RV Educational Institutions ®
RV College of Engineering ®

Autonomous Institution Affiliated to Visvesvaraya Technological University, Belagavi

Approved by AICTE, New Delhi, Accredited By NAAC, Bengaluru And NBA, New Delhi

Go, change the world

DEPARTMENT OF MASTER OF COMPUTER APPLICATIONS

MACHINE LEARNING
18MCA343

Assignment – Phase I

LOAN PRIDICTION SYSTEM

By

Ambika Badiger [1RV19MCA06]

Snehal Hukkeri [1RD19MCA09]

Under the Guidance

Of

Dr. Andhe Dharani

TABLE OF CONTENT

Slno.	Title	Page no.
1	Problem Statement	3
2	Existing system on problem	3
3	Gaps identified in the existing system	3
4	Algorithm to be used	4
5	Sample dataset with explanation	6
6	Dataset and features considered	8

1.PROBLEM STATEMENT:

Loan prediction system uses details provided in online form by customers to analyze whether the applicant to be approved for loan. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. However doing this manually takes a lot of time. Hence it wants to automate the loan eligibility process (real time) based on customer information. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. So the final thing is to identify the factors/ customer segments that are eligible for taking loan. Banks would give loans to only those customers that are eligible so that they can be assured of getting the money back. Hence the more accurate we are in predicting the eligible customers the more beneficial it would be for the Loan Prediction system.

2.EXISTING SYSTEM ON THE PROBLEM:

Banks collect the details of applicants such as Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. After collecting details they verify and analyze data to each applicant manually and then decide on approving the loan to the particular applicant who's details are accurate and beneficial to the Banks.

3.GAPS IDENTIFIED IN THE EXISTING SYSTEM:

Collecting all the details of applicants, verifying and analyzing them all manually would take a lot of time. Along with this comparing similarly data of previous records would be difficult and time consuming. Unlike in manual system, Loan prediction system all the previous records are compared to new applications of applicants, which makes compression and verification easy to decide on approving the loan to applicant.

4.ALGORITHMS USED FOR THE SYSTEM:

The above problem is a clear classification problem as we need to classify whether the Loan_Status is yes or no. So this can be solved by any of the classification techniques like

1. Logistic Regression .
2. Decision Tree Algorithm.
3. Random Forest Technique.

1. Logistic Regression :

- Logistic regression is one of the most popular Machine learning algorithm that comes under Supervised Learning techniques.
- It can be used for Classification as well as for Regression problems, but mainly used for Classification problems.
- Logistic regression is used to predict the categorical dependent variable with the help of independent variables.
- The output of Logistic Regression problem can be only between the 0 and 1.
- Logistic regression can be used where the probabilities between two classes is required. Such as whether it will rain today or not, either 0 or 1, true or false etc.
- Logistic regression is based on the concept of Maximum Likelihood estimation. According to this estimation, the observed data should be most probable.
- In logistic regression, we pass the weighted sum of inputs through an activation function that can map values in between 0 and 1. Such activation function is known as sigmoid function and the curve obtained is called as sigmoid curve or S-curve.

2. Decision Tree Algorithm:

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.

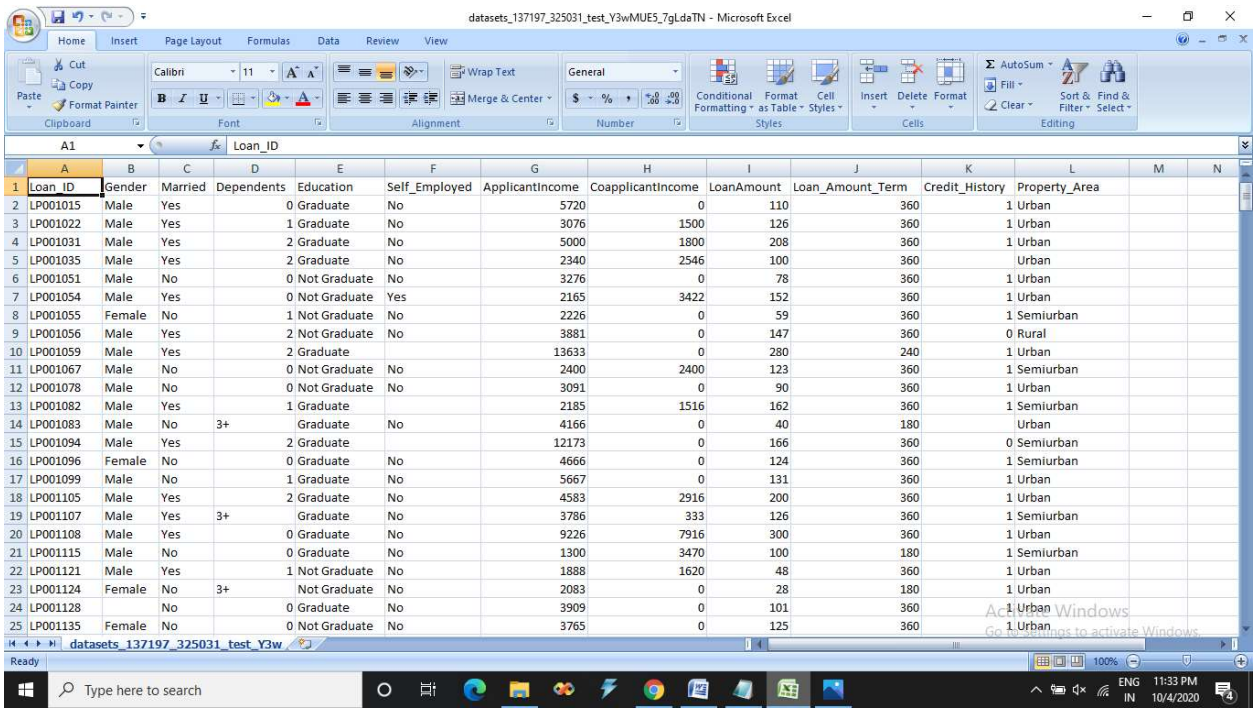
3. Random Forest Technique:

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

5.SAMPLE DATASETS WITH EXPLANATION TO BE GIVEN:

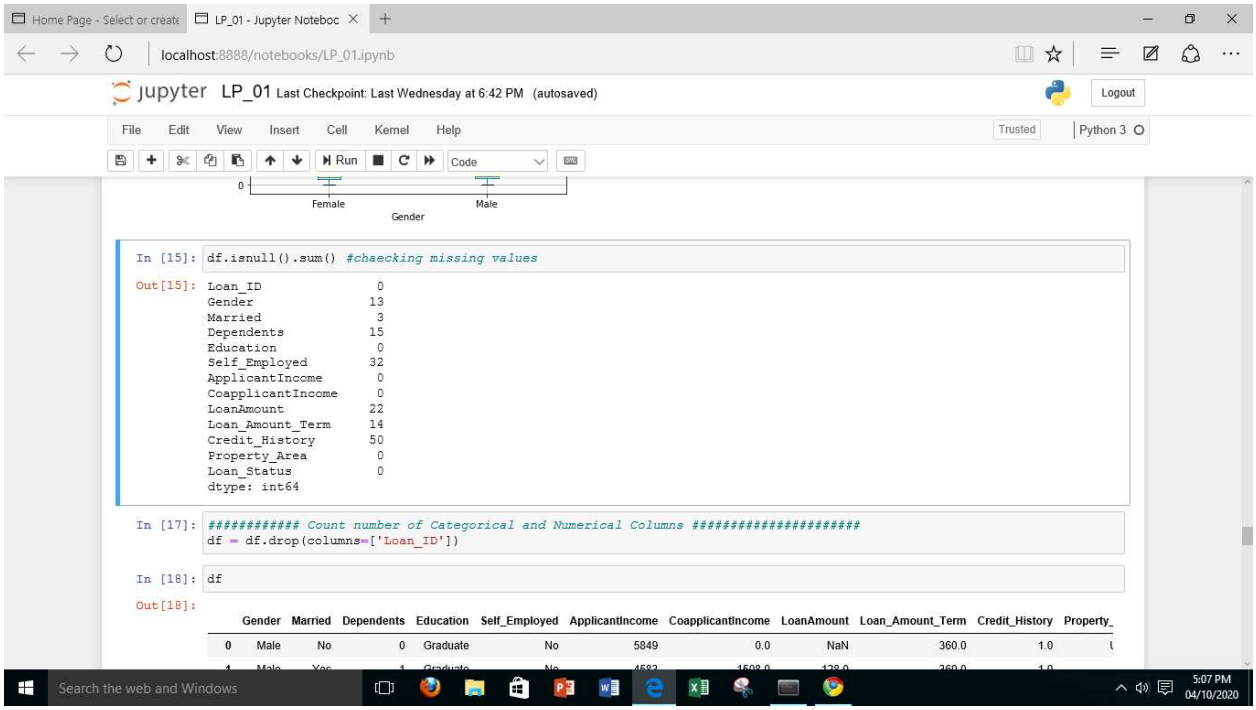
The sample data set consist of 13 columns and 614 rows, including target variable, all of them are self explanatory. Here columns represents the features of the dataset and row represents the sample entries of the dataset. The data types that features of dataset have are float, integer, object. We also see some missing values, lets take stock of missing columns and what are the possible values for categorical and numerical columns. The features of dataset are as follows

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:	Loan approved (Y/N)



Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
LP001015	Male	Yes	0	Graduate	No	5720	0	110	360	1	Urban
LP001022	Male	Yes	1	Graduate	No	3076	1500	126	360	1	Urban
LP001031	Male	Yes	2	Graduate	No	5000	1800	208	360	1	Urban
LP001035	Male	Yes	2	Graduate	No	2340	2546	100	360		Urban
LP001051	Male	No	0	Not Graduate	No	3276	0	78	360	1	Urban
LP001054	Male	Yes	0	Not Graduate	Yes	2165	3422	152	360	1	Urban
LP001055	Female	No	1	Not Graduate	No	2226	0	59	360	1	Semiurban
LP001056	Male	Yes	2	Not Graduate	No	3881	0	147	360	0	Rural
LP001059	Male	Yes	2	Graduate		13633	0	280	240	1	Urban
LP001067	Male	No	0	Not Graduate	No	2400	2400	123	360	1	Semiurban
LP001078	Male	No	0	Not Graduate	No	3091	0	90	360	1	Urban
LP001082	Male	Yes	1	Graduate		2185	1516	162	360	1	Semiurban
LP001083	Male	No	3+	Graduate	No	4166	0	40	180		Urban
LP001094	Male	Yes	2	Graduate		12173	0	166	360	0	Semiurban
LP001096	Female	No	0	Graduate	No	4666	0	124	360	1	Semiurban
LP001099	Male	No	1	Graduate	No	5667	0	131	360	1	Urban
LP001105	Male	Yes	2	Graduate	No	4583	2916	200	360	1	Urban
LP001107	Male	Yes	3+	Graduate	No	3786	333	126	360	1	Semiurban
LP001108	Male	Yes	0	Graduate	No	9226	7916	300	360	1	Urban
LP001115	Male	No	0	Graduate	No	1300	3470	100	180	1	Semiurban
LP001121	Male	Yes	1	Not Graduate	No	1888	1620	48	360	1	Urban
LP001124	Female	No	3+	Not Graduate	No	2083	0	28	180	1	Urban
LP001128		No	0	Graduate	No	3909	0	101	360	1	Urban
LP001135	Female	No	0	Not Graduate	No	3765	0	125	360	1	Urban

The above screenshot is of sample dataset. In this we can see 13 columns, as described earlier and rows of sample entries.



The above screen shot shows the number of missing values from the dataset used.

6.DATASET AND FEATURES CONSIDERED – FEATURE TRASFORAMTION TO BE CARRIED OUT:

Features of Dataset:

- Loan ID -> As the name suggests each person should have a unique loan ID.
- Gender -> In general it is male or female. No offence for not including the third gender.
- Married -> Applicant who is married is represented by Y and not married is represented as N. The information regarding whether the applicant who is married is divorced or not has not been provided. So we don't need to worry regarding all these.
- Dependents -> the number of people dependent on the applicant who has taken loan has been provided.
- Education -> It is either non -graduate or graduate. The assumption I can make is " The probability of clearing the loan amount would be higher if the applicant is a graduate".
- Self_Employed -> As the name suggests Self Employed means , he/she is employed for himself/herself only. So freelancer or having a own business might come in this category. An applicant who is self employed is represented by Y and the one who is not is represented by N.
- Applicant Income -> Applicant Income suggests the income by Applicant. So the general assumption that i can make would be "The one who earns more have a high probability of clearing loan amount and would be highly eligible for loan "
- Co Applicant income -> this represents the income of co-applicant. I can also assume that " If co applicant income is higher , the probability of being eligible would be higher "
- Loan Amount -> This amount represents the loan amount in thousands. One assumption I can make is that " If Loan amount is higher , the probability of repaying would be lesser and vice versa"
- Loan_Amount_Term -> This represents the number of months required to repay the loan.
- Credit_History -> When I googled it , I got this information. A credit history is a record of a borrower's responsible repayment of debts. It suggests → 1 denotes that the credit history is good and 0 otherwise.

- Property_Area -> The area where they belong to is my general assumption as nothing more is told. Here it can be three types. Urban or Semi Urban or Rural
- Loan_Status -> If the applicant is eligible for loan it's yes represented by Y else it's no represented by N.

Feature transformation to be carried out:

Input data needs to be pre-processed before we feed it to model. Following things need to be taken care:

1. Encoding Categorical Features.
2. Imputing missing values.
3. Deleting entire rows of missing values.

- Encoding Categorical Features:

  In machine learning projects, one important part is feature engineering. It is very common to see categorical features in a dataset. However, our machine learning algorithm can only read numerical values. It is essential to encoding categorical features into numerical values.

  Here we will cover three different ways of encoding categorical features:

  1. LabelEncoder and OneHotEncoder
  2. DictVectorizer
  3. Pandas get_dummies

- Imputing missing values:

  A basic strategy to use incomplete datasets is to discard entire rows and/or columns containing missing values. However, this comes at the price of losing data which may be valuable (even though incomplete). A better strategy is to impute the missing values, i.e., to infer them from the known part of the data.

- Deleting entire rows of missing values:

  It removes all the rows which had any missing value. It didn't modify the original data frame, it just returned a copy with modified contents.