution (VIP Chart)



Predictor Contribution (Summary Table)

Variable	Importance	Variable	Importance	Variable	Importance	Sig
pct_ontime	0.63189910	pct_ontime	1553.3398	total_num_of_loans	9.957707	NEG
max_age_at_loan	0.13161547	total_ontime	709.4010	total_ontime	9.836485	POS
max_approval_duration	0.08612703	max_age_at_loan	468.3609	termdays	8.507604	NEG
total_num_of_loans	0.07571182	max_approval_duration	431.4167	pct_ontime	5.482180	NEG
total_ontime	0.04938355	total_num_of_loans	333.5934	max_approval_duration	3.197470	POS
termdays	0.02526302	termdays	105.7097	max_age_at_loan	2.159325	POS

Previous Next Previous Next Previous Next

~ Thank you ~