MINI PROJECT REPORT

For the subject
Data Warehousing And Data Mining (trimester VIII)

LOAN PREDICTOR USING ORANGE TOOL

Submitted by:

Aishwarya Kulkarni PG32
1032180606

Lakshit Jain PG33
1032180611

Harshit Jain PG35
1032180629

Dishi Jain PG37
1032180725

Sharayu Guhe PG47
1032181226

Under the guidance of
Prof. Vaishali Suryawanshi

School of Computer Engineering And Technology
MIT World Peace University, Kothrud,
Pune 411038, Maharashtra - India

2020-2021

ABSTRACT

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this project we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this project is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing. In this paper we are predict the loan data by using some machine learning algorithms they are classification, logic regression, Decision Tree and gradient boosting.

INTRODUCTION

This Problem is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to a particular person will be safe or not. We have implemented this loan prediction problem using Decision tree algorithm and data cleaning in Orange as there are missing values in the dataset. We use map function for the missing values. The aim of this paper is to apply machine learning technique on dataset which has 1000 cases and 7 numerical and 6 categorical attributes. The creditability of a customer for sanctioning loan depend on several parameters, such as credit history, Instalment etc.

MOTIVATION

Loans default will cause huge loss for the banks, so they pay much attention on this issue and apply various method to detect and predict default behaviours of their customers. The loan is one of the most important products of the banking. All the banks are trying to figure out effective business strategies to persuade customers to apply their loans. However, there are some customers behave negatively after their application are approved. To prevent this situation, banks have to find some methods to predict customers' behaviours. Machine learning algorithms have a pretty good performance on this purpose, which are widely-used by the banking

PROBLEM DEFINITION

The loan is one of the most important products of the banking. All the banks are trying to figure out effective business strategies to persuade customers to apply their loans. However, there are some customers behave negatively after their application are approved. To prevent this situation, banks have to find some methods to predict customers' behaviours. Machine learning algorithms have a pretty good performance on this purpose, which are widely-used by the banking

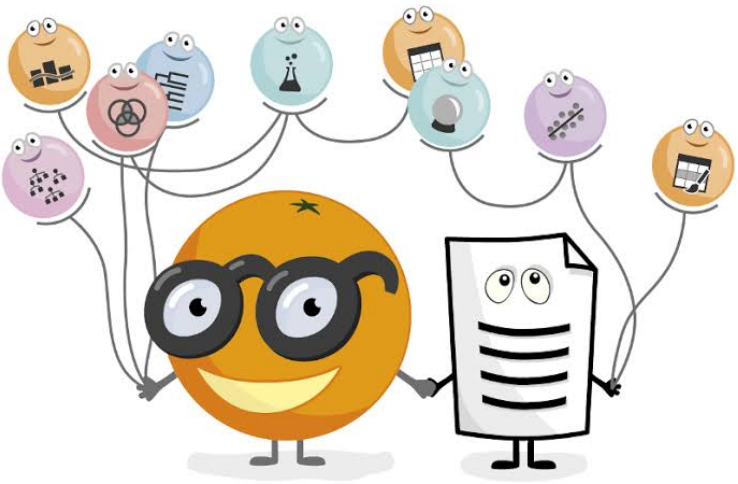
OBJECTIVES

Loan Prediction is very helpful for employee of banks as well as for the applicant also.
The aim of this:

- Paper is to provide quick, immediate and easy way to choose the deserving applicants. It can provide special advantages to the bank.
- The Loan Prediction System can can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight.
- A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not.
- Loan Prediction System allows jumping to specific application so that it can be check on priority basis.
- This Paper is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing.
- Result against particular Loan Id can be send to various department of banks so that they can take appropriate action on application.
- This helps all others department to carried out other formalities.

TOOLS USED

ORANGE3

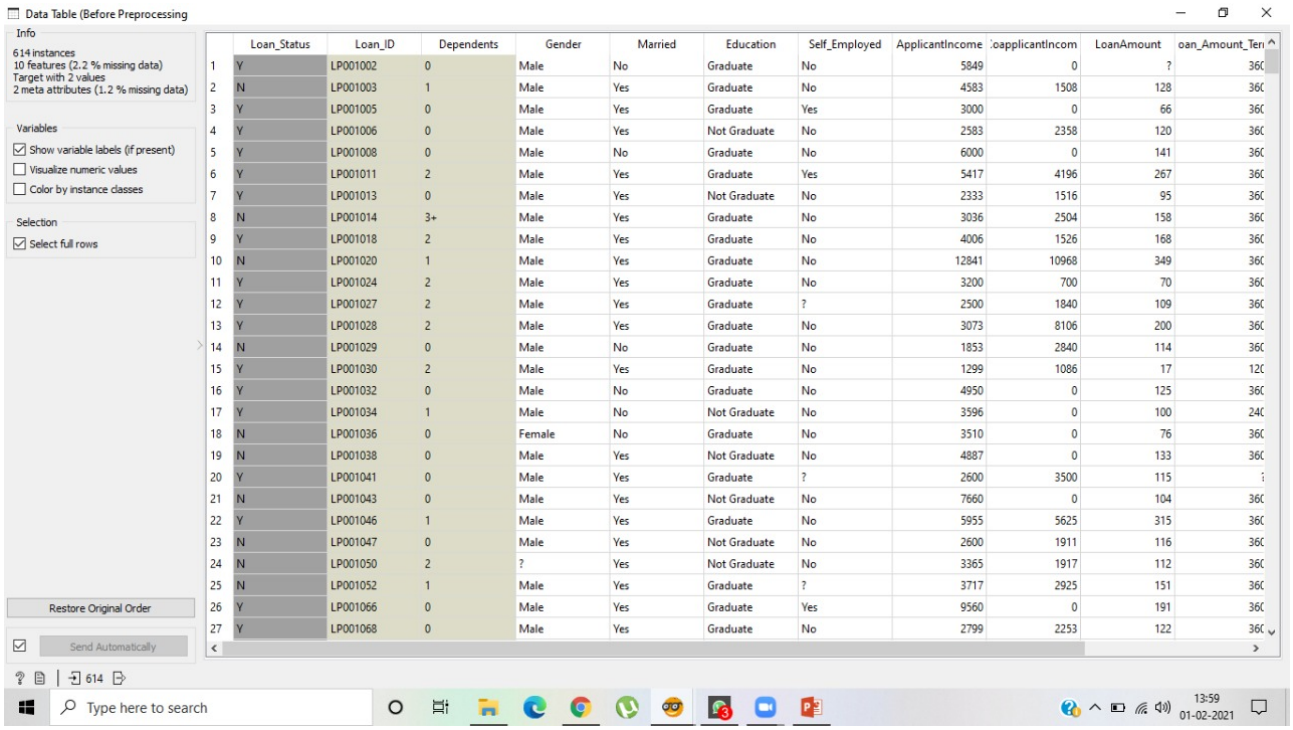


DATASET DESCRIPTION

Data Dictionary

Train file: CSV containing the customers for whom loan eligibility is known as 'Loan_Status'

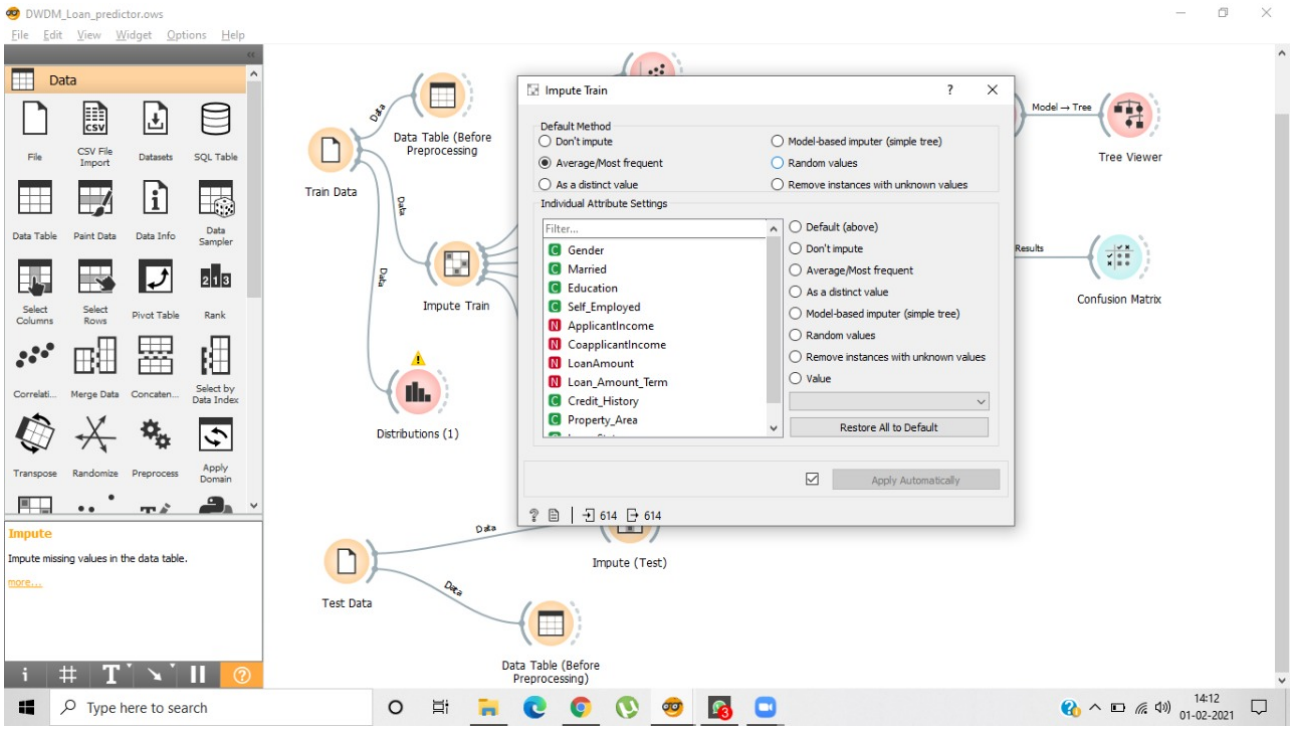
Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	(Target) Loan approved (Y/N)



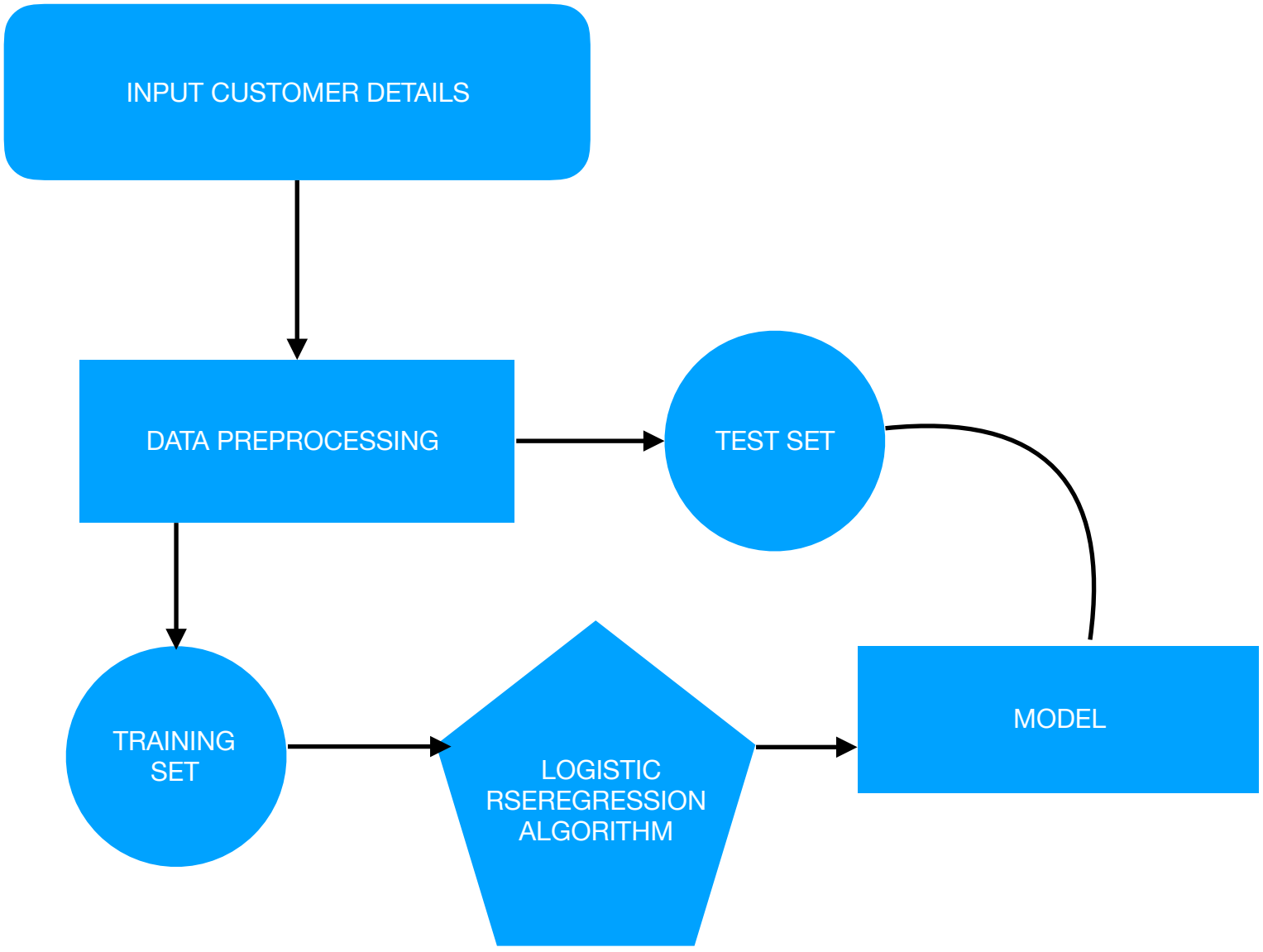
	Loan_Status	Loan_ID	Dependents	Gender	Married	Education	Self_Employed	ApplicantIncome	:oapplicantIncom	LoanAmount	oan_Amount_Ter
1	Y	LP001002	0	Male	No	Graduate	No	5849	0	?	360
2	N	LP001003	1	Male	Yes	Graduate	No	4583	1508	128	360
3	Y	LP001005	0	Male	Yes	Graduate	Yes	3000	0	66	360
4	Y	LP001006	0	Male	Yes	Not Graduate	No	2583	2358	120	360
5	Y	LP001008	0	Male	No	Graduate	No	6000	0	141	360
6	Y	LP001011	2	Male	Yes	Graduate	Yes	5417	4196	267	360
7	Y	LP001013	0	Male	Yes	Not Graduate	No	2333	1516	95	360
8	N	LP001014	3+	Male	Yes	Graduate	No	3036	2504	158	360
9	Y	LP001018	2	Male	Yes	Graduate	No	4006	1526	168	360
10	N	LP001020	1	Male	Yes	Graduate	No	12841	10968	349	360
11	Y	LP001024	2	Male	Yes	Graduate	No	3200	700	70	360
12	Y	LP001027	2	Male	Yes	Graduate	?	2500	1840	109	360
13	Y	LP001028	2	Male	Yes	Graduate	No	3073	8106	200	360
14	N	LP001029	0	Male	No	Graduate	No	1853	2840	114	360
15	Y	LP001030	2	Male	Yes	Graduate	No	1299	1086	17	120
16	Y	LP001032	0	Male	No	Graduate	No	4950	0	125	360
17	Y	LP001034	1	Male	No	Not Graduate	No	3596	0	100	240
18	N	LP001036	0	Female	No	Graduate	No	3510	0	76	360
19	N	LP001038	0	Male	Yes	Not Graduate	No	4887	0	133	360
20	Y	LP001041	0	Male	Yes	Graduate	?	2600	3500	115	?
21	N	LP001043	0	Male	Yes	Not Graduate	No	7660	0	104	360
22	Y	LP001046	1	Male	Yes	Graduate	No	5955	5625	315	360
23	N	LP001047	0	Male	Yes	Not Graduate	No	2600	1911	116	360
24	N	LP001050	2	?	Yes	Not Graduate	No	3365	1917	112	360
25	N	LP001052	1	Male	Yes	Graduate	?	3717	2925	151	360
26	Y	LP001066	0	Male	Yes	Graduate	Yes	9560	0	191	360
27	Y	LP001068	0	Male	Yes	Graduate	No	2799	2253	122	360

DATA PREPROCESSING

- The data which was collected might contain missing values that may lead to inconsistency.
- To gain better results data need to be preprocessed so as to improve the efficiency o the algorithm.
- The outliers have to be removed and also variable conversion need to be done.
- In order to overcoming these issues we use map function.



SYSTEM ARCHITECTURE



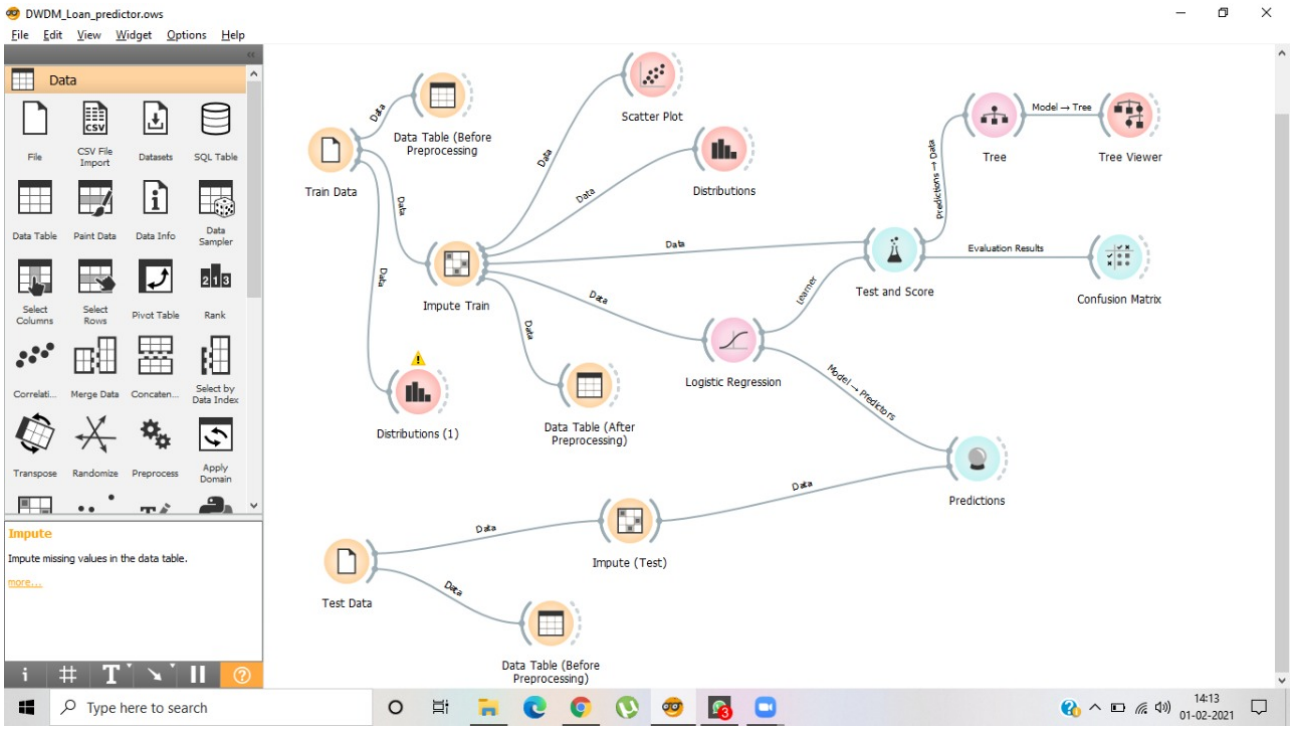
DATA MINING TASKS PERFORMED

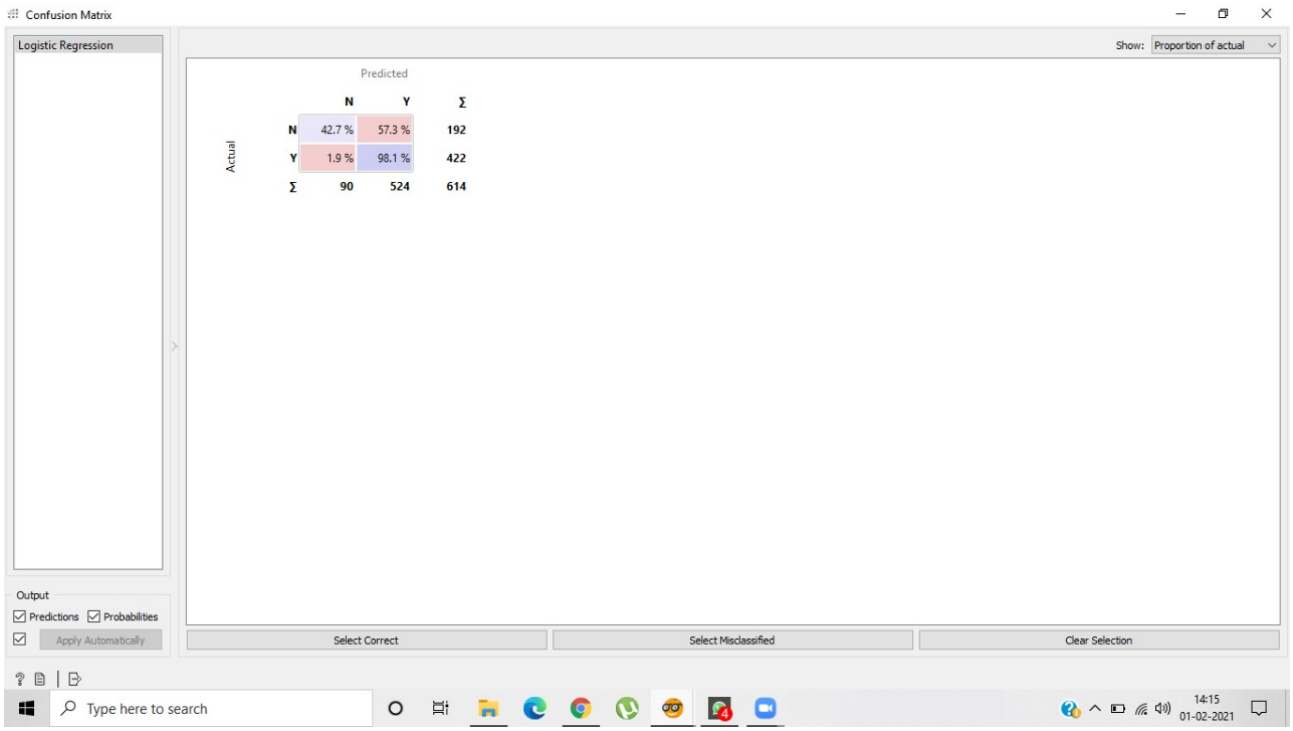
LOGISTIC REGRESSION

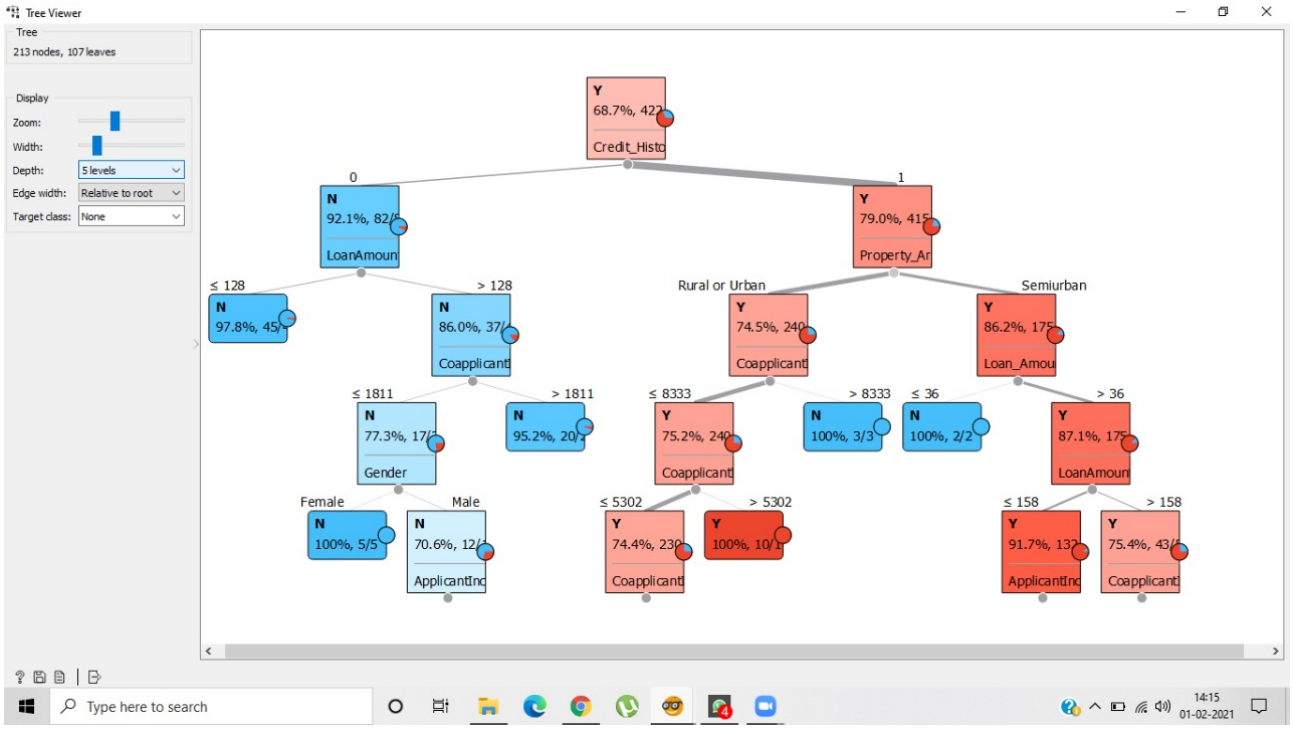
In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

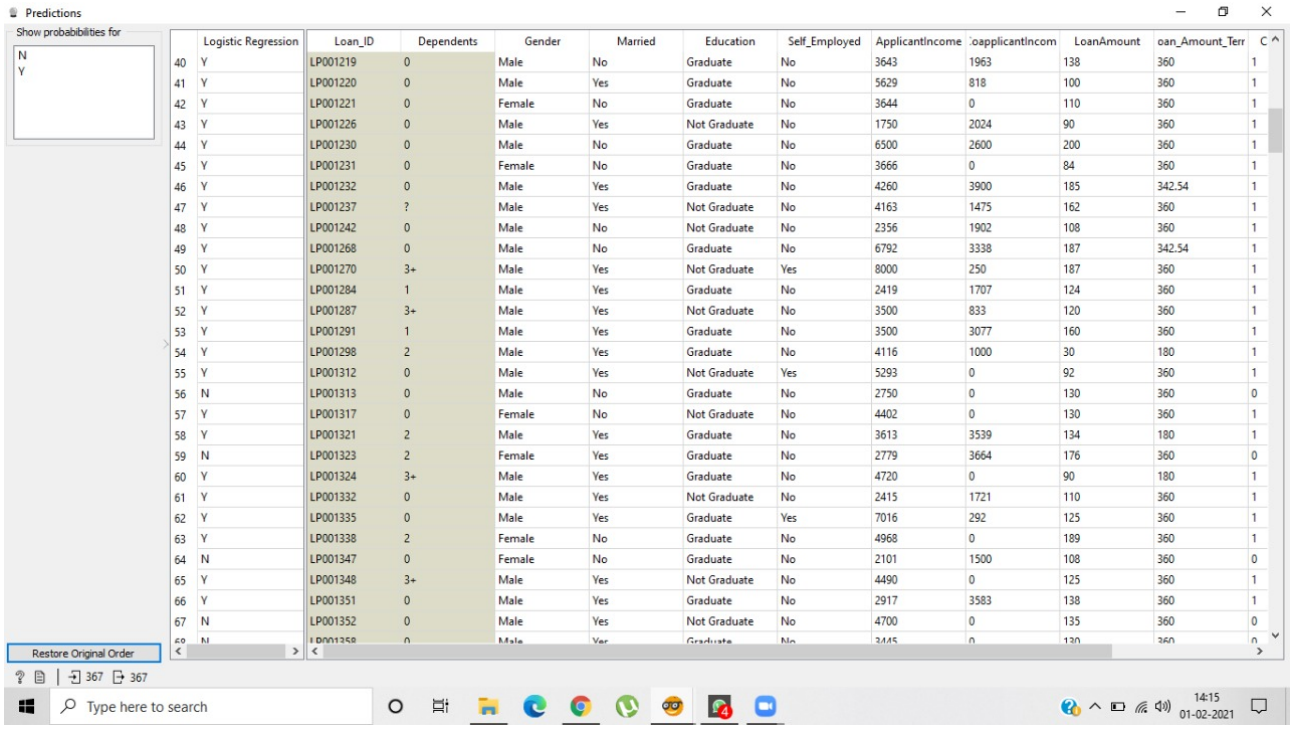
Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression.(or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labelling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names.

OUTPUT AND VISUALISATION









	Logistic Regression	Loan_ID	Dependents	Gender	Married	Education	Self_Employed	ApplicantIncome	oapplicantIncom	LoanAmount	oan_Amount_Terr	C
40	Y	LP001219	0	Male	No	Graduate	No	3643	1963	138	360	1
41	Y	LP001220	0	Male	Yes	Graduate	No	5629	818	100	360	1
42	Y	LP001221	0	Female	No	Graduate	No	3644	0	110	360	1
43	Y	LP001226	0	Male	Yes	Not Graduate	No	1750	2024	90	360	1
44	Y	LP001230	0	Male	No	Graduate	No	6500	2600	200	360	1
45	Y	LP001231	0	Female	No	Graduate	No	3666	0	84	360	1
46	Y	LP001232	0	Male	Yes	Graduate	No	4260	3900	185	342.54	1
47	Y	LP001237	?	Male	Yes	Not Graduate	No	4163	1475	162	360	1
48	Y	LP001242	0	Male	No	Not Graduate	No	2356	1902	108	360	1
49	Y	LP001268	0	Male	No	Graduate	No	6792	3338	187	342.54	1
50	Y	LP001270	3+	Male	Yes	Not Graduate	Yes	8000	250	187	360	1
51	Y	LP001284	1	Male	Yes	Graduate	No	2419	1707	124	360	1
52	Y	LP001287	3+	Male	Yes	Not Graduate	No	3500	833	120	360	1
53	Y	LP001291	1	Male	Yes	Graduate	No	3500	3077	160	360	1
54	Y	LP001298	2	Male	Yes	Graduate	No	4116	1000	30	180	1
55	Y	LP001312	0	Male	Yes	Not Graduate	Yes	5293	0	92	360	1
56	N	LP001313	0	Male	No	Graduate	No	2750	0	130	360	0
57	Y	LP001317	0	Female	No	Not Graduate	No	4402	0	130	360	1
58	Y	LP001321	2	Male	Yes	Graduate	No	3613	3539	134	180	1
59	N	LP001323	2	Female	Yes	Graduate	No	2779	3664	176	360	0
60	Y	LP001324	3+	Male	Yes	Graduate	No	4720	0	90	180	1
61	Y	LP001332	0	Male	Yes	Not Graduate	No	2415	1721	110	360	1
62	Y	LP001335	0	Male	Yes	Graduate	Yes	7016	292	125	360	1
63	Y	LP001338	2	Female	No	Graduate	No	4968	0	189	360	1
64	N	LP001347	0	Female	No	Graduate	No	2101	1500	108	360	0
65	Y	LP001348	3+	Male	Yes	Not Graduate	No	4490	0	125	360	1
66	Y	LP001351	0	Male	Yes	Graduate	No	2917	3583	138	360	1
67	N	LP001352	0	Male	Yes	Not Graduate	No	4700	0	135	360	0

CONCLUSION

The analytical process started from data cleaning and processing, Missing value imputation with mice package, then exploratory analysis and finally model building and evaluation. The best accuracy on public test set is 0.80. This brings some of the following insights about approval.

Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of a not paying back. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by the company.