BIL PROJECT

LOAN PREDICTION MODEL

NAME	CLASS	ROLL-NO.
SAHIL SHAH	TE-6	35
ABHAY RAJDE	TE-6	29
SALONI JAIN	TE-6	21
BHAVYA SHAH	TE-6	31

Guide
Prof. Ashwini Deshmukh

Department of Information Technology
Shah & Anchor Kutchhi Engineering College, Mumbai

1.Problem Statement :-

Company wants to automate the loan eligibility process (real time) based on customer detail provide while filling online application form. We have collected all the necessary details. We have to automate this process by identifying the customer segments and those who are eligible for loan amount so that they can specifically target these customers.

INTRODUCTION :-

The aim of our model is to predict loan eligibility of different categories of people. For model building we have used the “ORANGE” tool which is an open source data visualization and analysis tool, where data mining is done through visual programming or Python scripting. The tool has components for machine learning.

2.Dataset :-

In [1]:
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns #for plotting
```

In [4]:
```python
df = pd.read_csv("S:\\train_ctrUa4K.csv")
```

In [5]:
```python
df
```

Out[5]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_Histor
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.
...	...	...	...	...	...	...	...	...	...	...	.
609	LP002978	Female	No	0	Graduate	No	2900	0.0	71.0	360.0	1.
610	LP002979	Male	Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.
611	LP002983	Male	Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.
612	LP002984	Male	Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.
613	LP002990	Female	No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.

614 rows × 13 columns

In [4]:
```python
df = pd.read_csv("S:\\train_ctrUa4K.csv")
```

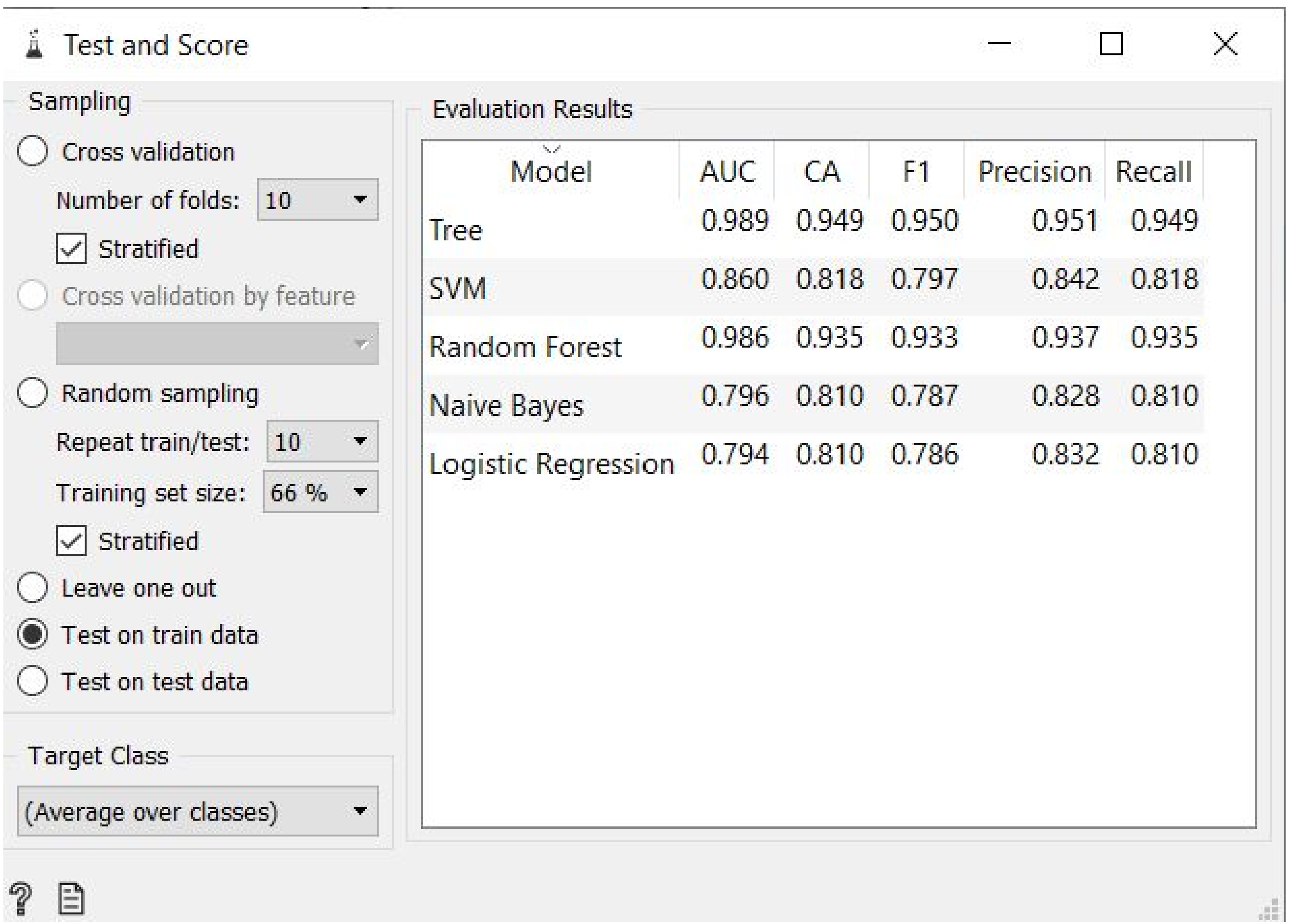
In [5]:
```python
df
```


Out[5]:

ied	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y
...	...	...	...	...	...	...	...	...	...	...
No	0	Graduate	No	2900	0.0	71.0	360.0	1.0	Rural	Y
Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.0	Rural	Y
Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.0	Urban	Y
Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.0	Urban	Y
No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.0	Semiurban	N

3.Algorithm :-

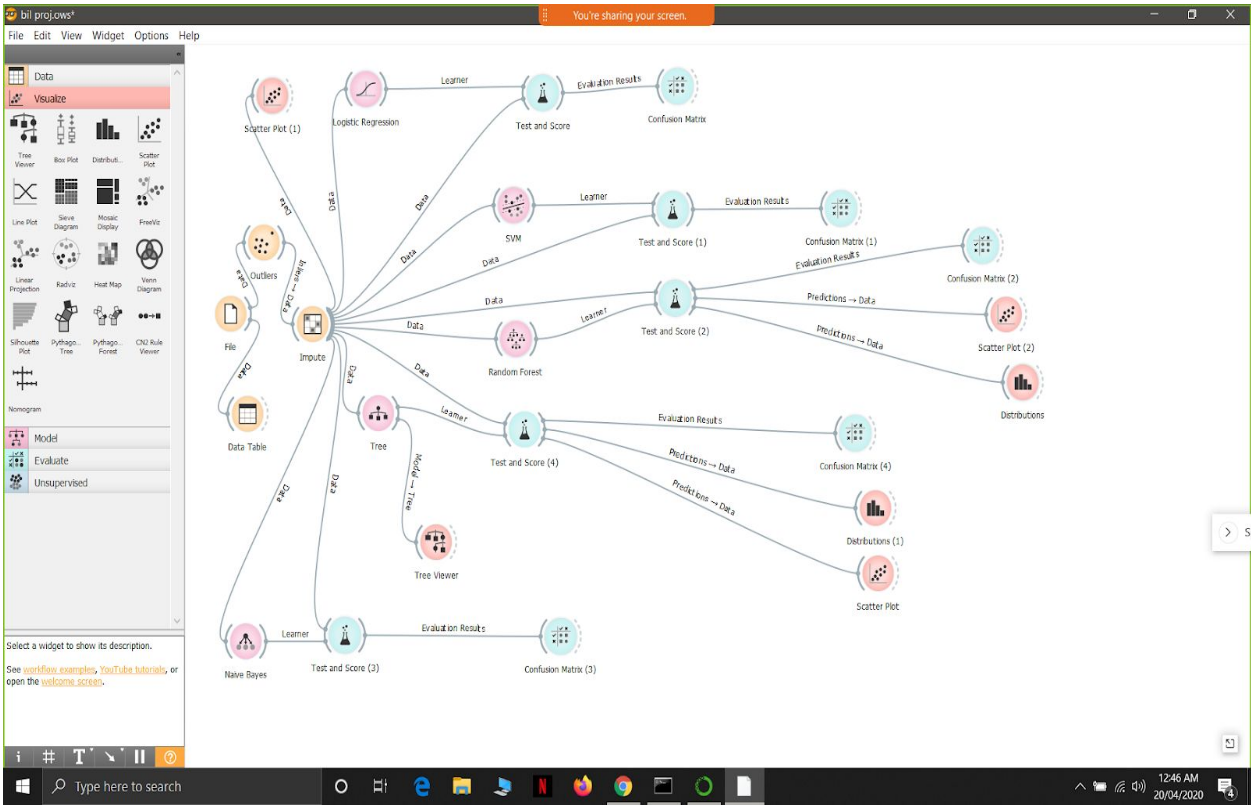
We have used 5 different classification algorithms which are Tree, Naive Bayes, Logistic Regression, Random Forest and SVM(Support Vector Machines). All the algorithms performed had different precision levels. Among them the highest classification accuracy was shown by Tree and Random Forest.

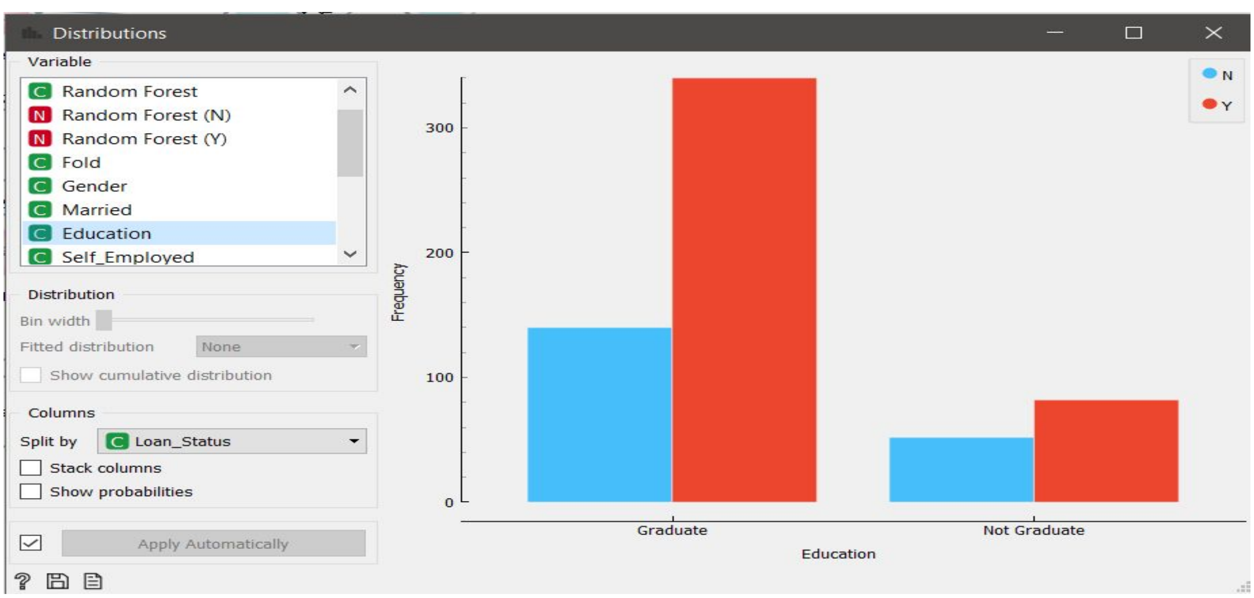


Model	AUC	CA	F1	Precision	Recall
Tree	0.989	0.949	0.950	0.951	0.949
SVM	0.860	0.818	0.797	0.842	0.818
Random Forest	0.986	0.935	0.933	0.937	0.935
Naive Bayes	0.796	0.810	0.787	0.828	0.810
Logistic Regression	0.794	0.810	0.786	0.832	0.810

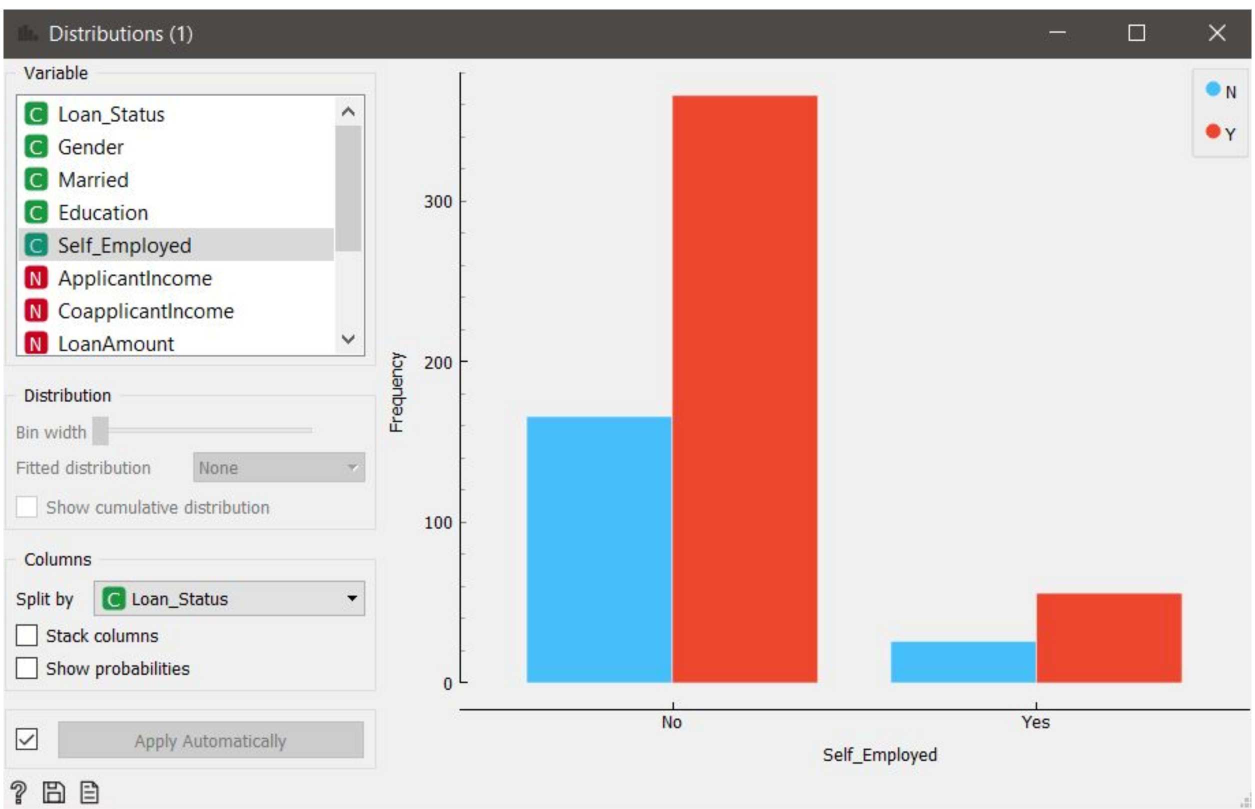
4. Visualization :-

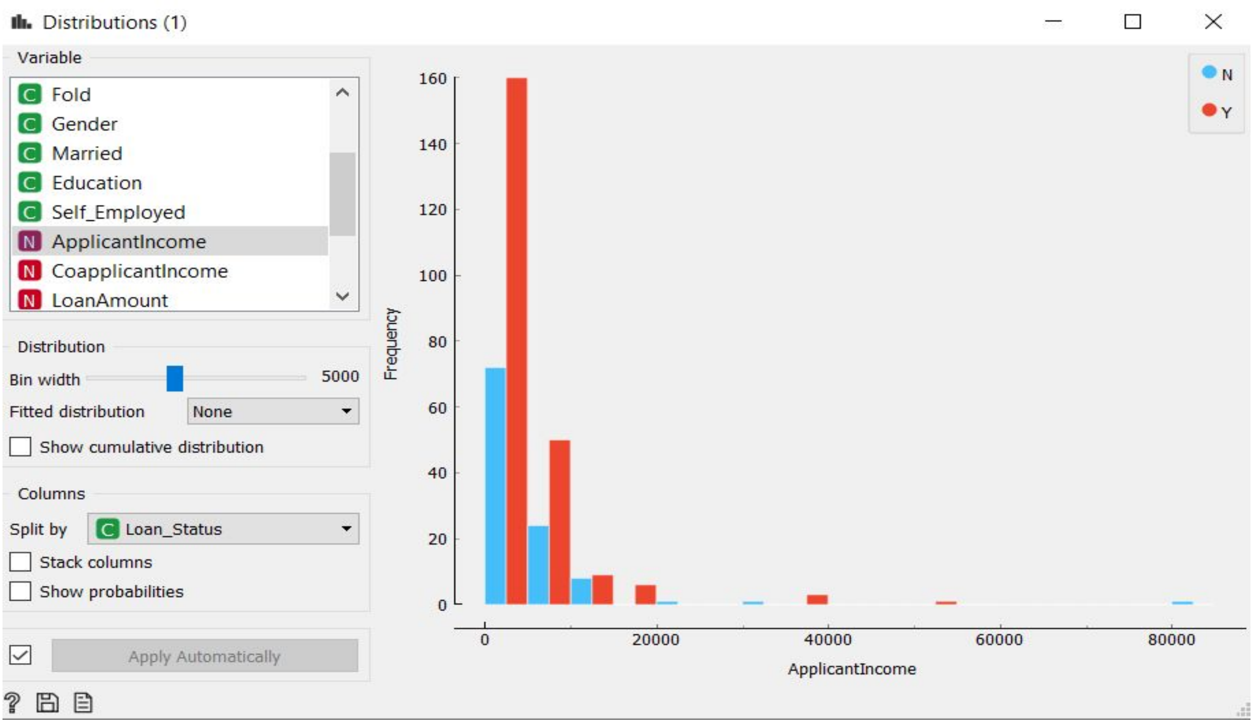
Model of loan prediction system

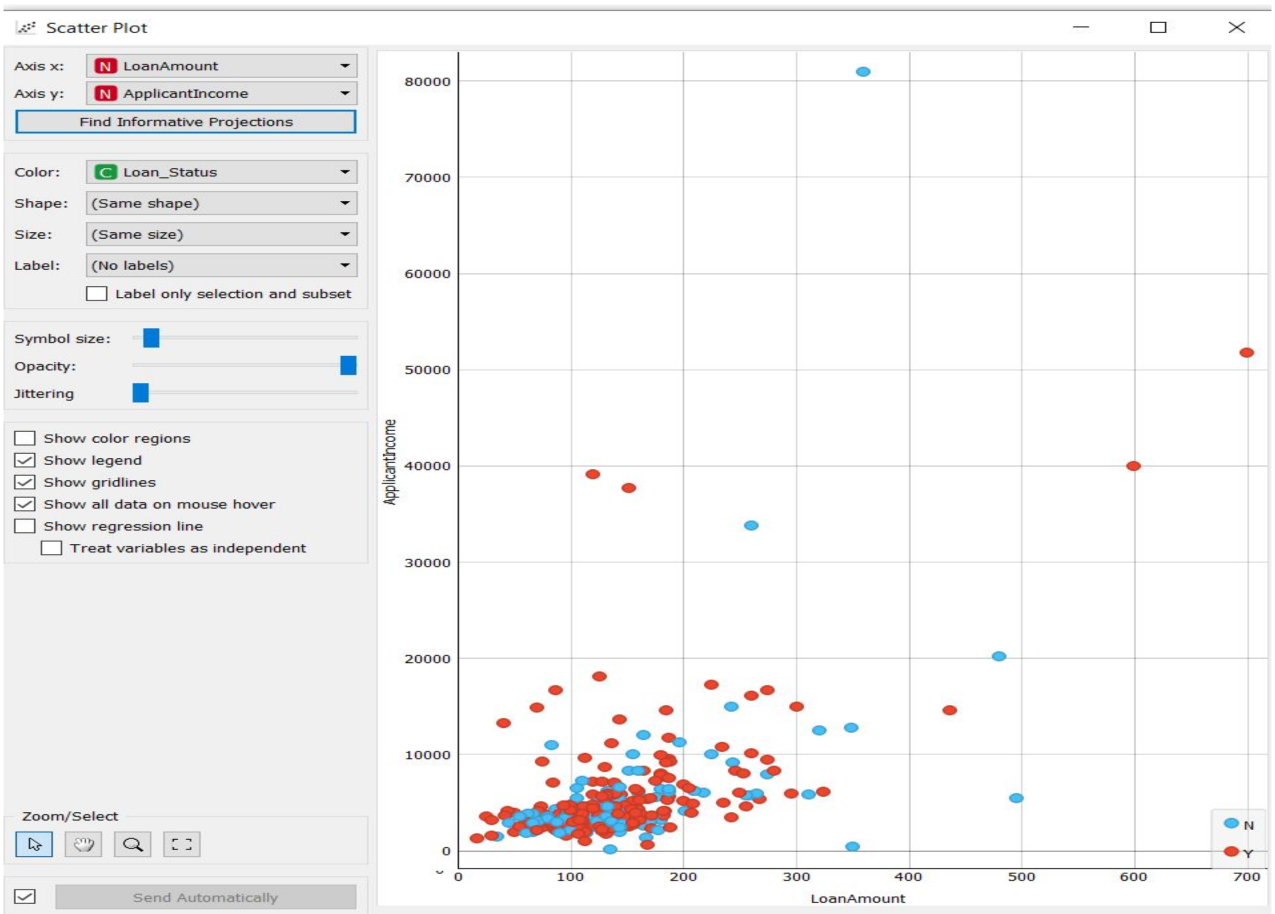


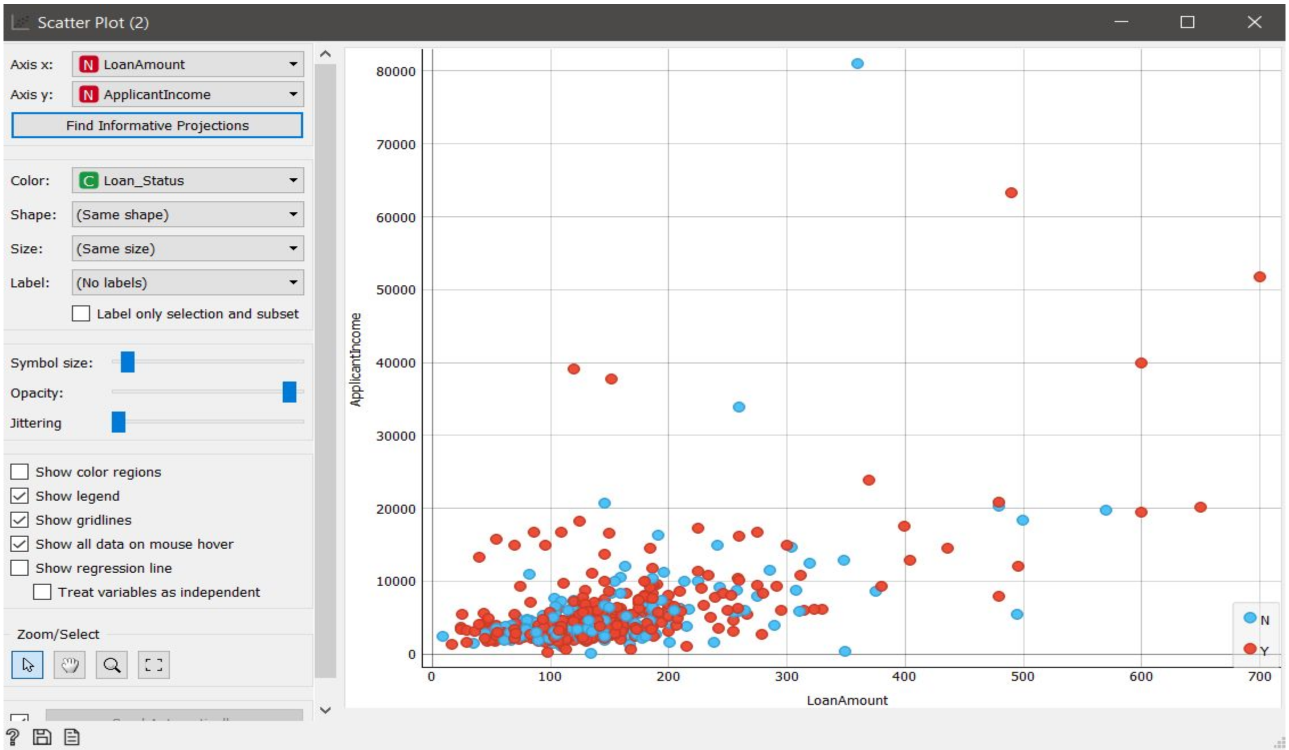


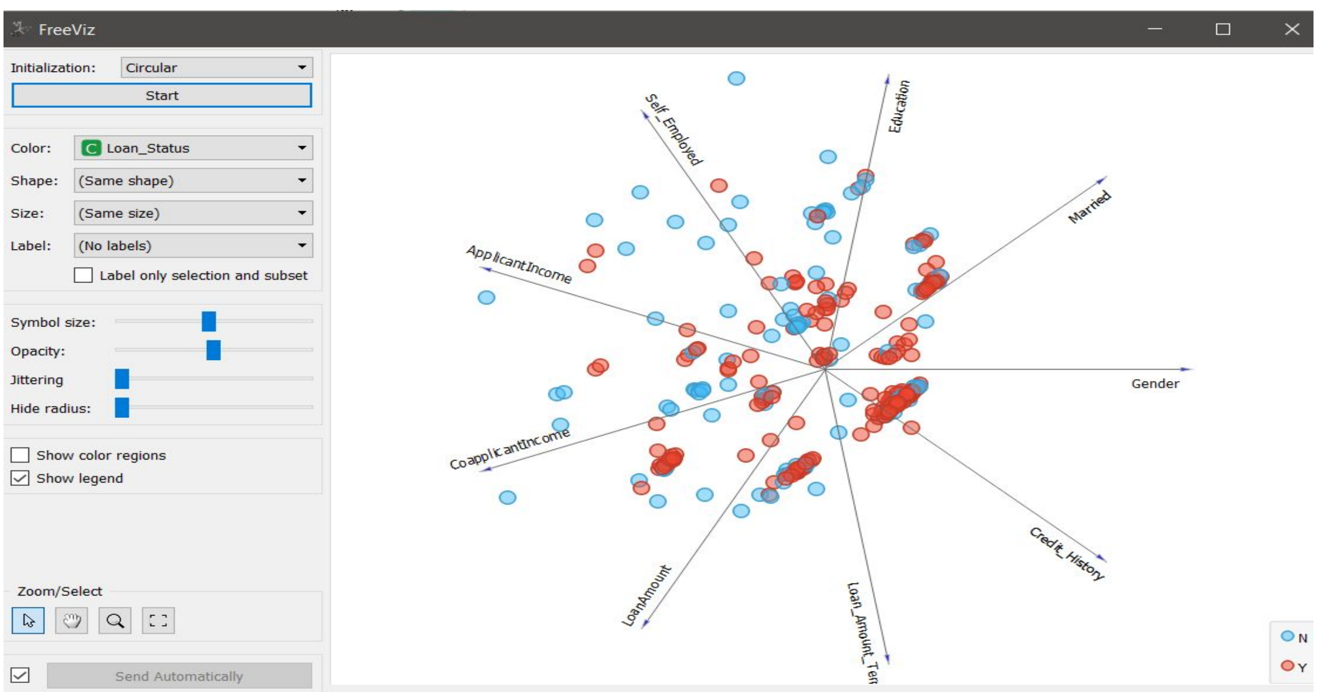

Distributions (1)
Variable
Loan_Status
Gender
Married
Education
Self_Employed
ApplicantIncome
CoapplicantIncome
LoanAmount
Distribution
Bin width
Fitted distribution
None
Show cumulative distribution
Columns
Split by
Loan_Status
Stack columns
Show probabilities
Apply Automatically
N
Y
Frequency
0
100
200
300
Female
Male
Gender


Distributions (1)
Variable
Loan_Status
Gender
Married
Education
Self_Employed
ApplicantIncome
CoapplicantIncome
LoanAmount
Distribution
Bin width
Fitted distribution
None
Show cumulative distribution
Columns
Split by
Loan_Status
Stack columns
Show probabilities
Apply Automatically
N
Y
Frequency
0
100
200
300
No
Yes
Self_Employed


Distributions (1)
Variable
Fold
Gender
Married
Education
Self_Employed
ApplicantIncome
CoapplicantIncome
LoanAmount
Distribution
Bin width
5000
Fitted distribution
None
Show cumulative distribution
Columns
Split by
Loan_Status
Stack columns
Show probabilities
Apply Automatically
Frequency
ApplicantIncome
N
Y


Scatter Plot
Axis x:
LoanAmount
Axis y:
ApplicantIncome
Find Informative Projections
Color:
Loan_Status
Shape:
(Same shape)
Size:
(Same size)
Label:
(No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering
Show color regions
Show legend
Show gridlines
Show all data on mouse hover
Show regression line
Treat variables as independent
Zoom/Select
Send Automatically
ApplicantIncome
LoanAmount
N
Y





After removing the outliers

5.Exploration Sheet

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Exploration Sheet of Loan Prediction												
2	Serial.No	Attribute	Type	Missing Values	Distinct Values	Unique Values	Min	Max	Mean	Std. Dev	Label	Count	Description
3	1	Loan_id	Nominal	0	614	614	-	-		-	LP001002	1	Unique identification number for loan
4	2	Gender	Nominal	13	2	0	-	-	-	-	Male	489	Gender of customers
5											Female	112	
6	3	Married	Nominal	3	2	0	-	-	-	-	NO	213	Maratial status of customers
7											YES	398	
8	4	Dependents	String	15	4	0	-	-	-	-	-	-	Dependents of customer
9	5	education	Nominal	0	2	0	-	-	-	-	Graduate	480	Qualification status of customer
10											Not Graduate	134	
11	6	Self_Employed	Nominal	32	2	0	-	-	-	-	yes	82	Employement Status
12											no	500	
13	7	ApplicantIncome	Numeric	0	505	445	150	81000	5403.459	6109.82	-	-	Income of Applicant
14	8	CoapplicantIncome	Numeric	0	287	247	-	41667	1621.246	2926.248	-	-	Income of Coapplicnat
15	9	LoanAmount	Numeric	22	203	93	9	700	146.412	85.587	-	-	Amount of loan
16	10	Loan_Amount_Term	Numeric	14	10	1	12	480	342	65.12	-	-	Term of Loan Amount
17	11	Credit_History	Numeric	50	2	0	0	1	0.842	0.365	-	-	Customer credit history
19	12	Property_Area	Nominal	0	3	0	-	-	-	-	Urban	202	Area where property is located
20											Rural	179	
21											Semiurban	233	
22	13	Loan_Status	nominal	0	2	0	-	-	-	-	Y	422	Status of loan,i.e it is approved or not
23											N	192	

6. BI Decision & Inference

From the above graphs and histograms we can interpret and understand that ApplicantsIncome plays an important role in determining whether the person is eligible for loan or not and also helps the company to target suitable customers for the loan. After applicantsincome, employement status plays an important role in determing the loan status. Thus, we can analyze it through the screenshots mentioned in visualization above.

Hence comparing to other attributes ApplicantsIncome and Employement status play an important role in decision factor of Loan Prediction Model.