Loan Prediction Report

INDEX

1. About Company………………………………………………………………………1
2. Problem……………………………………………………………………………….1
3. Goal…………………………………………………………………………………...1
4. Issues With Solution………………………………………………………………….2
5. Analysis of Data………………………………………………………………………2
6. Predictive Modeling…………………………………………………………………..2
7. Data…………………………………………………………………………………...3
8. Hypothesis…………………………………………………………………………….3
9. Univariate Analysis…………………………………………………………………...5
10. Bivariate Analysis…………………………………………………………………….9
11. Missing Data and Outlier Treatment………………………………………………...20
12. Feature Engineering…………………………………………………………………20
13. Regression Modeling………………………………………………………………..22
14. Model Improvement………………………………………………………………...23

About Company

Dream House Finance is amongst the leading providers in the housing loan industry in New Delhi. They have presence across all urban, semi urban and rural areas. A customer first applies for a home loan, after which the company validates the customer’s eligibility for the loan.

Problem

The company wants to automate the loan eligibility process (in real time) based on a customer’s details provided while filling in an 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 want to use their existing data from previous customers in order to create a model which can help them in doing so. They have provided us with a partial data set in order to simplify the problem for us.

Goal

The goal of this loan prediction problem is to create a model using their existing data in order to automate the process of loan approval. Once this process is automated, they can save time by specifically targeting these customers, as well as several man hours on reviewing non prospective customer details. First I will start by conducting a thorough analysis the of the dataset in order to identify which of the customer details (variables) have the highest correlation to the loan approval. Following which I will create this predictive model using different forms of regression and finally choosing the model with the highest accuracy rate.

Issues With the Solution

Regression modeling is a great way to make predictions. First off, we need a big enough data set for the model to be accurate. Second, we need to manage our expectations about the accuracy of the model. No model will be extremely reliable in its early stages. All such models mature in time as they are trained with larger amounts of incoming data. The con about making a predictive model in such a situation can be very heavy. Consider the situation where the model makes a wrong prediction and gives a thumbs up for loan approval to a customer that may not be able to pay back the loan. In such a case, the company will lose a substantial amount of money. On the other hand, if the model rejects a customer that may have been able to pay back the loan, the company stands to lose out on a big margin of profits.

Analysis of Data

First, we cover a univariate analysis of all the variables in order to better understand who our customers are. Through analysis each of the variables individually, we get a better picture about what kind of people are applying for loans. For example, we can tell that most applicants are males who are married, are not self-employed, do not have children, are graduates and have a good credit history. Through simple analysis of each of the variables individually we were able to know a lot about our customer base. Next, we cover a bivariate analysis of all the customer details against the loan status variable. This gives us an idea about which of the customer details are most correlated to the loan approval. After a thorough analysis of the variables, I inferred customer details which were highly correlated to the loan approval – applicant income, co-applicant income and credit history. These details can be used to create additional statistics such as Total Income (Applicant Income + Co-applicant Income), EMI (Loan Amount / Loan Time in Months) and Balance Income (Total Income - EMI). These variables constituted as an integral factor in loan approval only second to the Credit History Variable.

Predictive Modeling

The first step in creating a predictive model using regression is to figure out what sort of regression fits our dataset the most. Since our target variable is binary in nature and is dependent on several independent variables, my first pick is logistic regression. By splitting the training data into two halves – training and testing we are able to create the model using the training half and validate the accuracy of our model on the testing half. We get an accuracy of ~80% which is a great number for a new, immature model. To further improve the accuracy of the model without the use of extra data or time for the model to mature, we use the stratified k fold cross validation technique. This improves our model’s accuracy to 81% which may not seem like much, but is a big jump without the maturation of our model.

Data

|  |  |
| --- | --- |
| **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) |

Hypothesis

Null Hypothesis (H\_0) : The loan approval does not depend on 1) Gender, 2) Marital Status, 3) Number of Dependents, 4) Education Level, 5) Self Employment, 6) Applicant Income, 7) Co-applicant Income, 8) Loan Amount, 9) Loan Term, 10) Property Area and 11) Credit History.

Alternate Hypothesis (H\_1) : The loan approval depends on (some of) the factors listed above.

These are some of the factors I think will affect the Loan Status approval:

Education Level: Applicants who are graduates should have higher chances of loan approval

Applicant/Co-applicant Income: Applicants with high income should have more chances of loan approval.

Credit history: Applicants who have repaid their previous debts should have higher chances of loan approval.

Loan amount: Loan approval should also depend on the loan amount. If the loan amount is less, chances of loan approval should be high.

Loan term: Loan for less time period and less amount should have higher chances of approval.

After conducting bivariate analysis on each of these 11 fields against the Loan Status field from our training data we will be able to better say which of these fields does the Loan Status (correlate) depend on.

Python Output (analyzing data fields):

Training file data types:

Loan\_ID object

Gender object

Married object

Dependents object

Education object

Self\_Employed object

ApplicantIncome int64

CoapplicantIncome float64

LoanAmount float64

Loan\_Amount\_Term float64

Credit\_History float64

Property\_Area object

Loan\_Status object

dtype: object

Testing file data types:

Loan\_ID object

Gender object

Married object

Dependents object

Education object

Self\_Employed object

ApplicantIncome int64

CoapplicantIncome int64

LoanAmount float64

Loan\_Amount\_Term float64

Credit\_History float64

Property\_Area object

dtype: object

Training file dimensions:

(614, 13)

Testing file dimensions:

(367, 12)

Loan Status Frequency:

Y 422

N 192

Name: Loan\_Status, dtype: int64

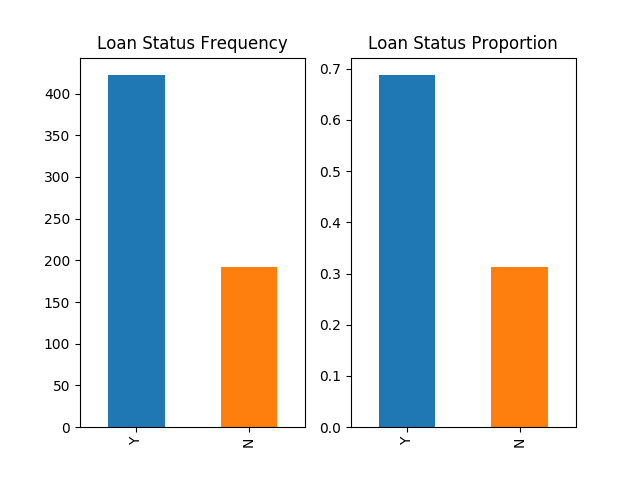
Loan Status Proportion:

Y 0.687296

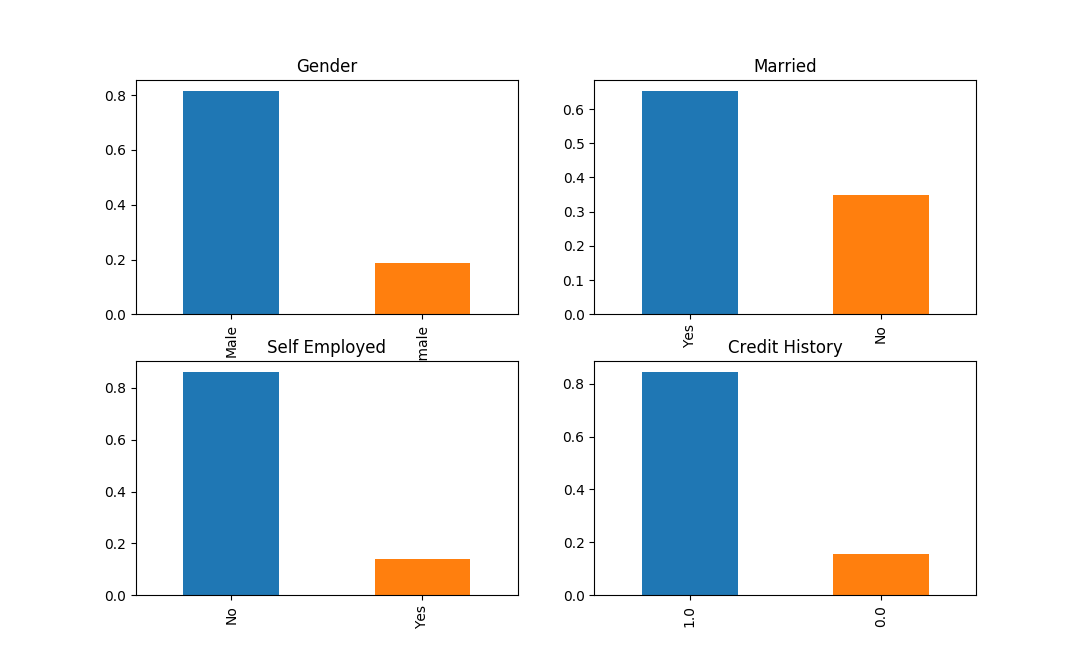
N 0.312704

Name: Loan\_Status, dtype: float64

Univariate Analysis



It can be inferred that ~69% of loans are approved and ~31% of loans are denied.

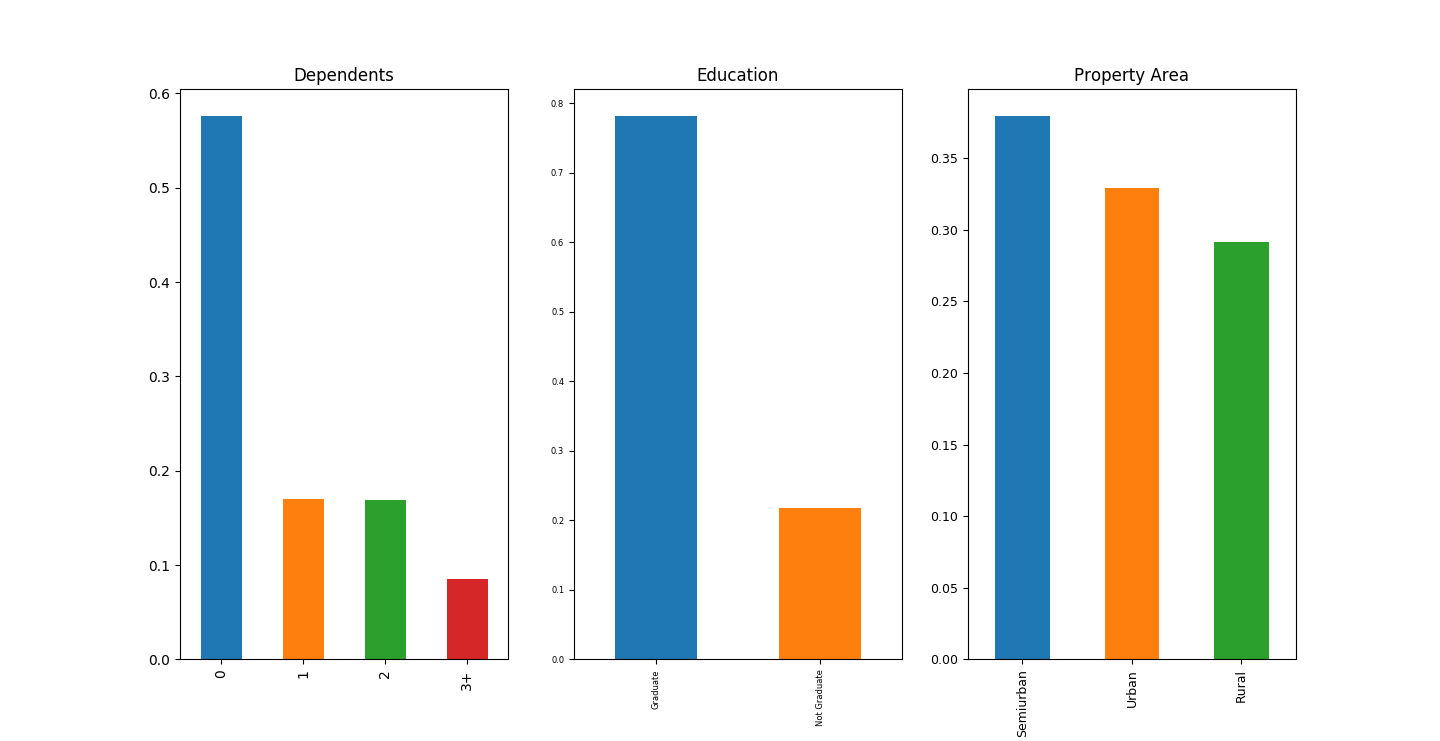


It can be inferred that ~80% of applicants are male and ~20% are female.

It can be inferred that ~65% of applicants are married and ~35% are married.

It can be inferred that ~85% of applicants are not self-employed and ~15% are self-employed.

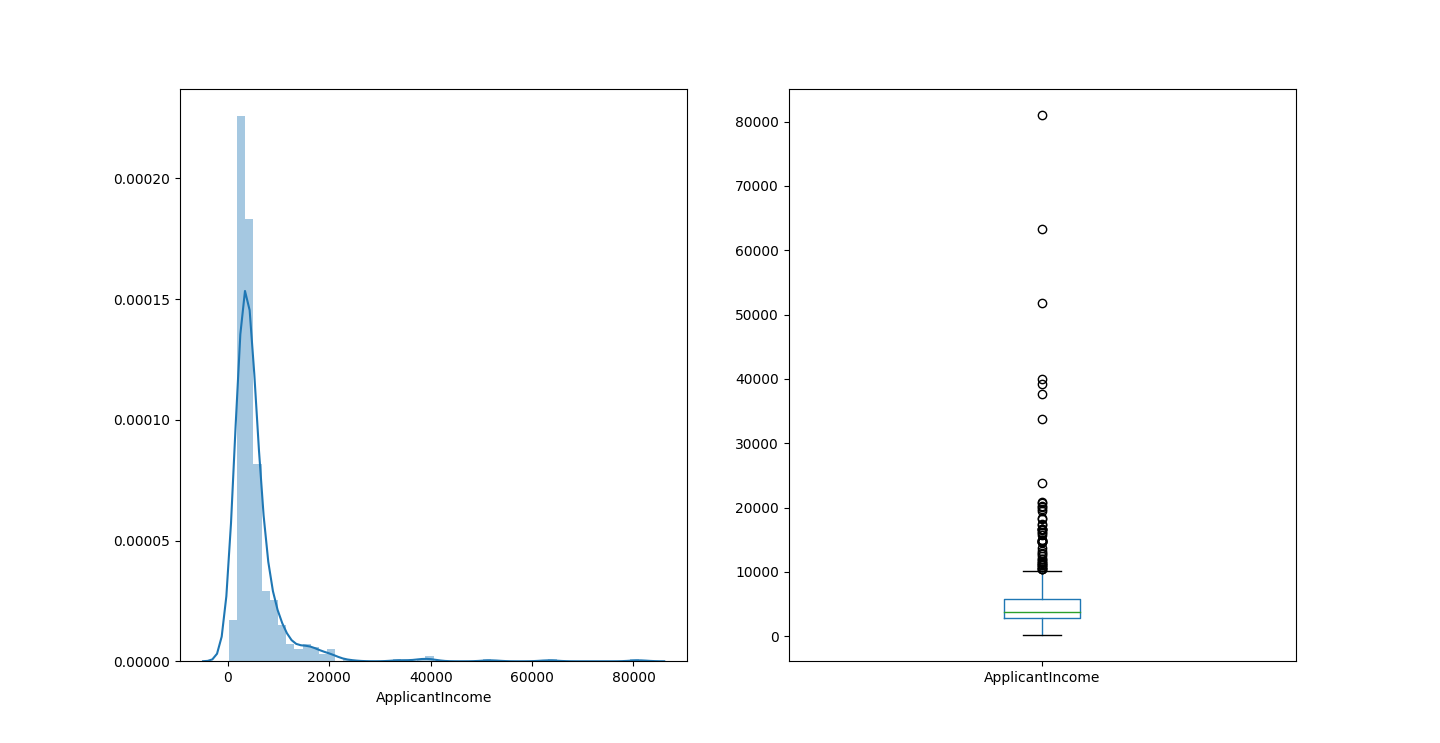
It can be inferred that ~85% of applicants have repaid their debts and ~15% have not.



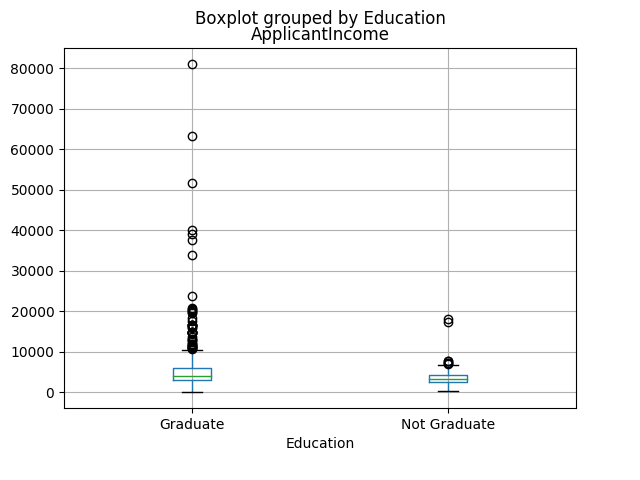
It can be inferred that majority of the applicants have no dependents.

It can be inferred that ~78% of the applicants are graduates and ~22% are undergraduates.

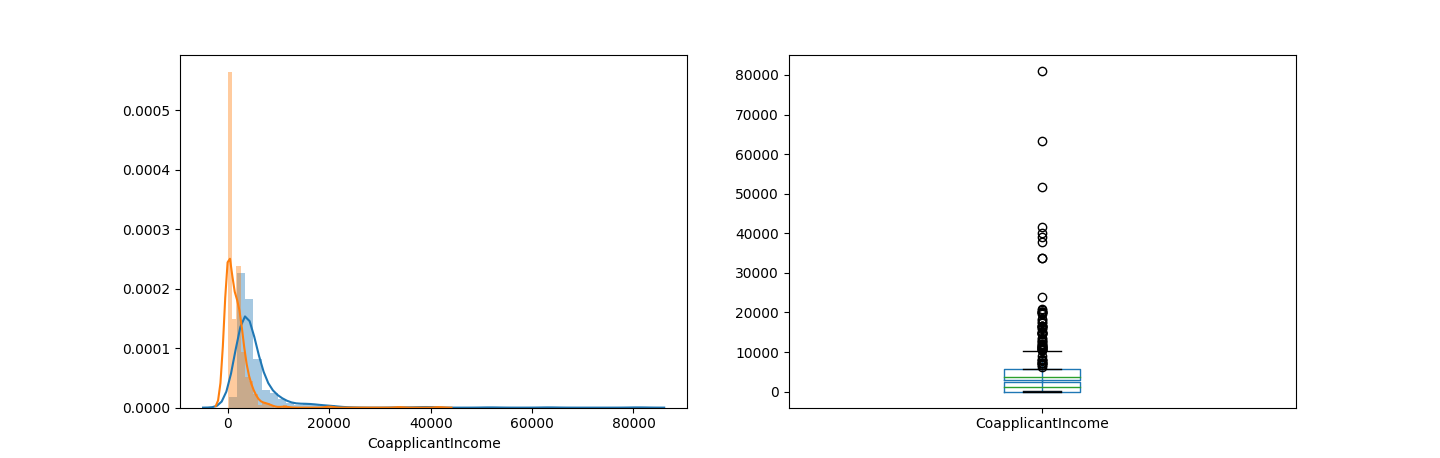
It can be inferred that majority of the applicants are from suburban areas (although urban and rural are not that far behind).



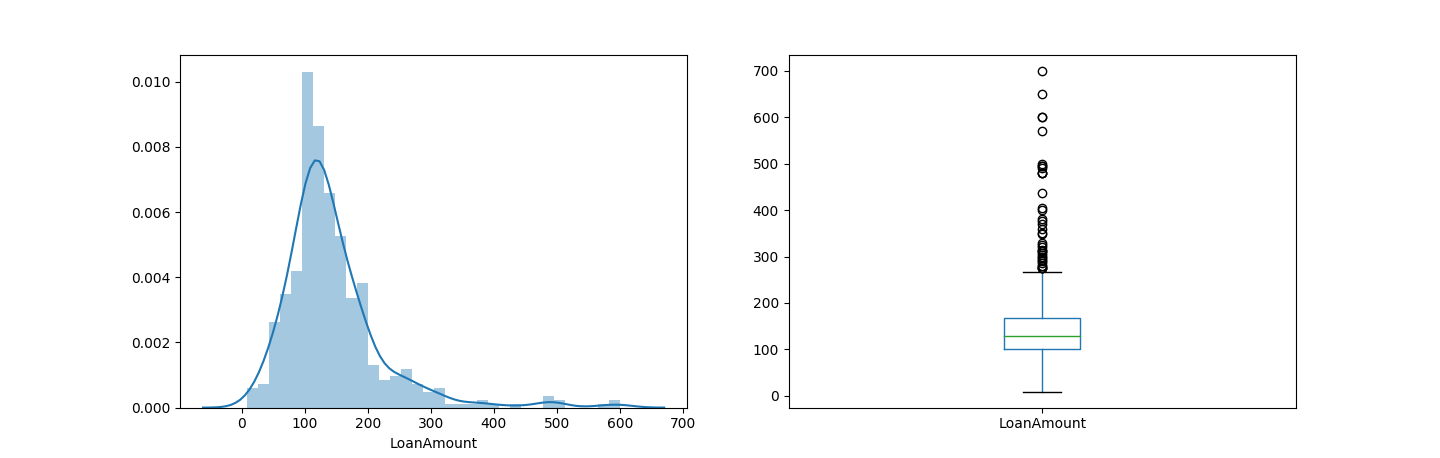
The data looks heavily skewed to the left and does not look normal. The box plot tells us that there are a significant number of outliers which we will have to take into consideration. This could be due to difference in education level. To confirm this, we will plot the applicant income according to education level down below.



As we can see, the vast majority of the outliers are lie in graduate applicants. Thus the high ‘outlier’ income can be attributed to the education level.

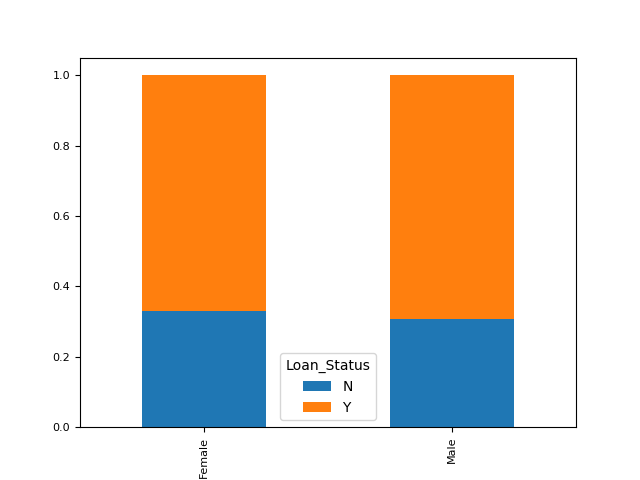


Similarly, the co-applicants income is not normally distributed and has a lot of outliers as well.

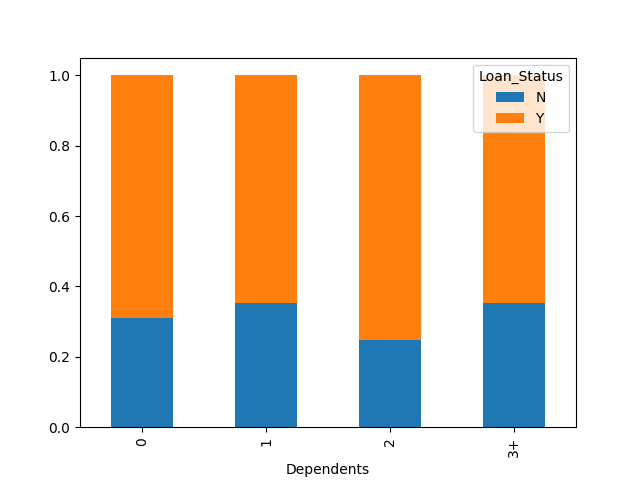


Although not perfectly normal, the Loan Amount distribution is fairly normal compared to the applicant and co-applicant income distributions. There are a good amount of outliers here too.

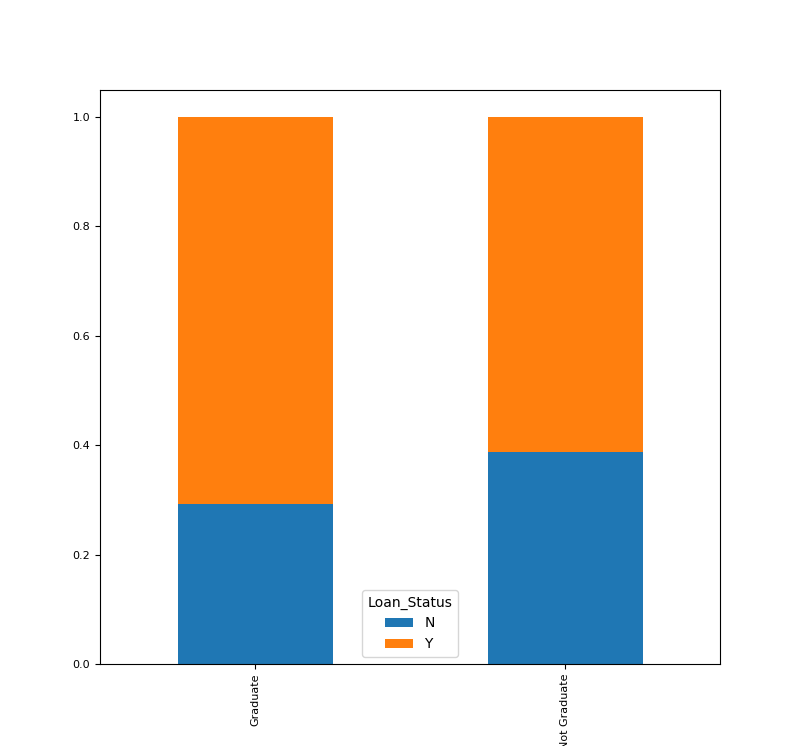
Bivariate Analysis



Since ~30% of both males and females are rejected we can infer that Loan Status approval is not effected by gender.

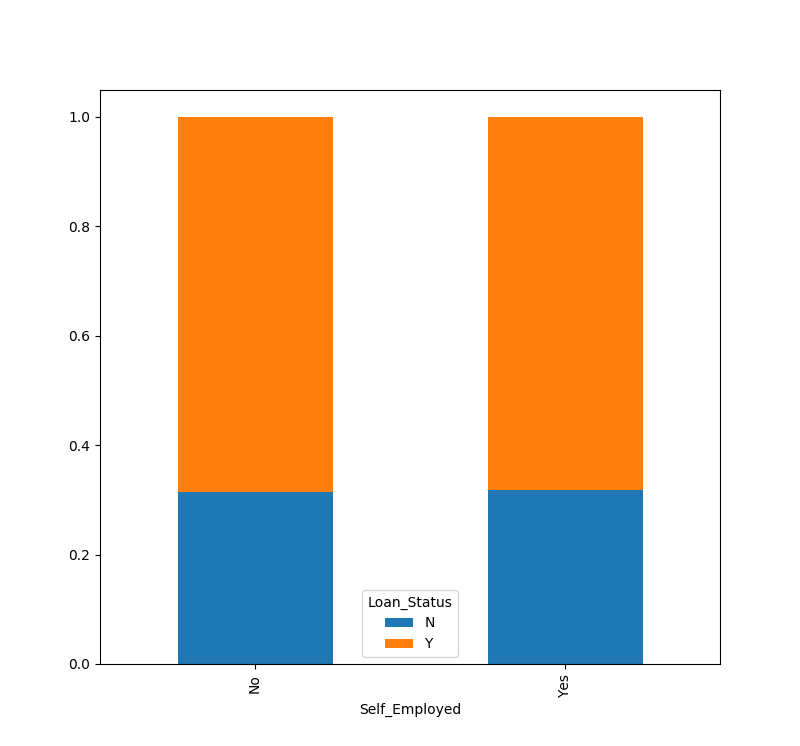
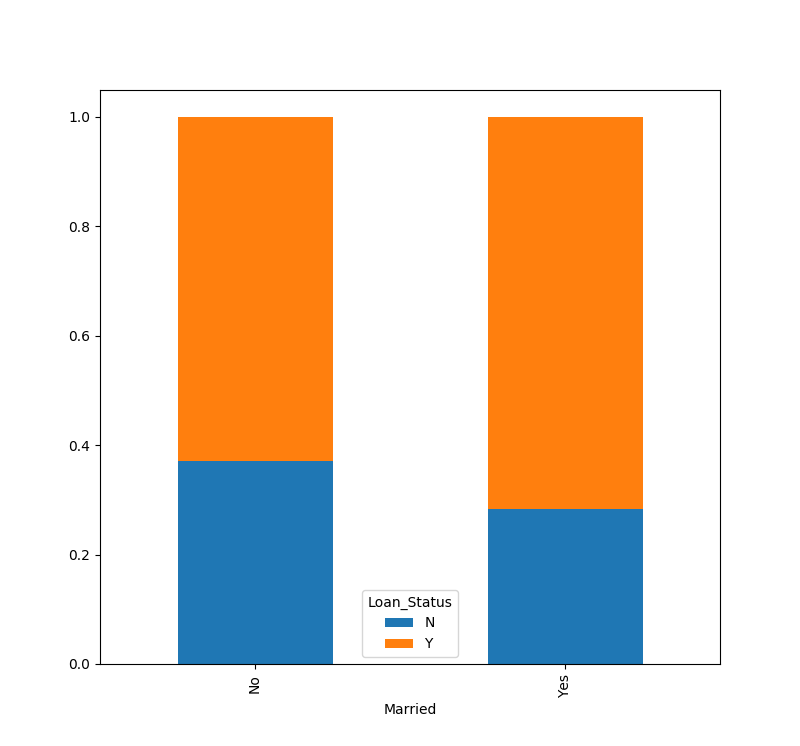


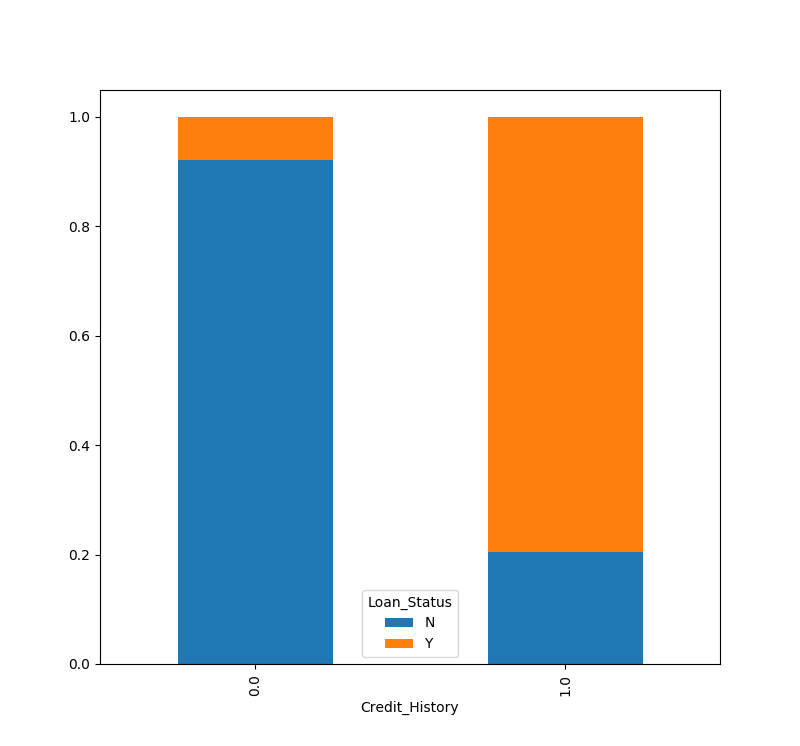
The Loan Status approval rate seems to be higher (~70%) for applicants with 0 or 2 dependents while it seems to be lower(~65%) for applicants with 1 or 3+ dependents. We can infer that there is some sort of relationship (although unclear) between Dependents and Loan Status.



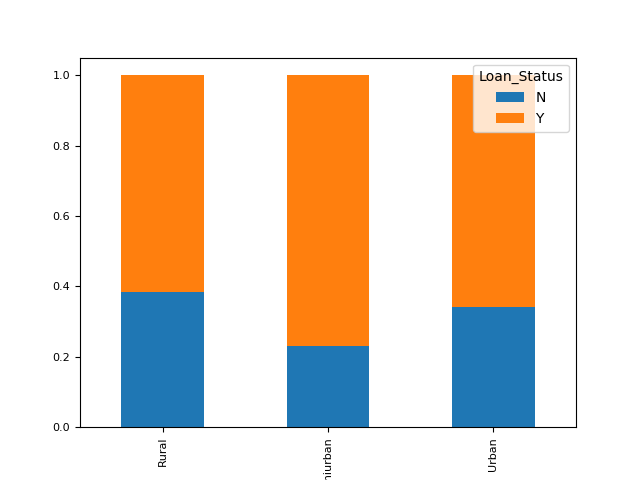
Education

The Loan Status approval is significantly higher (~70%) for graduate applicants rather than undergraduate applicants. We can infer that there is a direct correlation between Education and Loan Status.

The Loan Status approval is the same for self-employed and non elf employed applicants. We can say that there is no correlation between self-employment and Loan Status.The Loan Status approval rate seems to be higher (~70%) for married applicants rather than single applicants. We can infer that Loan Status is correlated to marital status.

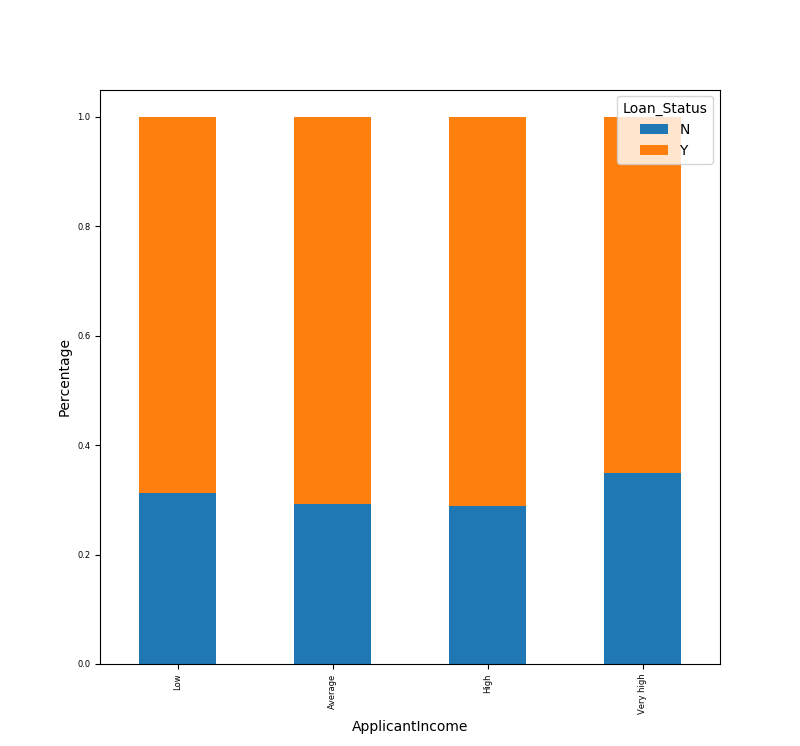


People who have paid back previous debts have a significantly higher rate of Loan Status approval (~80%). We can say that Credit History and Loan Status are highly correlated.

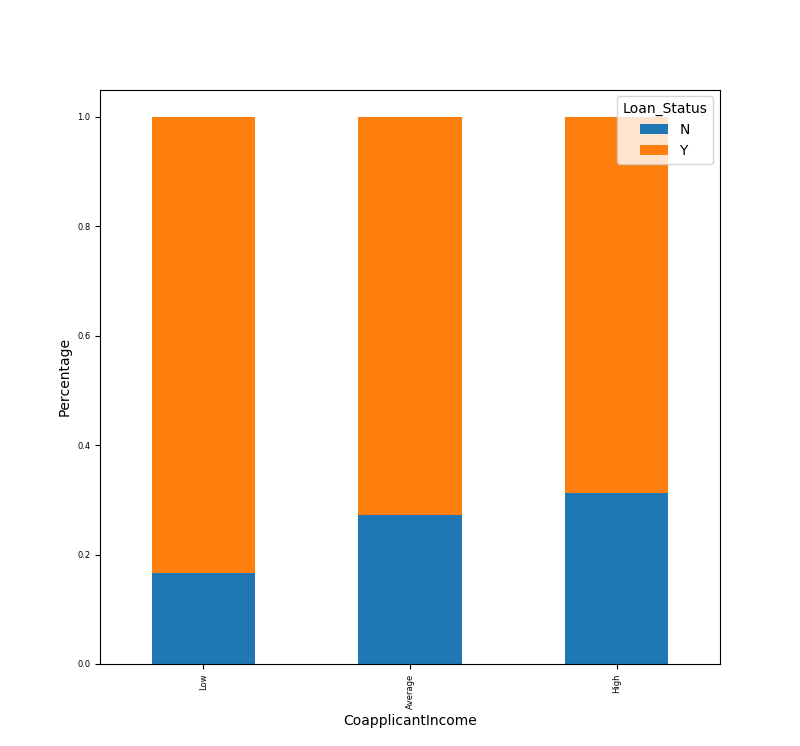


Property Area

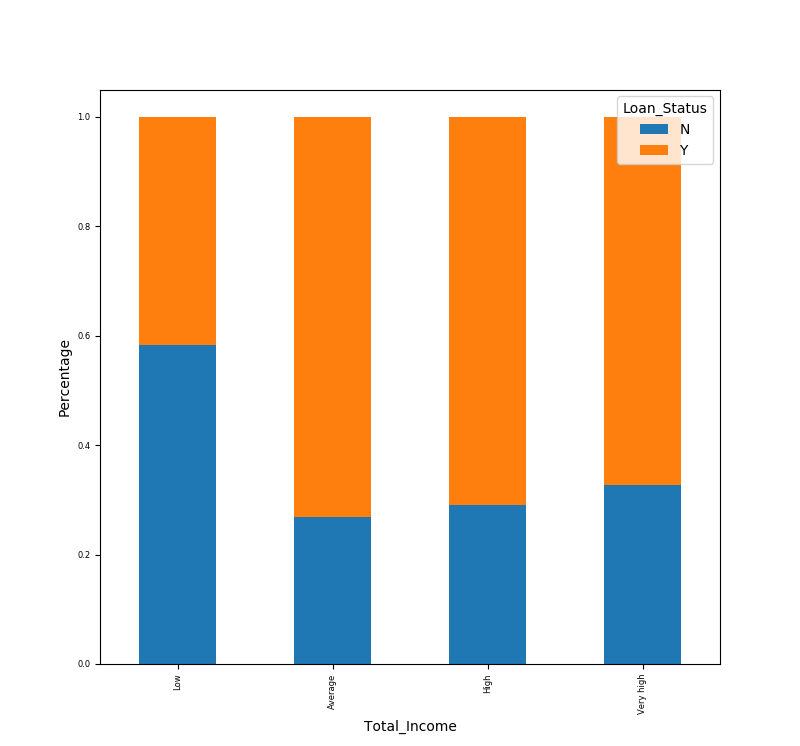
It seems that people living in semi urban areas have the highest (~75%) acceptance rate of Loan Status approval. There is a correlation between property area and loan status.



It seems that applicant income more or less does not affect loan status approval rate which contradicts our hypothesis.



It seems that lower co-applicant income makes for a higher loan status approval rate. But this doesn’t seem right. This may be due to the fact that many co-applicants income is 0, making the low bin of a higher frequency compared to the others. To correct this, let us try to plot the total income instead.



The lowest bracket in the total income has a very low loan status acceptance rate compared to the other 3 brackets.



From this complete bivariate analysis, we can see from the heat map that the most correlated variables with Loan Status are credit history and applicant income.

This concludes our bivariate analysis. But now, we need to take into consideration the missing values and the outliers because they affect our data in a big manner.

Missing Data and Outlier Treatment

Python Output for missing values in our training data set:

Loan\_ID 0

Gender 13

Married 3

Dependents 15

Education 0

Self\_Employed 32

ApplicantIncome 0

CoapplicantIncome 0

LoanAmount 22

Loan\_Amount\_Term 14

Credit\_History 50

Property\_Area 0

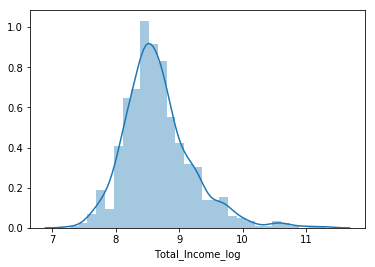
Loan\_Status 0

To fill these missing values, we shall fill the categorical variables using mode and numerical variables using mean. We shall do the same for the testing data set.

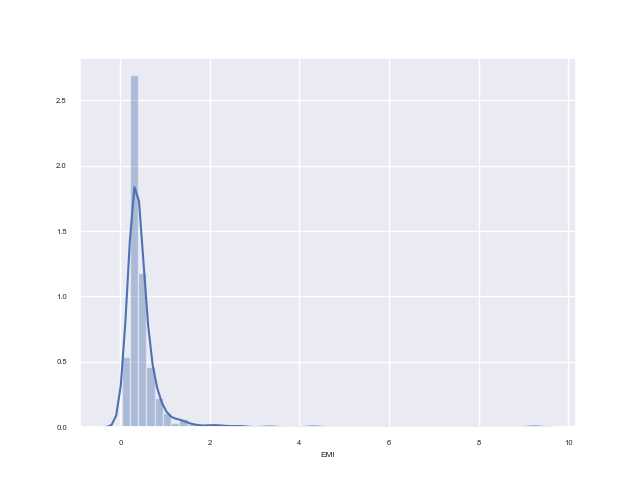
Outliers can significantly throw off the statistics such as mean and standard deviation. It is imperative that we deal with them to avoid skewing our data. To treat these outliers and avoid skewness, we can use logarithmic transformations as taking logs doesn’t affect small numbers too much but it reduces larger values significantly.

Feature Engineering

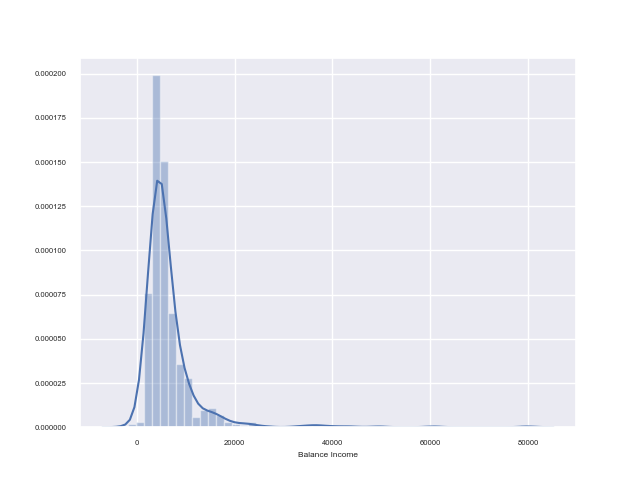
* **Total Income** - We will combine the Applicant Income and Co-applicant Income. If the total income is high, chances of loan approval might also be high.



* **EMI** - EMI is the monthly amount to be paid by the applicant to repay the loan. Idea behind making this variable is that people who have high EMI’s might find it difficult to pay back the loan. We can calculate the EMI by taking dividing the loan amount with respect to loan amount term.



* **Balance Income** - This is the income left after the EMI has been paid. Idea behind creating this variable is that if this value is high, the chances are high that a person will repay the loan and hence increasing the chances of loan approval.



Regression Modelling

In order to test the accuracy of our model, we separate the training data into a training and a testing part. We will make the model based on the training part and then validate the accuracy of our model on the testing part. After which we will be able to approximate what the accuracy of predictions of the actual testing data file will be.

Finally, making a regression model on the training half of our training, we get the an accuracy score of 80% on the testing half of our training data.

Accuracy Score for training data:

0.8054054054054054

Improving the Model / Using Stratified K-Folds Cross Validation

Python Output:

1 of kfold 5

('accuracy\_score', 0.7983870967741935)

2 of kfold 5

('accuracy\_score', 0.8306451612903226)

3 of kfold 5

('accuracy\_score', 0.8114754098360656)

4 of kfold 5

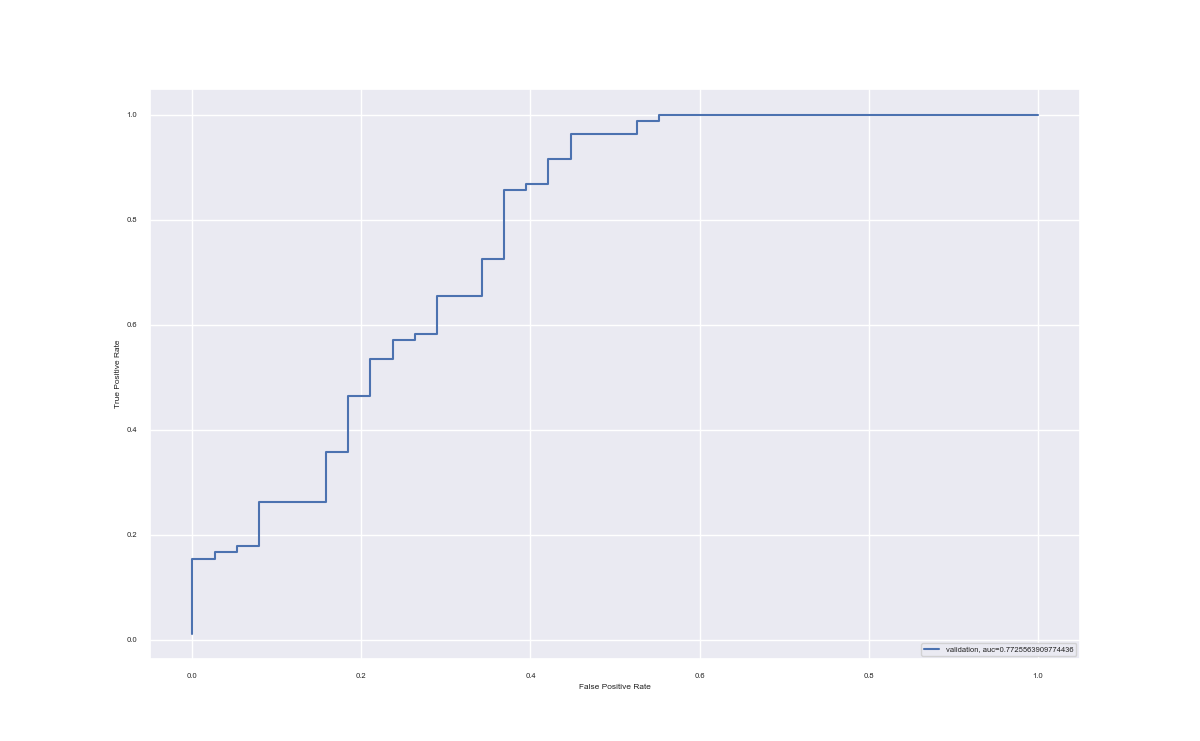
('accuracy\_score', 0.7950819672131147)

5 of kfold 5

('accuracy\_score', 0.8278688524590164)

The mean accuracy score turns out to be ~81%.

Visualizing the ROC curve:



AUC value = 0.77