Loan Prediction III

Problem Statement

The Dream Housing Finance Company deals with Home loans. The company wants to automate the Loan Eligibility process in real time based on inputs/details provided by the customer.

Hypothesis

H0 – Null hypothesis : There is no feature exists which has impact on the dependent variable.
Ha – Alternate hypothesis : There exists feature which has impact on the dependent variable.
Main Factors that determine the Loan Approval (as per study) :

Credit history : Credit score is an indication of aptplicants creditworthiness and is numerical in nature. Better the score, better are chances of Loan approval
Income (Applicant & CoApplicant): Higher the income, better are chances of Loan approval
Loan term : Shorter the Loan term, higher are chances of Loan Approval
Age : More the number of Working Years, better are chances of Loan Approval
Professional background : Most banks prefer professional from Govt./Corporate background than "Self-Employed". If applicants are from Govt./Corporate professional background, higher are chances of Loan Approval.
Existing loans : Existing loans contribute to Credit History, in turn it affects the loan approvals
Attributes of the property (New, how old, urban/rural and so on) : Locality, age of the property, market value impacts the Loan Approval
Marital status : If married and spouse is also earning (tax payer), chances of Loan Approval is higher
Dependents : Smaller the family, better are chances of Loan Approval. Less dependents means more borrowing power.
Loan Amount : Smaller the Loan Amount, higher are chances of Loan Approval
EMI : Smaller the EMI, higher are chances of Loan Approval
Debt-to-income ratio : ratio is the amount of debt you have relative to income—including your mortgage payments. Ex: ratio is the amount of debt you have relative to income—including your mortgage payments.
Lower the DTI, chances of getting the loan approval is better.
Household Expenditure Measure or HEM : Lower HEM, better is the borrowing capacity.

Input Details/Parameters – Getting Data

Download the data from Analytics Vidhya [Practice Problem : Loan Prediction III]

Below are main parameters:

Predictors :

Loan_ID
Gender
Married
Dependents
Education
Self_Employed
ApplicantIncome
CoapplicantIncome
LoanAmount
Loan_Amount_Term
Credit_History
Property_Area

Target :

Loan_Status

DATA EXPLORATION

Data Types

FEATURE	DATA TYPE	INFERENCE
Loan_ID	object	Categorical
Gender	object	Categorical
Married	object	Categorical
Dependents	object	Categorical/Ordinal
Education	object	Categorical/Ordinal
Self_Employed	object	Categorical
ApplicantIncome	int64	Continuous/Integer
CoapplicantIncome	float64	Continuous/ Float
LoanAmount	float64	Continuous/Float
Loan_Amount_Term	float64	Categorical/Float
Credit_History	float64	Categorical/Float
Property_Area	object	Categorical
Loan_Status	object	Categorical

Variable Identification

Variable Category	
Categorical:	
	Gender
	Married
	Education (Ordinal)
	Self_Employed
	Credit_History
	Dependents (Ordinal)
	Loan_Amount_Term
	Property_Area (Ordinal)
Continuous:	
	ApplicantIncome
	CoapplicantIncome
	LoanAmount

Size of sets

Train set has 614 Rows and 13 Columns
Test set has 367 Rows and 12 Columns

UniVariate Analysis

Check central tendencies mean, median, mode, min, max, quantiles of continuous features using the method “describe()”

train.describe()

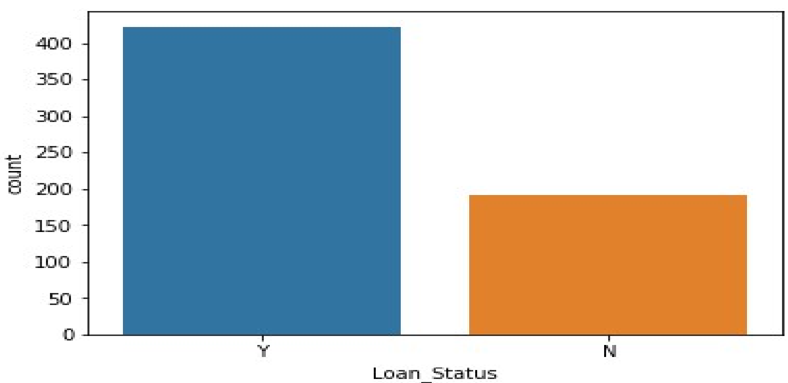
TRAIN	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614	614	592	600	564
mean	5403.459283	1621.245798	146.412162	342	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150	0	9	12	0
25%	2877.5	0	100	360	1
50%	3812.5	1188.5	128	360	1
75%	5795	2297.25	168	360	1
max	81000	41667	700	480	1

test.describe()

TEST	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	367	367	362	361	338
mean	4805.599455	1569.577657	136.132597	342.537396	0.825444
std	4910.685399	2334.232099	61.366652	65.156643	0.38015
min	0	0	28	6	0
25%	2864	0	100.25	360	1
50%	3786	1025	125	360	1
75%	5060	2430.5	158	360	1
max	72529	24000	550	480	1

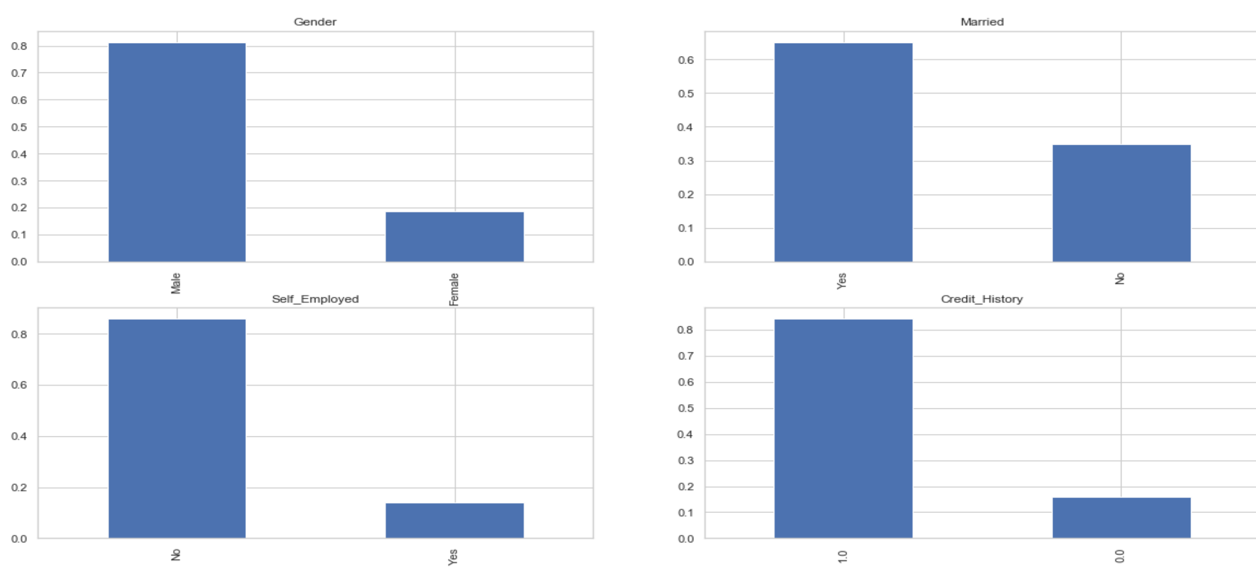
Check that for ApplicanIncome and CoapplicantIncome standard deviation is High in both test and train sets, which indicates that values are spread across wider range.

Target – Loan_Status



From the above count plot, “Loan_status” is not normally distributed. They are not 50:50 distributed.
N=192, Y=422 => Distribution Ration N:Y is **31:69**
Normal distributed target features will help in better modeling of data.

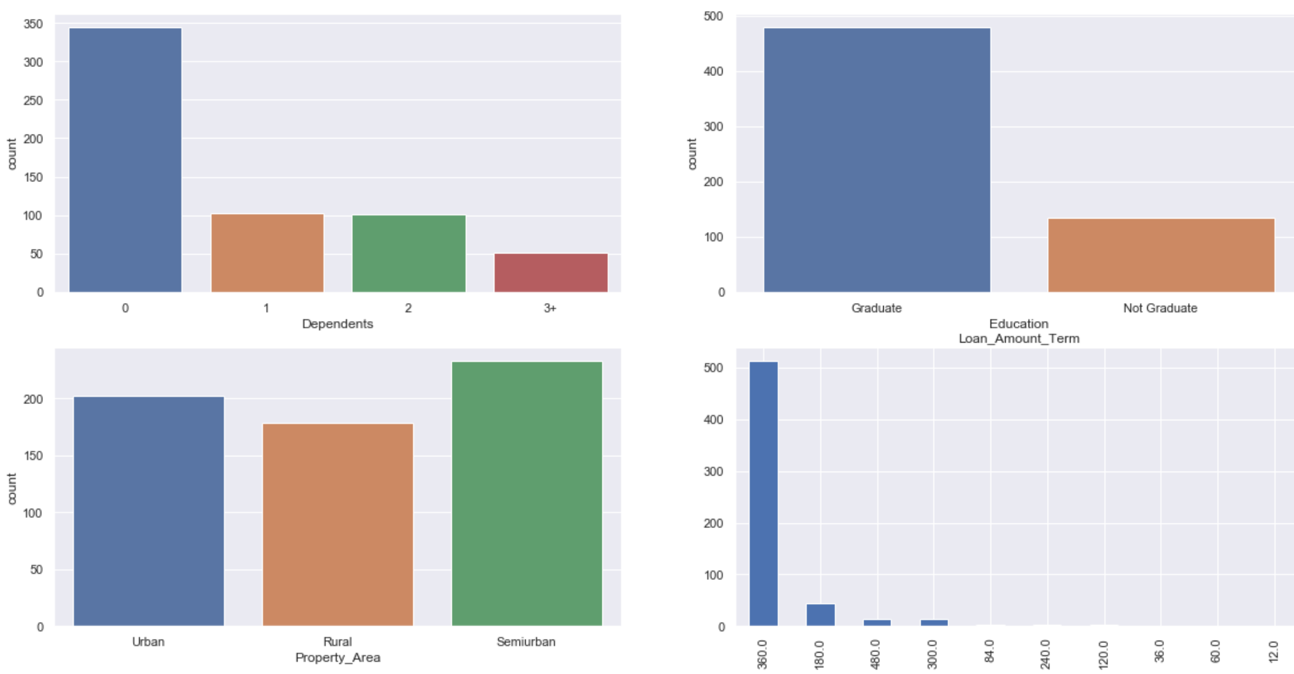
Categorical Features (Independent)



From the above count plot, it can be inferred that features are not normally distributed

1. ~80% applicants are male (Male : Female ratio => 80:20)
2. ~65% applicants are married (Married:Unmarried ratio => 65:35)
3. ~85% applicants are not self-employed (Self-Employed:Non Self-Employed ratio => 85:15)
4. ~85% applicants have good Credit_History

Ordinal Features (Independent)



From the above count plot, it can be inferred that

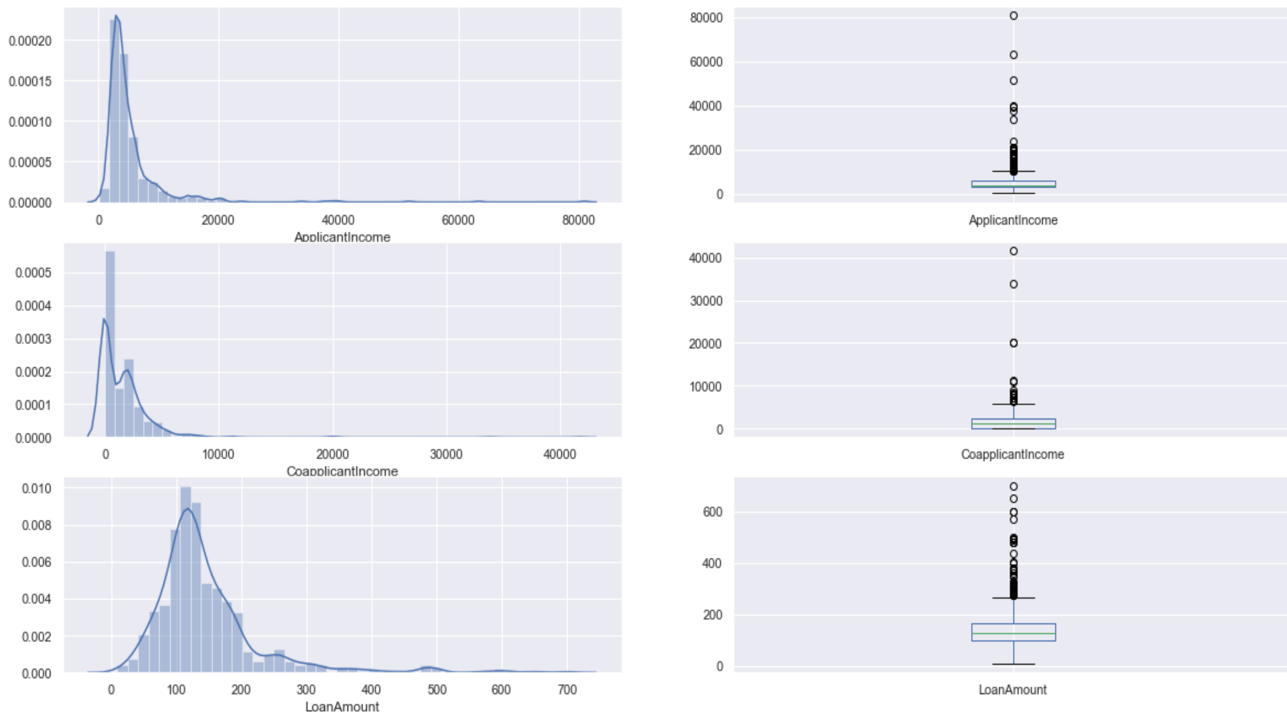
1. Most of the applicants have no children
2. Most of applicants are 'Graduates'
3. Most of applicants are from 'Semi Urban' area
4. Most of loan applications are for term – 360 months

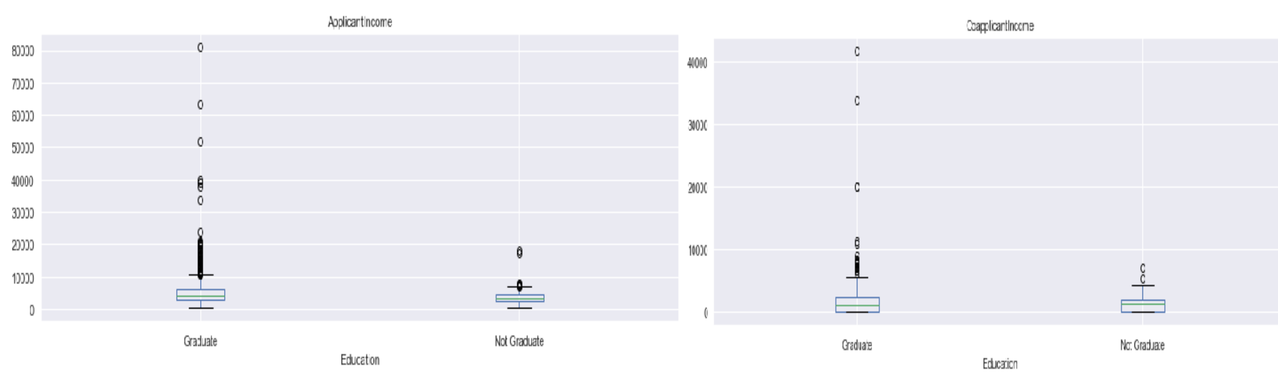
Numerical Features (Independent)

Below points can be infered from the plots of features – 'ApplicantIncome', 'CoapplicantIncome' and 'LoanAmount'

1. All of the feature data are not normally distributed.
2. All of them have right skew and they have outliers (box plot)
3. Data of 'ApptlicantIncome', 'CoapplicantIncome' are segregated by 'Education' and can be observed that, higher number of graduates are with high income (appearing as outliers)

4.

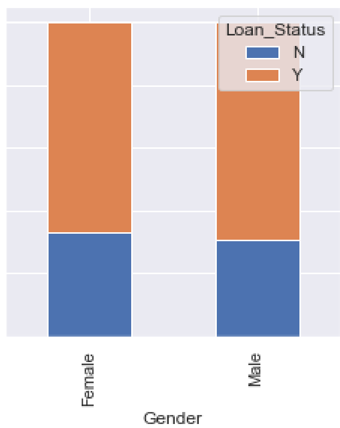

ApplicantIncome
CoapplicantIncome
LoanAmount
ApplicantIncome
CoapplicantIncome
LoanAmount

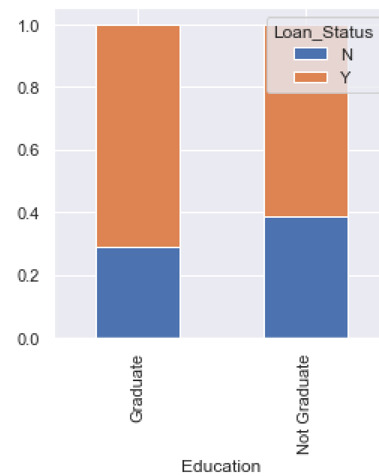

ApplicantIncome
Graduate
Not Graduate
Education
CoapplicantIncome
Graduate
Not Graduate
Education

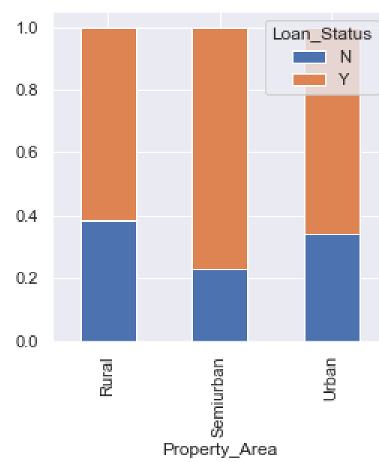
BiVariate Analysis

We are interested in finding how the features are affecting the target variable 'Loan_Status'. So, will try to plot the categorical, numerical features against target variable to draw inference

Categorical Variable vs **Target Variable**

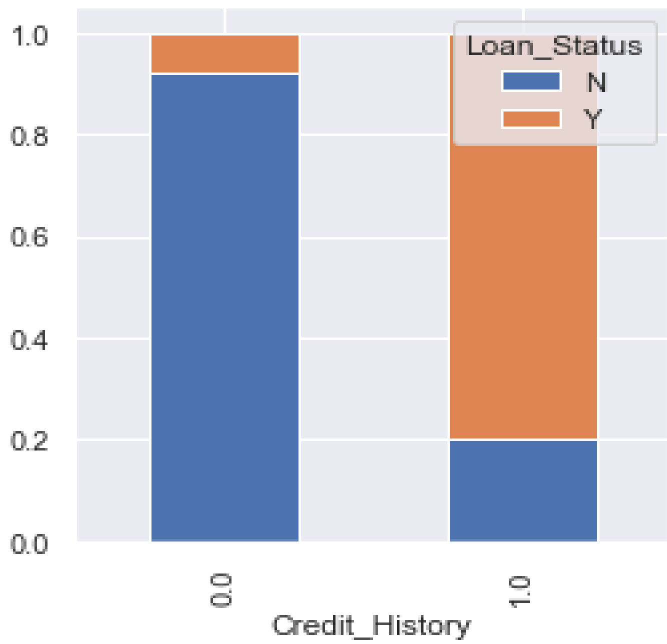






From above plots we can infer that

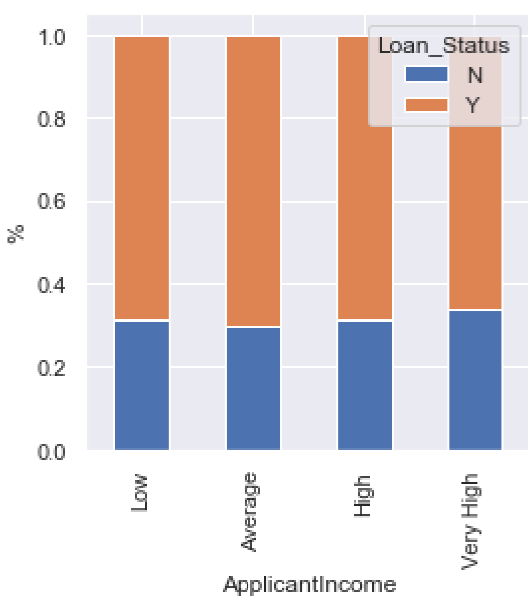
1. Proportion of male and female applicants are almost same for approved/unapproved loans
2. No much difference between proportion of self-employed/non self-employed for approved/unapproved loans
3. Married applicants have higher proportion of approved loans
4. Proportion of loan approvals is more in case of Graduate applicants
5. Proportion of loan approvals is more in SemiUrban area
6. Loan approvals are distributed equally for dependents 1, 3+. Proportion of loan approvals is high in case of dependents 2.

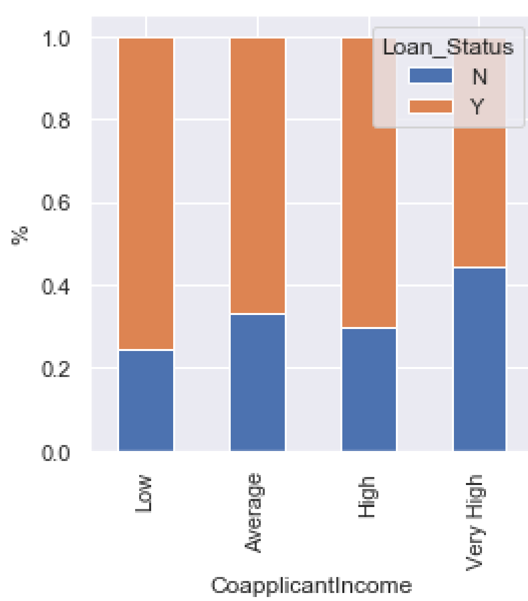


From above plot it looks like, applicants with Credit_History good (1), have very high chances of loan approval.

Numerical Variable vs **Target Variable**

Now, will try to find how loan approvals are distributed related to income. Will bin the income into categories and plot them against the loan approvals



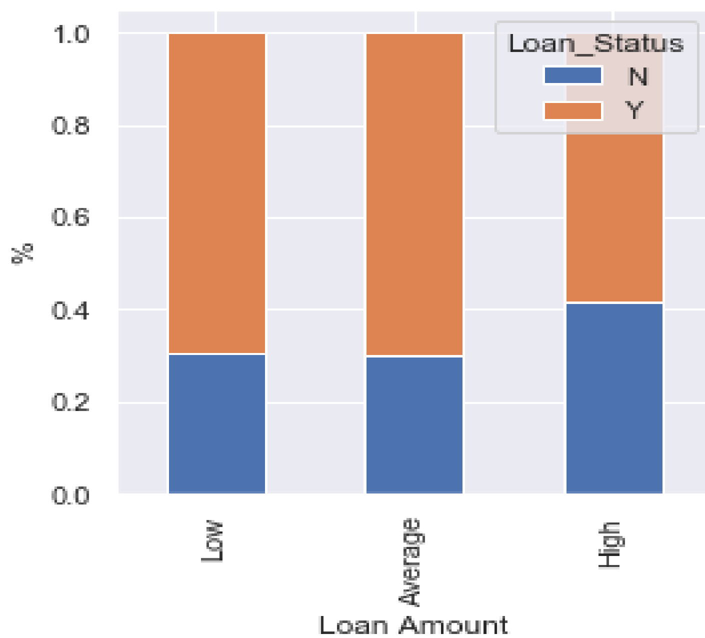


1. From ApplicantIncome plot, it can be observed that the distribution of loan

approvals doesn't vary much b/w income groups

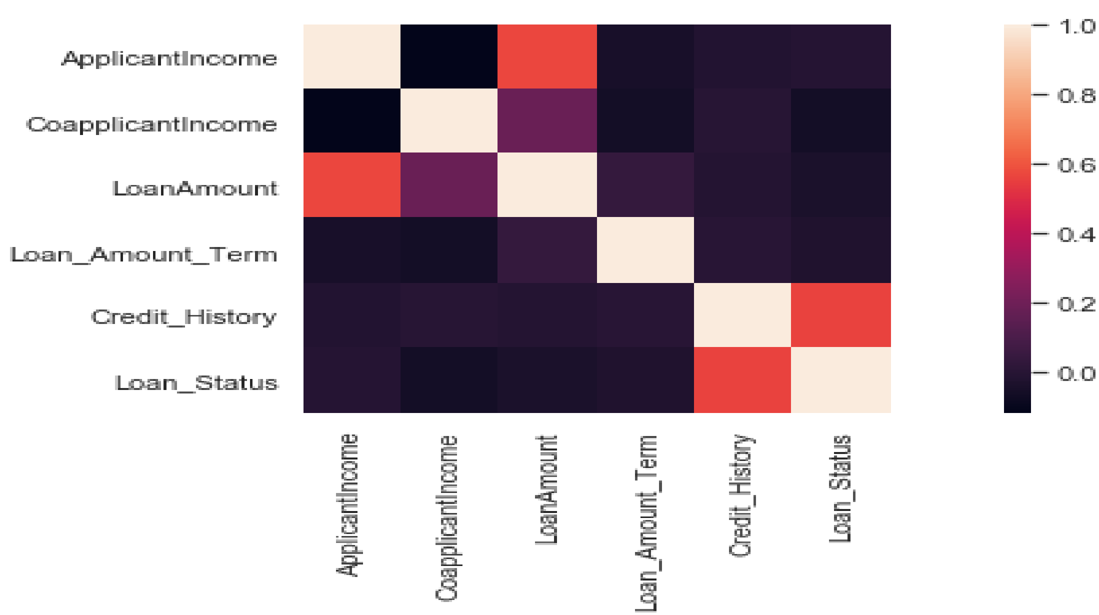
2. CoapplicantIncome shows that with "Low" income, chances of Loan Approval is high. This is against our hypothesis that high income leads to high loan approvals. The reason for this behavior is, CoapplicantIncome is 0 in many cases (314/614) and indicates that many of Applicants doesn't have CoApplicants. So, quite possible that this feature may not be an key dependent for Loan Approvals.
3. We will derive new feature 'TotIncome" which is sum of " ApplicantIncome" and "CoapplicantIncome" and see the effect on Loan Approvals. It can be observed that Low income leads to low Loan Approvals, and Average and High have better chances of Loan Approvals.

Analyze LoanAmount data: From the plot it is clear that Lower the LoanAmont higher are chances for the Loan Approvals, which supports our hypothesis.



Correlation map

It will be interesting to find the correlation of numerical features against the target "Loan_Status". Since only train set has the target "Loan_Status", will work on correlation on train set



Observe the row 'Loan_Status', it seems it is strongly correlated to 'Credit_History'. 'Loan_Amount' is strongly correlated to 'ApplicantIncome' and also fairly correlated to 'CoapplicantIncome'. Let us print the correlated values of Loan_Status in descending order:

Loan_Status	**1**		**LoanAmount**	**1**
Credit_History	0.561678		ApplicantIncome	0.570909
ApplicantIncome	-0.00471		CoapplicantIncome	0.188619
Loan_Amount_Term	-0.021268		Loan_Amount_Term	0.039447
LoanAmount	-0.037318		Credit_History	-0.008433
CoapplicantIncome	-0.059187		Loan_Status	-0.037318

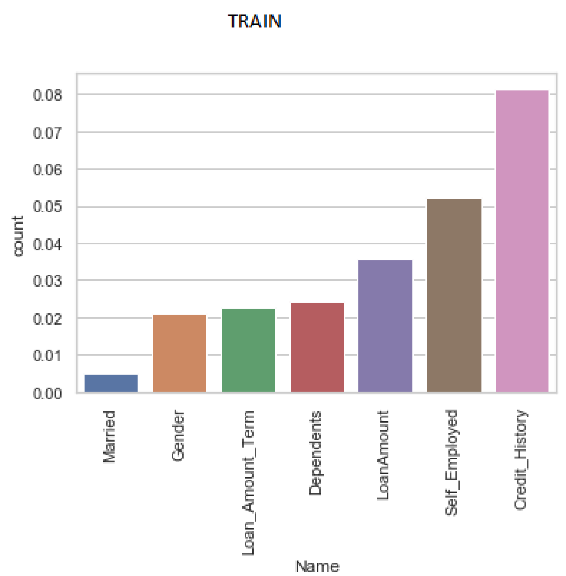
Missing Values

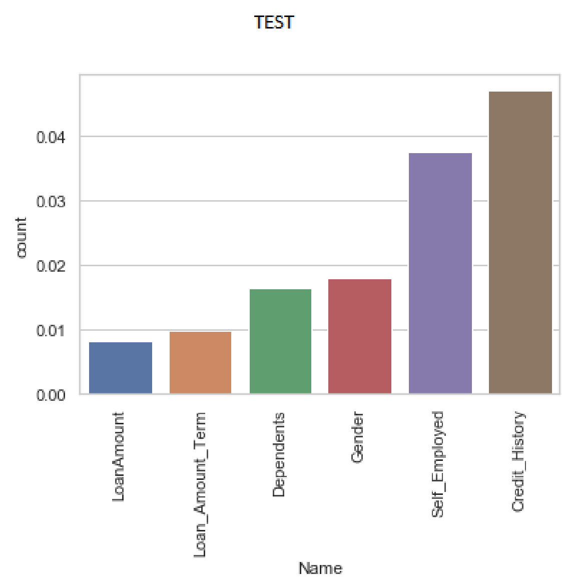
Check the missing values of Train and Test sets

TRAIN		
Features (Missing Vals)	**%**	**Count**
Credit_History	0.081433	50
Self_Employed	0.052117	32
LoanAmount	0.035831	22
Dependents	0.024431	15
Loan_Amount_Term	0.022801	14
Gender	0.021173	13
Married	0.004886	3

TEST		
Features (Missing Vals)	**%**	**Count**
Credit_History	0.047231	29
Self_Employed	0.037459	23
Gender	0.017915	11
Dependents	0.016287	10
Loan_Amount_Term	0.009772	6
LoanAmount	0.008143	5
Married	0	0

Notice that "Credit_History" has the highest missing values (Train – 8.14%; Test – 4.72%).





For missing values treatment, will consider below strategy:

1. Numerical Variables – imputation with mean/median
2. Categorical Variables – imputation with mode

Gender, Married, Dependents, Credit_History and Self_Employed variables are imputed with mode.

The numerical variable 'Loan_Amount_Term' is also imputed with mode, since value count for '360' is highest (most common)

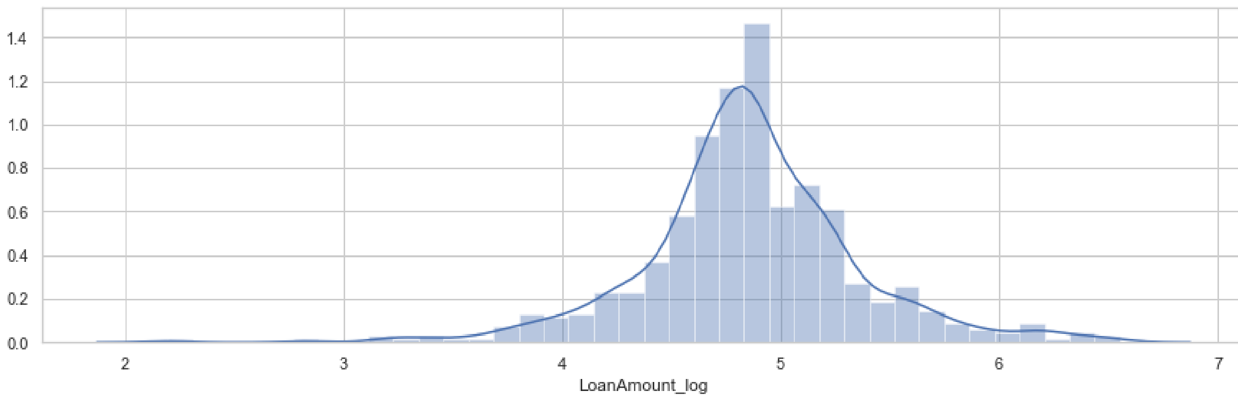
The numerical variable 'LoanAmount' is imputed with median, because of outliers.

Combining Train and Test

To make pre processing steps easy and uniform, will combine both Test and Train data into an single data frame.
Later once pre processing is completed, will split this data into Test and Train.

Outliers Treatment

Outliers have very significant effect on mean and standard deviation, which affects the distribution. We have seen that LoanAmount has very large number of outliers, will treat them with scaling using “minmaxscaler”



Sometimes, outlier treatment using algorithms doesn't provide good results. In that cases, we may need to manually verify the train data set and make modifications manually. For example:
in Loan_ID LP002317, ApplicantIncome is 81000, and area is rural. Quite possible that this could be a mistake. Changed “81000” to “8100”
in Loan_ID, Loan_Amount was “9” (typo mistake), changed to “99”
in LP002949, CoapplicantIncome is too high in comparison to ApplicantIncome. The CoapplicantIncome could be a typo error, so changed from “41667” to “4167”

Building Models

1. Logistic Regression

Logistic Regression is

1. Binary classification algorithm, since Loan Prediction is binary clarification issue will use this modeling
2. Estimation of logit function which is log of odds in favor of an event
3.

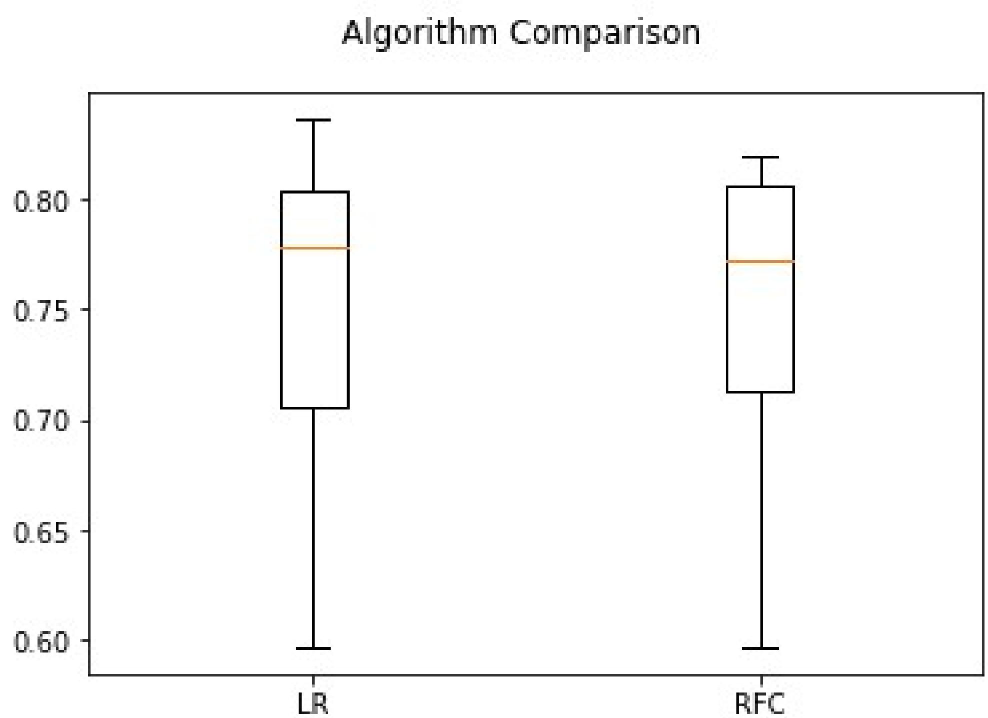
2. Random Forest Classification

RFC is

1. Classification algorithm which has better performance in controlling over-fitting
2. a meta estimator that fits a number of decision tree classifiers on various sub-samples of the data-set and uses averaging to improve the predictive accuracy and control over-fitting.

Did an initial comparison of both the algorithms. From the below picture it looks like LR accuracy is better compared to RFC. But, with unseen data (test data), it is observed that RFC algorithm performs better.

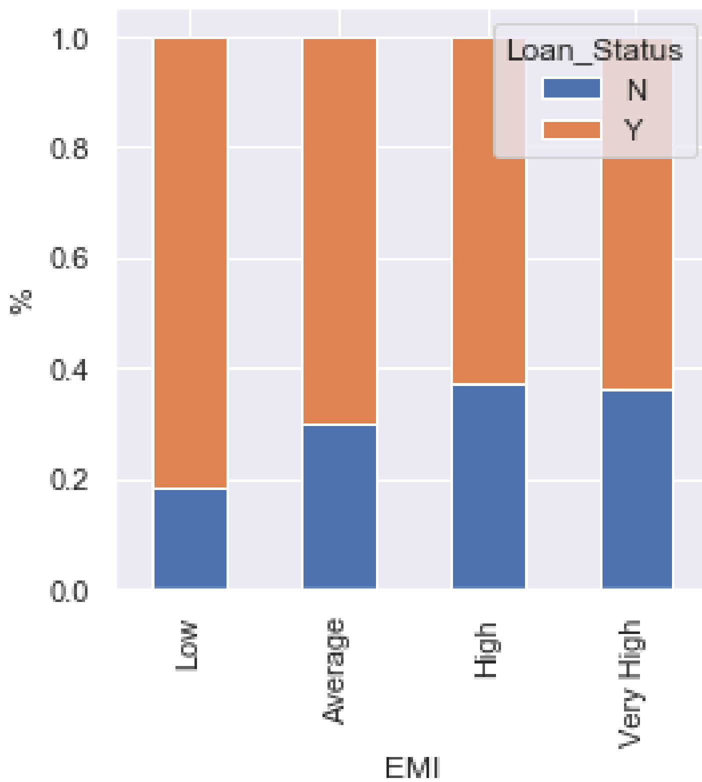
So will work on RFC algorithm to solve the Loan Prediction problem.



Feature Engineering

During the BiVariate analysis we had derived new feature “TotIncome” and observed that it justifies our hypothesis.

Will create new feature “EMI” and below plot 'EMI' vs 'Loan_Status' confirms that lower the EMI higher are chances of Loan Approval. It justifies our hypothesis.



Results

Accuracy with trian data : 0.7529878371232153

Acuuracy with test data (Score from submission): 0.8125