Loan Prediction Problem

Rahul Gupta

Problem Statement

About Company

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

Problem

Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customer's segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

Data Dictionary

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	Loan approved (Y/N)

Evaluation Metric

Percentage of loan approval correctly predicted.

Provided Dataset

1. Train file in the csv format
2. Test file in the csv format

Tools Used

RStudio, Microsoft Excel, SAS Enterprise Guide and SAS Miner

Exploratory Data Analysis

Few of the observations of train dataset

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N

Summary Statistics of the train data

```
 Loan_ID          Gender         Married    Dependents          Education
 LP001002:  1           : 13          :3           : 15       Graduate    :480
 LP001003:  1    Female:112       No :213    0  :345       Not Graduate:134
 LP001005:  1    Male  :489       Yes:398    1  :102
 LP001006:  1                                  2 :101
 LP001008:  1                                  3+: 51
 LP001011:  1
 (Other) :608

Self_Employed     ApplicantIncome      CoapplicantIncome   LoanAmount
     : 32         Min.   :  150        Min.   :    0      Min.   :  9.0
 No :500          1st Qu.: 2878        1st Qu.:    0      1st Qu.:100.0
 Yes: 82          Median : 3812        Median : 1188      Median :128.0
                  Mean   : 5403        Mean   : 1621      Mean   :146.4
                  3rd Qu.: 5795        3rd Qu.: 2297      3rd Qu.:168.0
                  Max.   :81000        Max.   :41667      Max.   :700.0
                                                          NA's   :22

Loan_Amount_Term Credit_History       Property_Area Loan_Status
Min.   : 12      Min.   :0.0000       Rural    :179   N:192
1st Qu.:360      1st Qu.:1.0000       Semiurban:233   Y:422
Median :360      Median :1.0000       Urban    :202
Mean   :342      Mean   :0.8422
3rd Qu.:360      3rd Qu.:1.0000
Max.   :480      Max.   :1.0000
NA's   :14       NA's   :50
```

Observations

1. There are 614 observations with 13 attributes.
2. Loan_ID will not be used in the further analysis as it is an ID variable for each loan application.
3. Loan_Status is the dependent variable.
4. There are 11 covariates which includes both categorical as well as continuous.
5. There are two types of missing values one is ‘NA’ and other is ‘’ as highlighted above

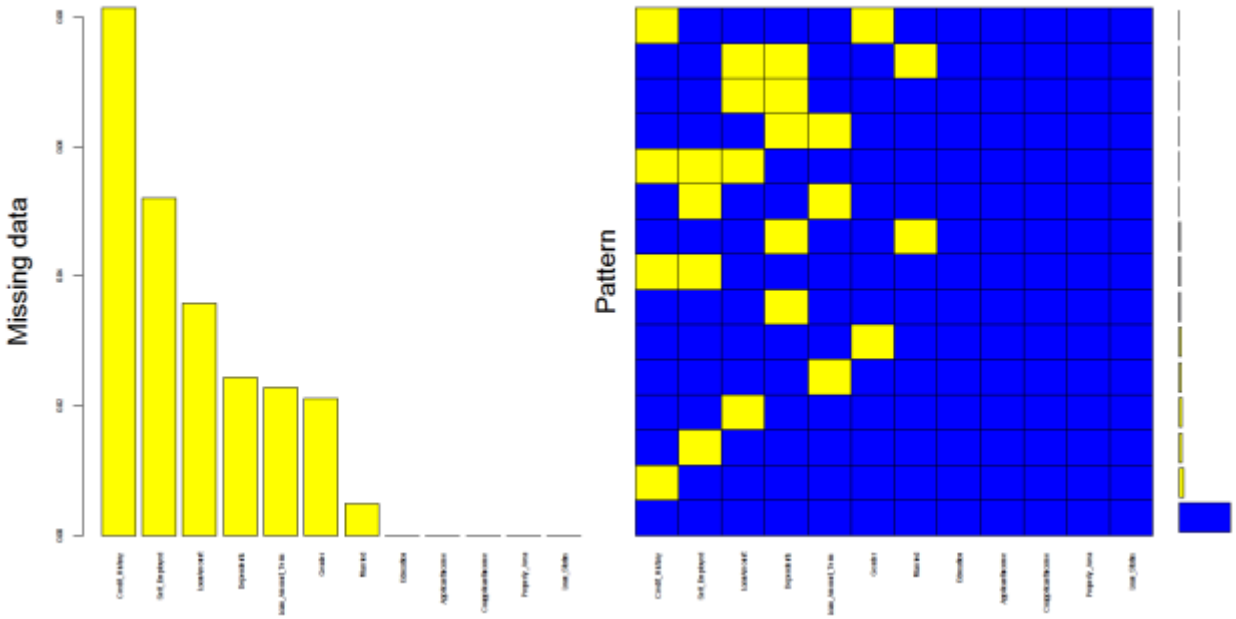
Handling Missing Data

Step 1. Making all missing values as 'NA' because R programming understands NA as missing whereas '' will be treated as a value not as a missing.

Step 2. Calculating missing value statistics from the train dataset and plotting them.

Variables sorted by number of missingness:

Variable	Count
Credit_History	50
Self_Employed	32
LoanAmount	22
Dependents	15
Loan_Amount_Term	14
Gender	13
Married	3
Education	0
ApplicantIncome	0
CoapplicantIncome	0
Property_Area	0
Loan_Status	0



Observations

1. Overall there are 149 observations which has missing values, which is approximately 25% of the data, so list wise deletion option will not be considered.
2. None of the observation have more than 30% missing values.
3. Credit_History variable is missing in 50 observations followed by Self_Employment, LoanAmount, Dependents, Loan_Amount_Term, Gender, Married.

Step 3. Conducting Little's Test to know the pattern of missing data

Null Hypothesis of Little's Test is missing values in dataset is of type MCAR (Missing Completely At Random)

H_0 : Missing values are of type MCAR

Using 'LittleMCAR' function present in 'BaylorEdPsych' package of R, got the **p-value of 0.0209** which is less than 0.05. So rejecting the Null Hypothesis.

Hence missingness in dataset is not of type MCAR.

Because the data is not missing completely at random, it is not safe to list wise delete cases with missing values or singly impute missing values. So Multiple Imputation methods will be used to impute the missing values.

Step 4. Preparing the data for imputation.

Assigning numerical values to categorical variables.

For Gender variable : 'Male' = 1 and 'Female' = 0

For Education variable : 'Graduate' = 1 and 'Not Graduate' = 0

For Dependents variable : '0' = 0, '1' = 1, '2' = 2 and '3+' = 3

For Married variable : 'Yes' = 1 and 'No' = 0

For Self_Employed variable : 'Yes' = 1 and 'No' = 0

For Property_Area variable : 'Urban' = 1, 'Semiurban' = 2 and 'Rural' = 0

For Loan_Status variable : 'Yes' = 1 and 'No' = 0

Credit_History is already coded as 1 and 0

This how the structure of train dataset looks like :

```
'data.frame':   614 observations with 12 variables
Gender            : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
Married           : Factor w/ 2 levels "0","1": 1 2 2 2 1 2 2 2 2 2 ...
Education         : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 1 2 2 2 ...
Self_Employed     : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 1 1 1 ...
Property_Area     : Factor w/ 3 levels "0","1","2": 2 1 2 2 2 2 2 3 2 3 ...
Dependents        : Factor w/ 4 levels "0","1","2","3": 1 2 1 3 1 4 3 2 ...
ApplicantIncome   : int  5849 4583 3000 2583 6000 5417 2333 4006 12841 ...
CoapplicantIncome : num  0 1508 0 2358 0 ...
LoanAmount        : int  NA 128 66 120 141 267 95 158 168 349 ...
Loan_Amount_Term  : int  360 360 360 360 360 360 360 360 360 360 ...
Credit_History    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 1 2 2 ...
Loan_Status       : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 1 2 1 ...
```

Observations

1. Dataset is not an MCAR, as it is clear by Little's Test
2. Multiple Imputation techniques will be used to impute the missing values

Step 5. Imputing the missing values using Multiple Imputation techniques

Using the 'mice' package in R, all types of missing values (Binary, Multinomial categorical and continuous) present in the dataset can be handled.

i. Imputing Binary categorical variables which includes *Gender, Self_Employed, Credit_History and Married.*

Logistic Regression technique is used to impute these variables. Tuning parameters of ***mice*** function includes number of iterations = 5 and method = 'logreg'.

VARIABLE	**BEFORE IMPUTATION**			**AFTER IMPUTATION**		
	0's	1's	NA's	0's	1's	NA's
GENDER	112	489	13	114	500	0
SELF_EMPLOYED	500	82	32	527	87	0
CREDIT_HISTORY	89	475	50	97	517	0
MARRIED	213	398	3	214	400	0

ii. Imputing multinomial categorical variable.

Proportional odds model technique is used to impute multi category variable *Dependent.* Tuning parameters of mice function includes, number of iterations = 10 and method = 'polr'.

	BEFORE IMPUTATION					**AFTER IMPUTATION**				
VARIABLE	0's	1's	2's	3's	NA's	0's	1's	2's	3's	NA's
DEPENDENT	345	102	101	51	15	355	105	102	52	0

iii. Imputing continuous variables which include *LoanAmount* and *Loan_Amount_Term*

Linear Regression technique is used to impute the missing values with 5 iterations which will make results more stable.

VARIABLE	**BEFORE IMPUTATION**		**AFTER IMPUTATION**	
	Mean	Std Deviation	Mean	Std Deviation
LOANAMOUNT	146.4122	85.587	146.227	84.301
LOAN_AMOUNT_TERM	342	65.120	341.962	64.376

Preparing datasets for modeling and validation

The dataset has been treated for missing values and brought to the form where it can be used in the modeling.

For the purpose of checking the accuracy of the classification model built, the dataset is split into parts i.e. train and validation in the 70-30 ratio.

Setting the Naïve Rule

Initial dataset has 614 observations with Loan_Status =1 (Yes) as a majority class, having approx. proportion of 69:31. This leads to setting of the majority rule for any of the model built using this data, that is it should surpass the majority class rule, which is of 69% for the given dataset.

Final datasets for modelling and validation

Training dataset has 431 observations with 69:31 proportion of Yes/No for Loan_Status.

Validation dataset has 183 observations with same 69:31 proportion of Yes/No for Loan_Status.

Model Building and Validating

Model for the purpose of classification: Used two famous algorithms for classification viz. '*Random Forest*' and *'Decision Tree*'. Both belongs to the family of CART. Random forest results will be more robust because it will be based on number of trees built and taking majority vote, whereas Decision Tree will give us a simple classification rule which can be used to classify the applicants without any statistical software (e.g. if income > A and age < B, eligible)

Model for the purpose of explanation: Apart from classifying the applicant into eligible or not, bank personnel's also need some insights on the variables which play crucial role in the loan application to be considered as eligible or not. For this purpose, used the '*Binary Logistic Regression'* model which will give the odds ratio of the variables involved in predicting the loan eligibility.

Model I – Random Forest

Using '**randomForest'** package in R with following specifications

Number of trees grown = 500

Node size = 35; this will act as a pruning parameter and prevent the trees from growing to overfitting

Built the model on training dataset which has 431 observations and validating the results on validation dataset which has 183 observations.

Below is the R code snippet

```
randomForest(Loan_Status~.,data=train,ntree=500,nodesize=35,strata=train$Loan_Status)
confusionMatrix(predict(model,valid[,-12]),valid$Loan_Status,positive = '1')
```

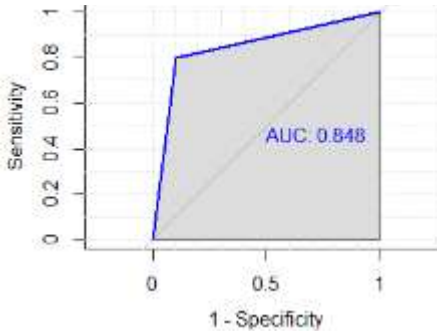
Confusion Matrix

	ACTUAL/REFERENCE	
PREDICTION	0	1
0	26	3
1	31	123

Accuracy : (26 + 123)/(26 + 3 + 31 +123) = 0.8142

Sensitivity : 0.9762; **Specificity** : 0.4561

ROC Curve along with AUC

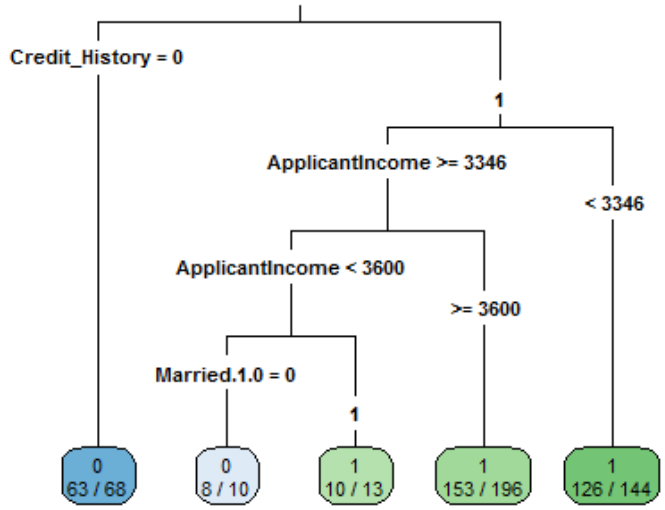


Variable Importance

VARIABLE	GINI INDEX
CREDIT HISTORY	47.51
APPLICANTINCOME	6.95
LOANAMOUNT	5.47
COAPPLICANTINCOME	4.58

Model II – Decision Tree

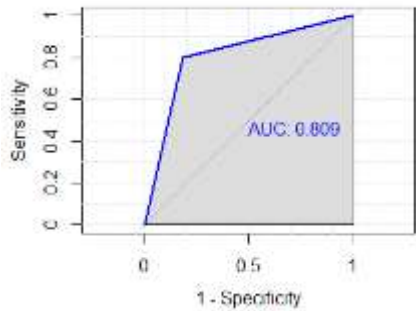
Decision Tree is built using 'rpart' and 'party' packages in R



Variable Importance

VARIABLE
CREDIT_HISTORY
APPLICANTINCOME
MARRIED

ROC Curve along with AUC



Confusion Matrix

	ACTUAL/REFERENCE	
PREDICTION	0	1
0	27	6
1	30	120

Accuracy : (27 + 120)/(27 + 6 + 30 +120) = 0.8033

Sensitivity : 0.9524; **Specificity** : 0.4737

Model III – Logistic Regression

Generalized linear model with binomial family and logit link is used to build logistic model.

In the first step Full model is compared with Null model to check Full model holds in comparison to Null Model. Below is the R code snippets for Full and Null model

glm(Loan_Status~.,data = train,family = "binomial")

glm(Loan_Status~1,data = train,family = "binomial")

Comparing Null model with the Full model on the basis of deviance

Null Model Deviance: 535.87 on 430 degrees of freedom
Full Model Deviance: 381.10 on 416 degrees of freedom

ΔDeviance = 154.77 with Δdf = 14

Chi-sq critical on df=14 with $\alpha = 0.05$ is approximately 24 which is far below than observed ΔDeviance, so there is evidence to reject the Null Hypothesis (H_0) and we can proceed with our full model as a baseline to search better nested model which is less complex.

Using forward, backward and stepwise approach for reduced model selection. Below is the R code snippet

step(logistic_NULL,scope=list(lower=logistic_NULL,upper=logistic_FULL),direction = "forward")

step(logistic_FULL,scope=list(lower=logistic_NULL,upper=logistic_FULL),direction="backward")

step(logistic_FULL,scope=list(lower=logistic_NULL,upper=logistic_FULL),direction = "both")

All the above procedures select the same model, which includes ***Credit_History, Married, Property_Area***

Comparing reduced model with the Full model

Full Model Deviance = 381.1 on df=416
Reduced Model Deviance =384.8 on df=426

ΔDeviance = 3.7 on Δdf = 10; Chi-sq critical at df=10 with $\alpha = 0.05$ is approximately 18 which means p-value >> 0.05. Hence the reduced model which includes ***Credit_History, Married, Property_Area*** holds in comparison to full model.

Final model:

glm(Loan_Status~Credit_History+Married.1.0+Property_Area.1.0,data=train,family = "binomial")

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.21225    0.54572  -5.886 3.95e-09 ***
Credit_History1    4.05575    0.49272   8.231  < 2e-16 ***
Married.1.01       0.56727    0.26236   2.162   0.0306 *
Property_Area.1.01 -0.09711   0.30912  -0.314   0.7534
Property_Area.1.02  0.71901   0.32633   2.203   0.0276 *
```

All these variables are categorical where Credit_History1 means applicant is meeting the credit history guidelines, Married.1.01 means applicant is married, Property_Area.1.01 means the property is in urban area and Property_Area.1.02 means property is in suburban area

This model gives an **accuracy of 0.8142 with sensitivity = 0.9762 and specificity = 0.4561;** cutoff used to classify the validation dataset observations is 0.5

Interpreting Odds ratio:

- Parameter estimate (β) of **Credit_History** : 4.056

Odds ratio = $e^{\beta} = e^{4.056} = 57.74$; the chances of loan application getting approved increases almost 58 times for applicant who meets the credit history guidelines (=1) in comparison to those who don't meet the guidelines, controlling for other variables.

- Parameter estimate (β) of **Married**: 0.567

Odds ratio = $e^{\beta} = e^{0.567} = 1.76$; the chances of loan application getting approved increases almost 1.76 times for applicant who are married (=1) in comparison to those who are not, controlling for other variables.

- Parameter estimate (β) of Property_Area

Since Property_Area is a multinomial variable having three category values i.e. 'Urban' = 1, 'Semiurban' = 2 and 'Rural' = 0

Parameter estimates of **Urban Property_Area** w.r.t. Rural = -0.097

Odds Ratio = $e^{\beta} = e^{-0.097} = 0.907 \approx 1$; having property in the urban area does not add to the chances of getting loan application approved in comparison to those having property in Rural area. This also makes sense as this parameter estimate is insignificant since its p-value >>0.05

Parameter estimates of **Suburban Property_Area w.r.t. Rural** = 0.719

Odds Ratio = $e^{\beta} = e^{0.719} = 2.05$; having a property in the Suburban area increases the chances of loan application getting approved to almost double in comparison to those having property in Rural area, controlling for other variables.

Final variables and Model Selection:

First model build was Random Forest with the aim of classification and to get an idea on the underlying behavior of the dataset. It gives a good accuracy with high sensitivity and average specificity. Though Random Forest is not designed for feature selection, just to know the importance of variables, it was found that Credit_History is the most important feature. Next model build was simple decision tree with the purpose of getting a simple rule to classify the applicants. The accuracy was good but lower than Random Forest. Decision tree also picks the Credit_History as the most important feature followed by ApplicantIncome and Marriage. Final model build was Logistic Regression with the purpose of interpretation and explanation i.e. how variables are impacting the approval of loan application. It picks three variables giving the same accuracy as that of Random Forest.

Final list of variables from multiple models that bank personnel's can use to decide on loan application :

- Credit_History
- Married
- ApplicantIncome
- Property_Area

Model selected: Random Forest for classification purpose and Logistic regression for interpretation.

Executive Summary:

Dream Housing Finance Company which deals in home loans want to automate their loan application process. The process will be like, the applicant will apply for the loan using their web portal and will fill up a form which has fields like Name, Age, Gender, Education, Number of Dependents, Credit History, Marital Status, Income, Co-applicant income, type of property they own, Loan Amount they are looking for etc. On filling up the form correctly, the background engine will process this information and tells the applicant in real time whether they are eligible or not, instead of waiting for number of days to hear from Company officials. This will also decrease the work load of employees and decrease the chances of human error, since this is completely automated process with no human intervention.

To achieve this goal what is needed is that engine which will process the information provided by the applicant and take decides accordingly. The engine will be a predictive model which will check the eligibility of applicants. Since this is a classification problem, so models designed for classification are used. Along with classification model, a list of important variables is also provided that will help the Finance Company to design their processes, be it in marketing or otherwise.

Recommendations on the basis of parameters having impact on loan eligibility are:

Credit History – whether the applicant is meeting the credit history guidelines or not. This is the most important criteria for loan eligibility. Applicants which are meeting the guidelines have almost 58 times more chances that their application will be approved. This being the most critical criterion in loan eligibility that Company can even put a disclaimer in the website that meeting the credit history guidelines are mandatory.

Marital Status – whether the applicant is married or not. The chances of loan application getting approved is almost 1.76 times if applicant is married. This may be because of social phenomena that person becomes more mature and responsible after marriage and there are less chances that they will default. This suggests that marketing strategies can be built around married people.

Property – the type of property owned by applicant also plays an important role in loan eligibility. Suburban property is preferred over urban and rural properties. The reason behind this may be, that suburban properties have more scope of increase in the value as compared to urban which has already achieved their maximum value. As far as statistics are concerned, the chances of loan application getting approved is more than two times if applicant has a suburban property as compared to applicant with urban or rural property. This can be a hint that the Finance Company needs to target suburban areas. Company can think of opening more branches in the suburban areas.

Applicant Income – income of applicant is also an important criterion in deciding the loan eligibility. This is quite intuitive that if applicant has good income chances of application getting approved are more.

Below are the takeaways:

1. Make Credit history, Marital Status, Property, Income as mandatory fields in the application form.
2. Open more branches/offices in the suburban areas.
3. Design the policies considering married people.
4. Give maximum weight to the credit history of applicants.
5. Gender, Education, Dependents have no influence in deciding the loan eligibility.