# Loan Approval Prediction Project

Anagha Kumbharkar


# Problem Statement

The Bank deals in all kinds of home loans. They have a presence across all urban, semi-urban and rural areas. The customer first applies for a home loan and after that, the company validates the customer eligibility for the loan.

The bank wants to automate the loan eligibility process based on customer details. These details are Loan ID, Gender, Marital Status, Education, number of Dependents, Income, Loan Amount, Credit History, and others.

To automate this process, they have provided a dataset to identify the customer that are eligible for loan so that they can specifically target these customers.

# Understanding The Data

- There are 614 observations with 13 fields in our dataset.
- Target label is Loan status (Binary categorical variable).
- Names of 13 columns is as following:

- Loan ID
- Gender
- Marital Status
- Number of Dependents
- Education
- Job type
- Income
- Tenure
- Credit History
- Property Area
- Loan Status

- We have 2 datasets: 1) Train dataset 2)Test dataset with following features.
  Targeted variable is absent in test dataset which we need to predict with model build on train dataset

| Categorical variables(9) | |
|---|---|
| Gender | Male/Female |
| Married | Yes/No |
| Number of dependents | Possible values:0,1,2,3+ |
| Education | Graduate / Not Graduate |
| Self-Employed | Yes/No |
| credit history | Yes/No |
| Property Area | Rural/Semi-Urban/Urban |
| Tenure(in months) | Possible values: |
| Loan Status(Target variable) | Yes/No |

| Numerical variables(4) | |
|---|---|
| Applicant Income | |
| Co-applicant Income | |
| Loan Amount | In thousands |
| Loan ID(Object type) | Unique ID |

# Challenges Faced

- Reading dataset and understanding meaning of some columns.
- While adding new columns we need to have core business knowledge of how banking system works.
- Handling missing values.
- Designing multiple visualization to summarize the information in dataset and drawing conclusions , understanding trends.
- Outlier detection and treatment.
- Changing hyper parameters to improve accuracy of model
- Model selection

# Project Workflow

## PREPROCESSING

Data cleaning, handling null values , missing value imputation, outlier treatment, mapping categorical variable , scaling and normalization.

## EDA

Data Summarization and Visualization for understanding the data distribution, discover patterns with the summary statistics and graphical representations understanding features

## FEATURE SELECTION

Selecting relevant features that contribute to the prediction of loan.

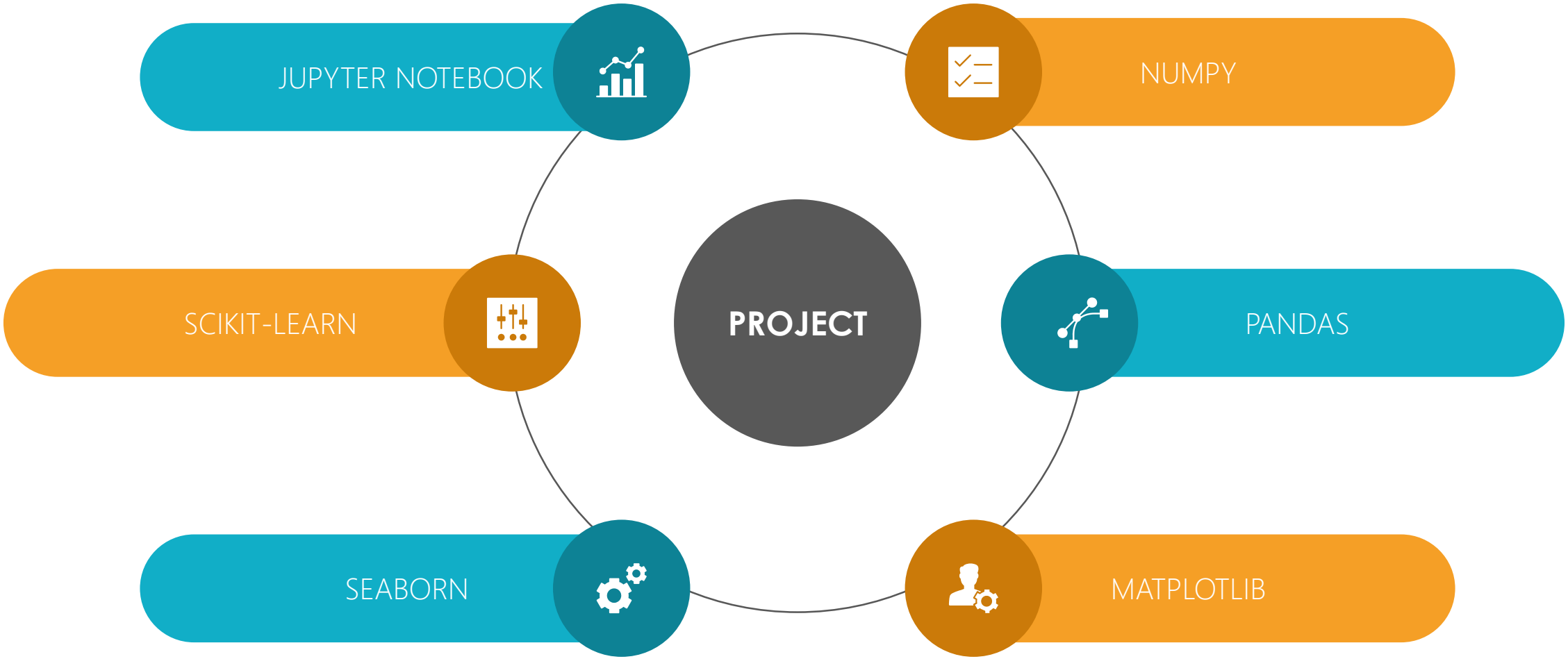
## MODEL SELECTION

Modeling and ensemble modeling. choosing the best generalized model

## EVALUATION

Accuracy score, Cross Validation, Confusion matrix

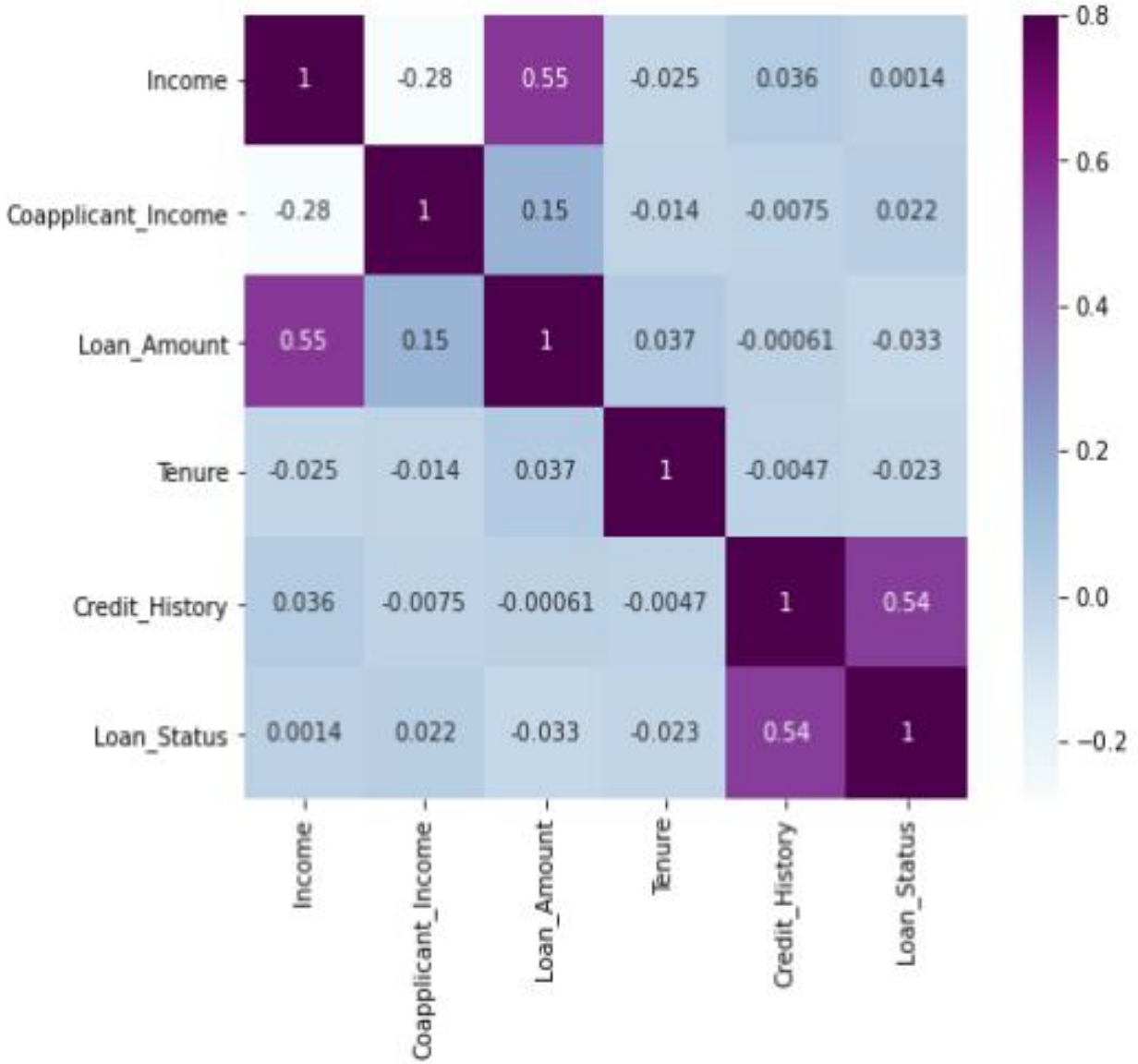
# Tools and packages



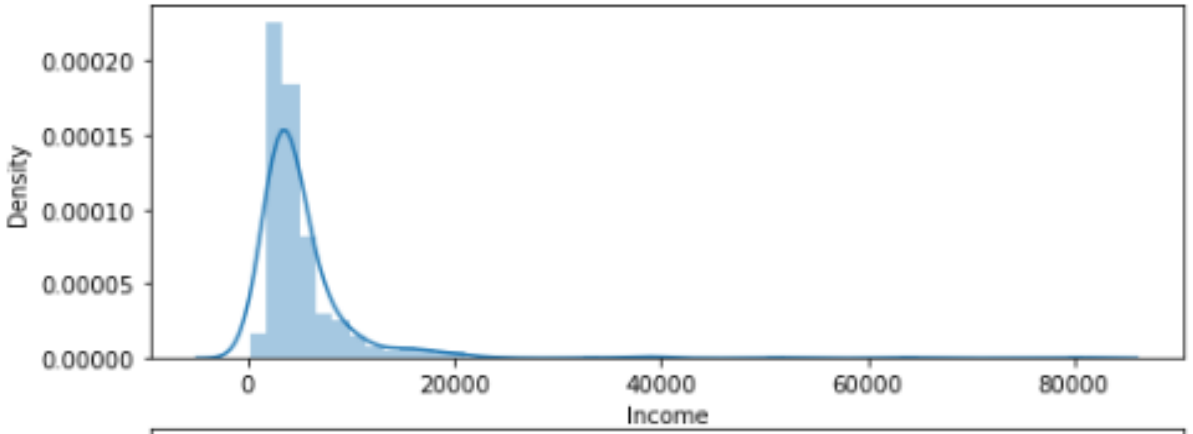
# Exploratory Data Analysis

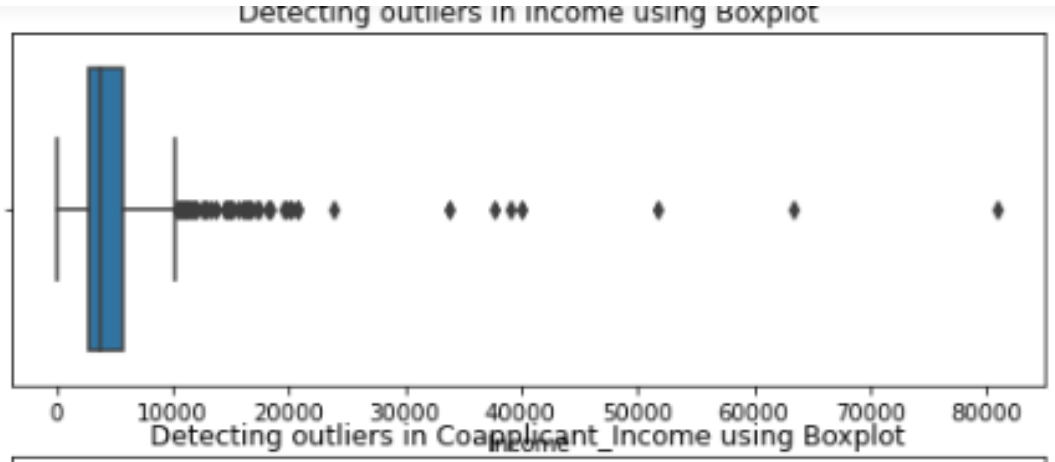
## Correlation Heatmap

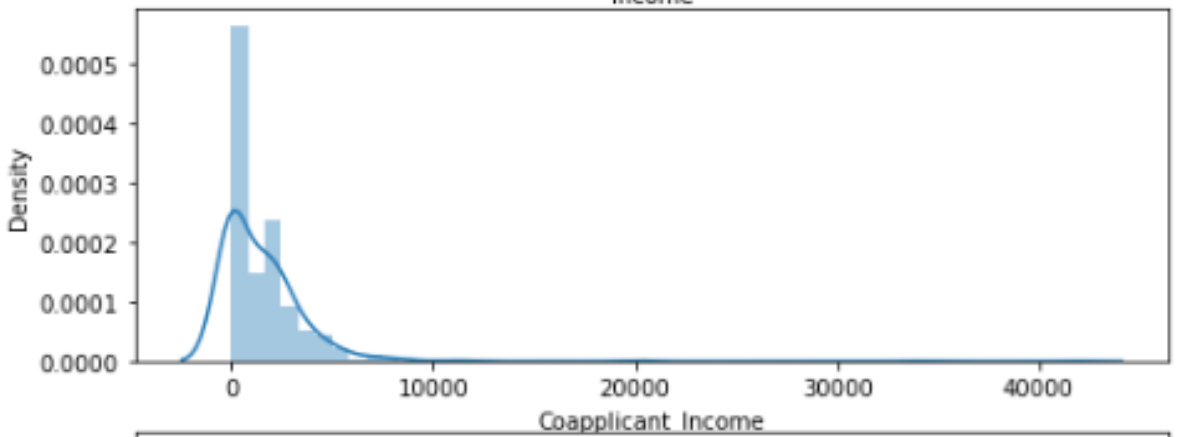
- The most correlated variables are Income - Loan_Amount and Credit_History - Loan_Status.
- Loan_Amount is also correlated with Coapplicant_Income.

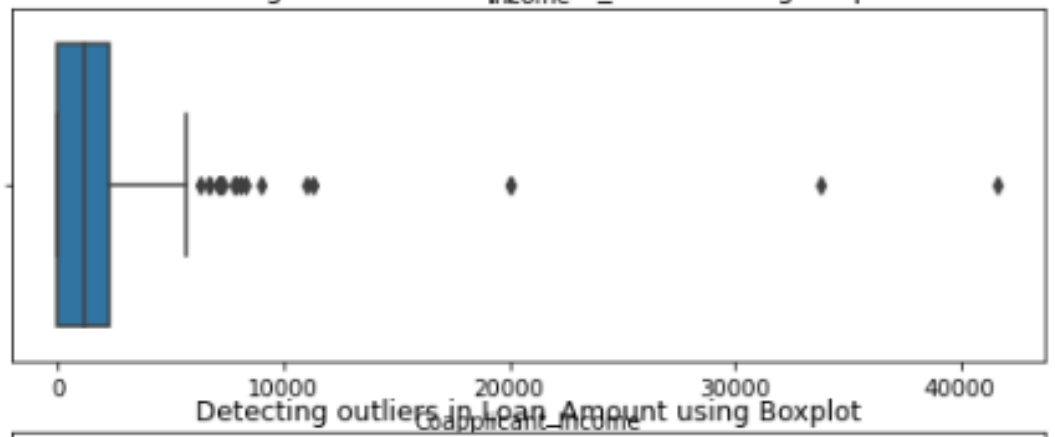


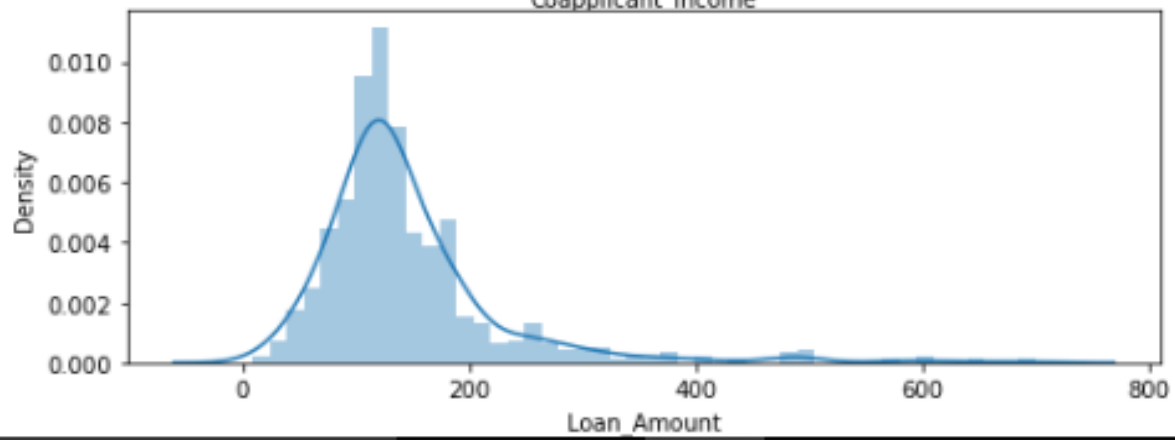
# Continuous variable and outliers

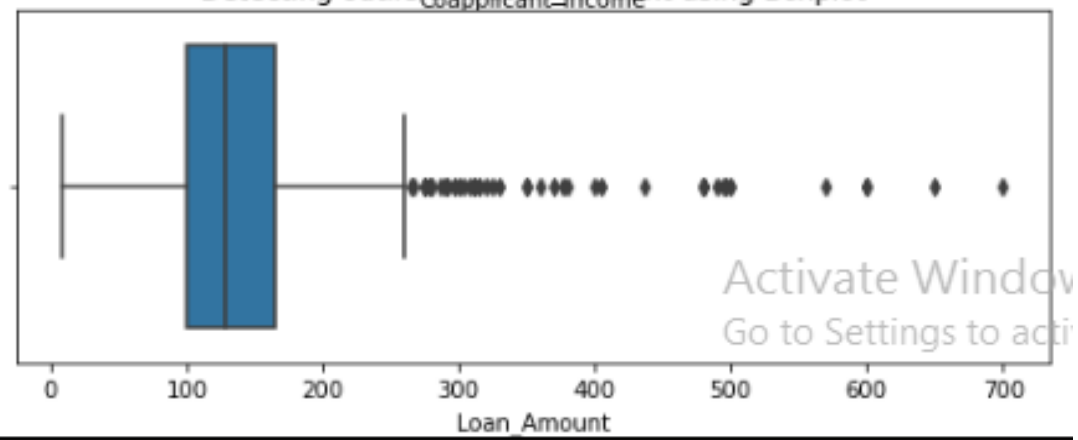


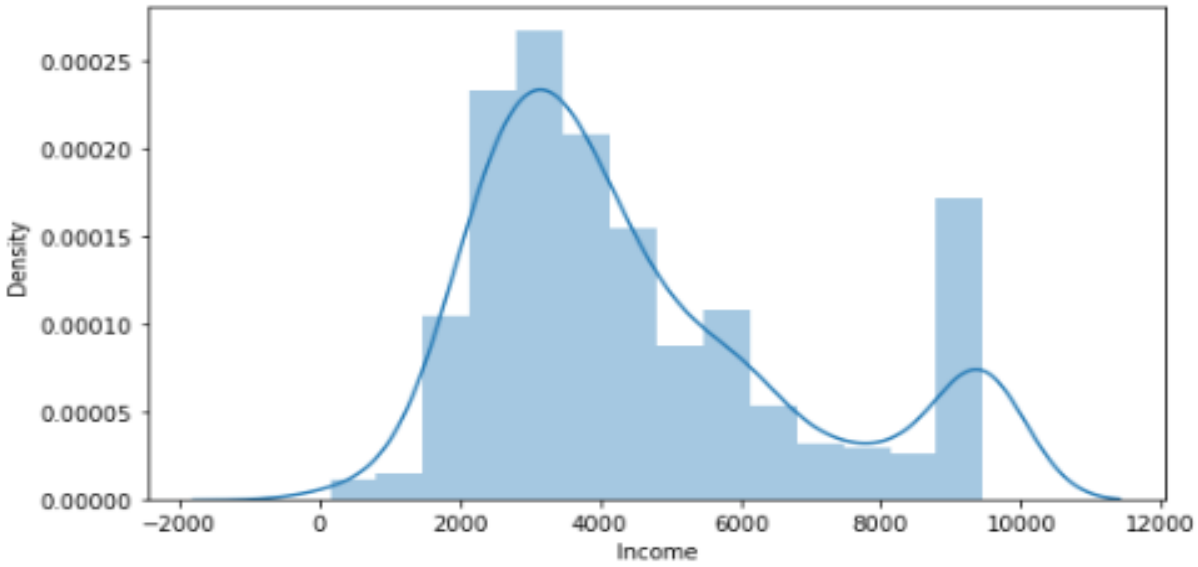



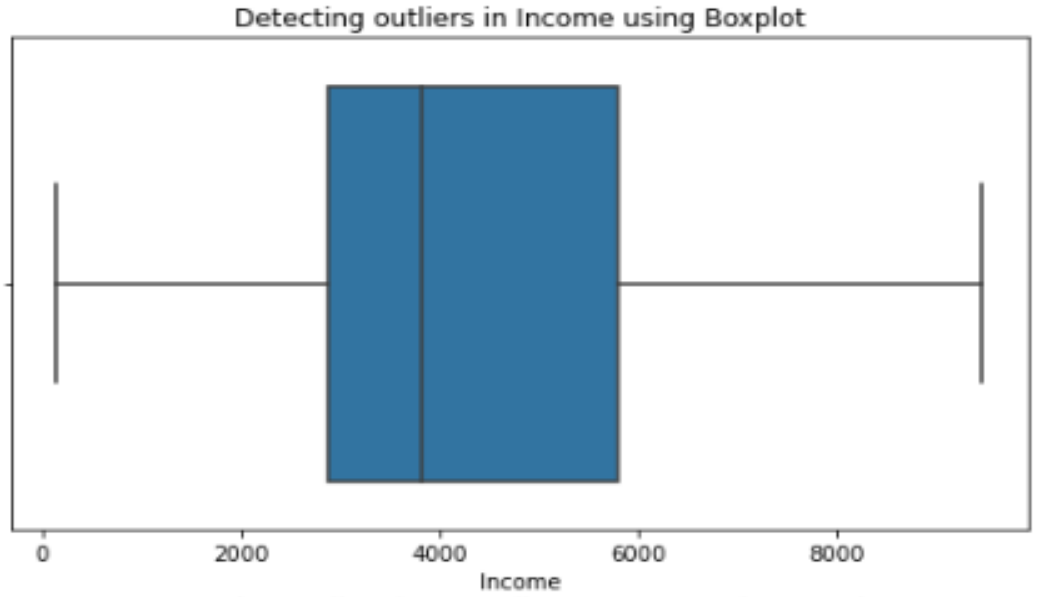
(1) Distribution of applicant income is right skewed and it has lots of outliers. This can be due to the high income differences in the society.

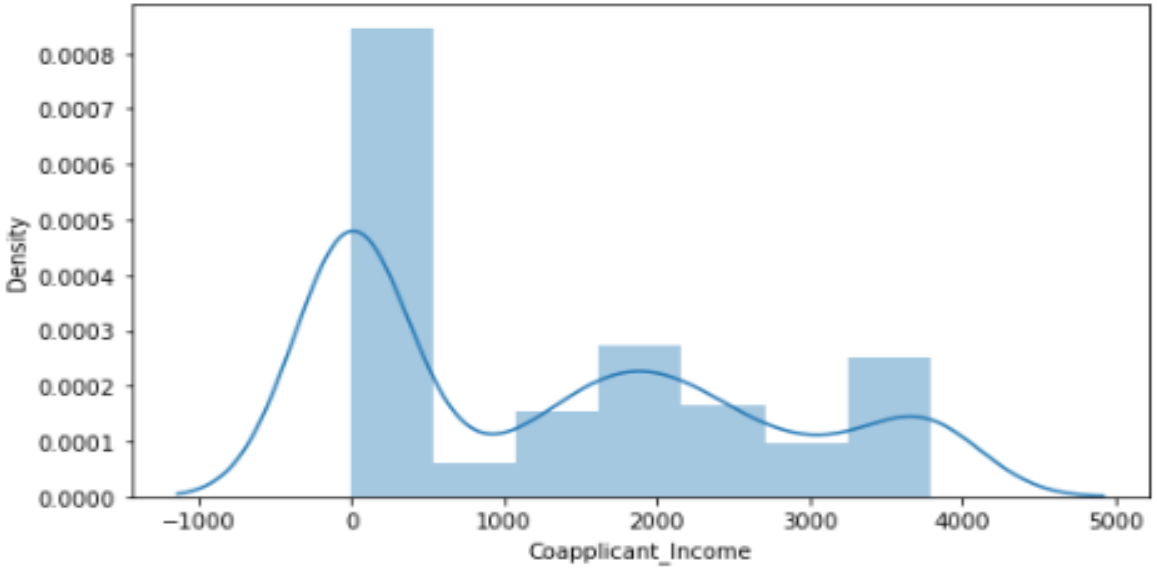
(2) Similar distribution with coapplicant income.

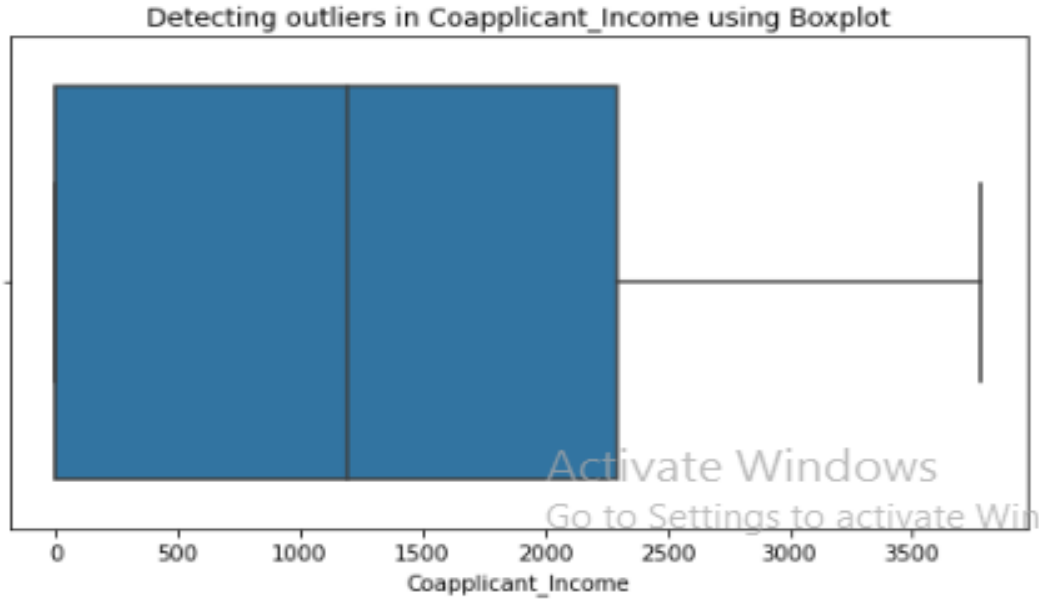
(3) loan amount distribution is fairly normal and still has lot of outliers but outliers in loan amount is possible because loan amount can vary depending upon requirement of applicant. as there can be different reasons why high amount loans were approved we will not treat outliers here.

# After outlier treatment



# Univariate analysis

1. Out of total applicants 82% are male and 18% are female i.e. Male applicants are more than Female

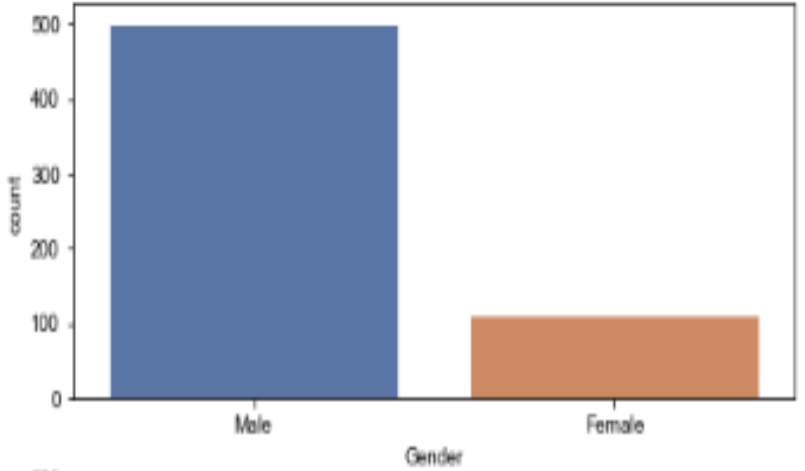
2. Out of total applicants 65% are married and 35% are unmarried i.e. Married applicant are more than Non-married

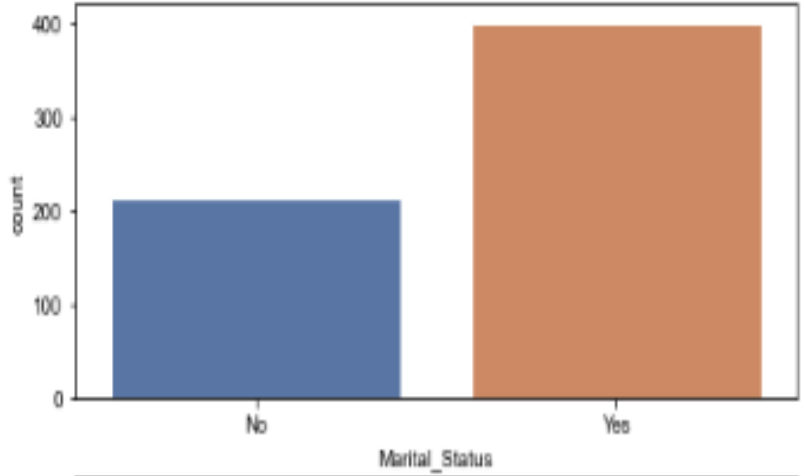
3. Out of total applicants 78% are Graduates and 22% are not graduate i.e.graduate applicant are more than not graduate

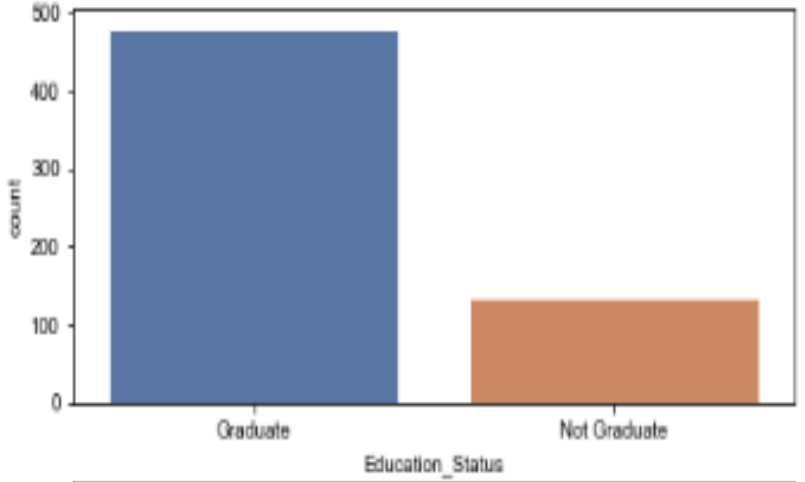
4. Out of total applicants 87% are Not self employed and 13% are self employed i.e.self-employed applicant are less than that of Non-Self-employed

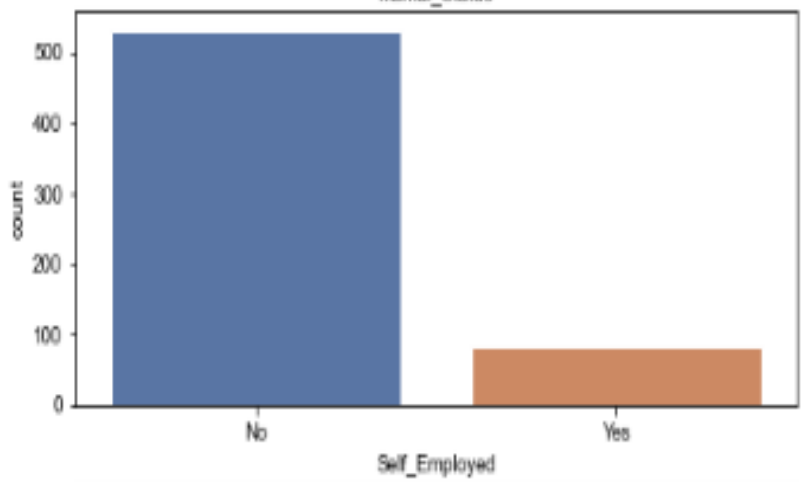
5. Out of total applicants 86% have credit history and 14% do not have credit history i.e.many applicants have Credit History

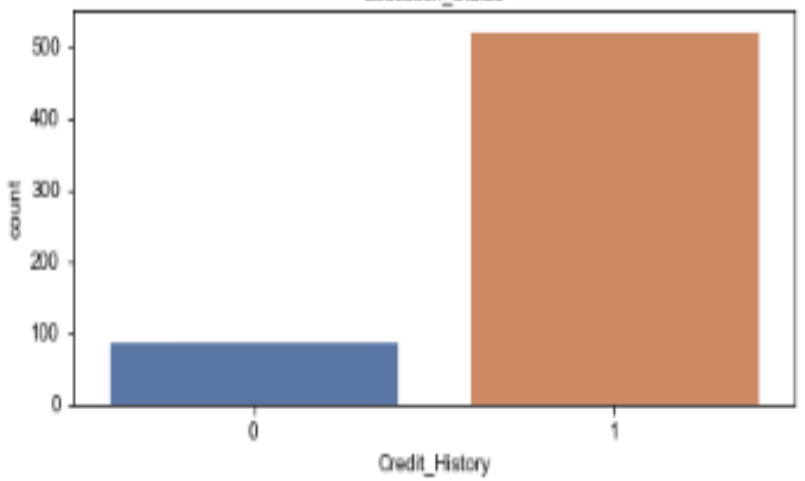
6. Out of total applicants 69% loans are approved and 31% loans are not approved i.e.more loans are approved than Rejected

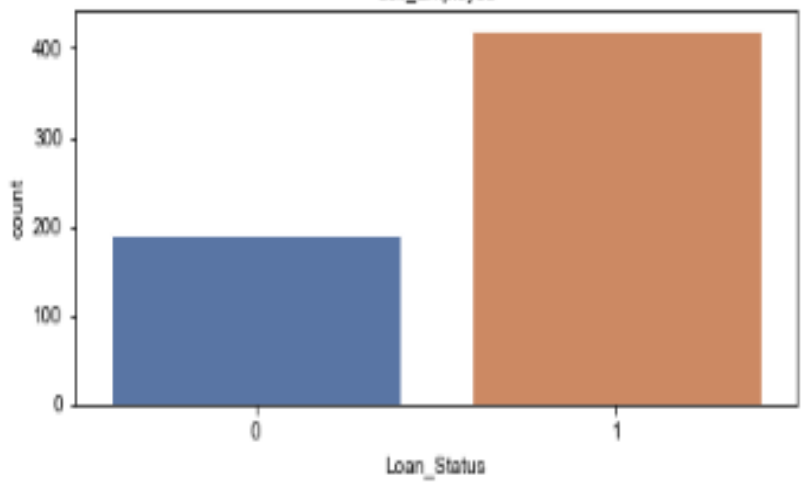


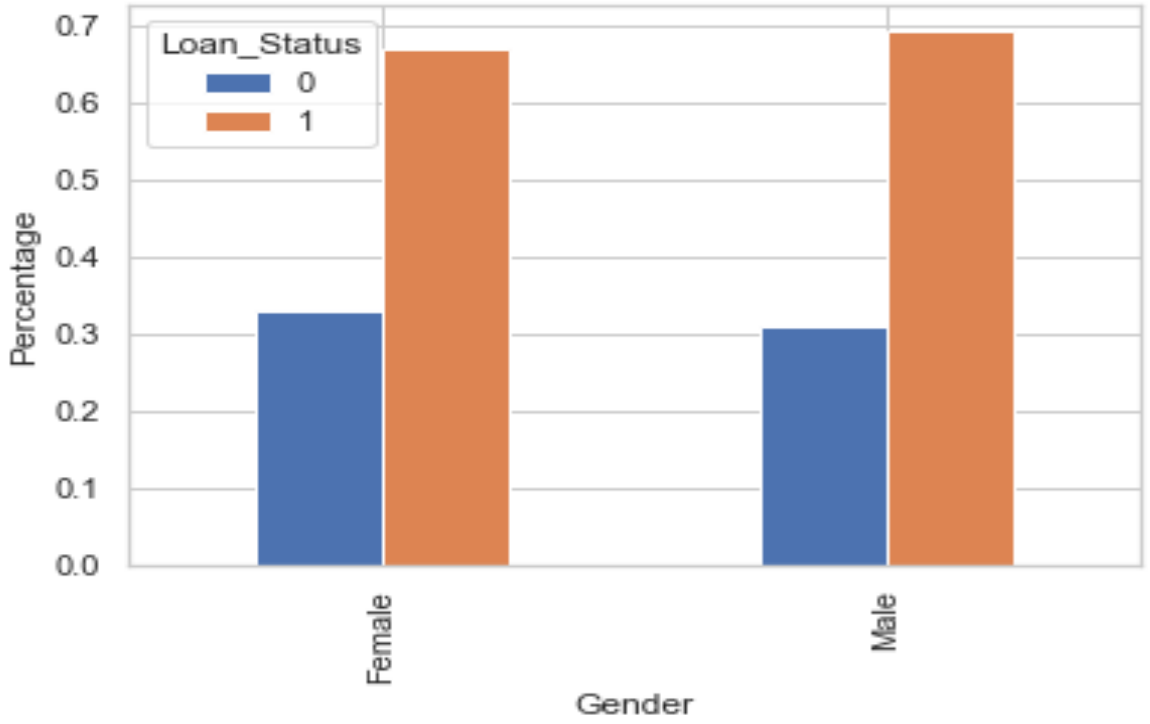



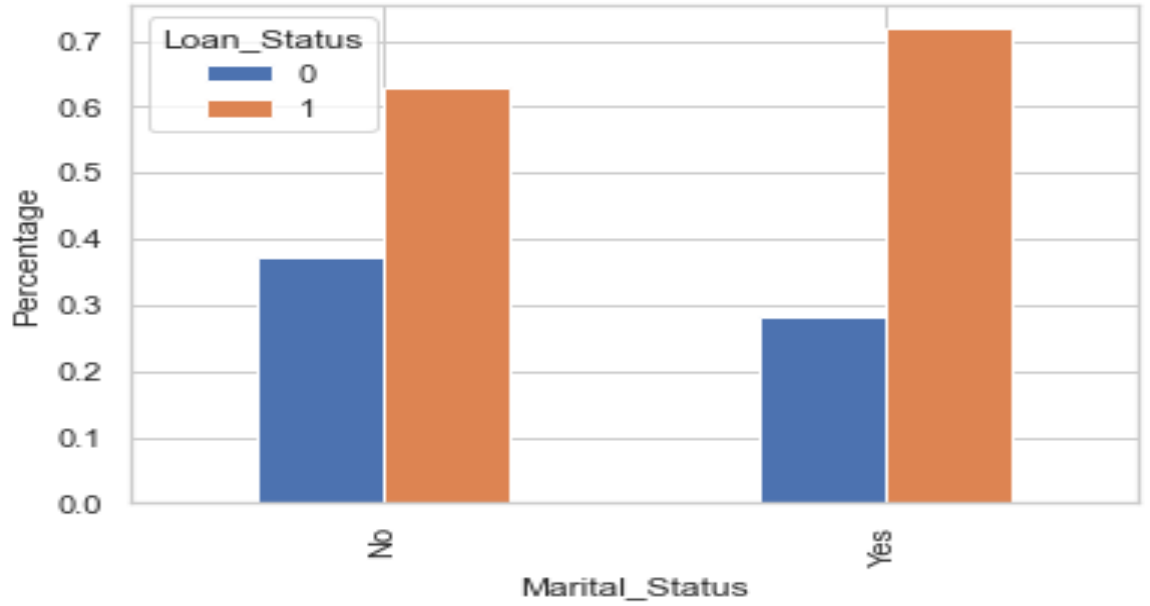



# Bivariate analysis categorical variable vs loan status

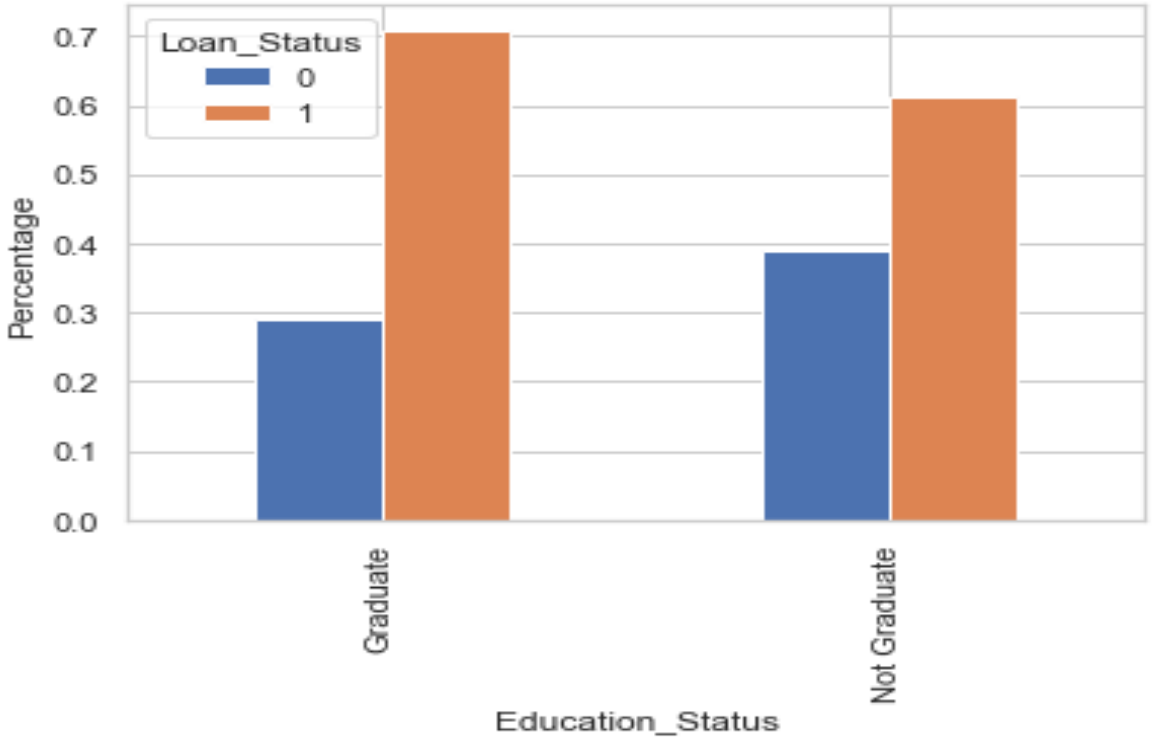
(1) Almost same percentage of male and female applicants have approved or disapproved loan. Gender is not significant deciding factor for loan status.



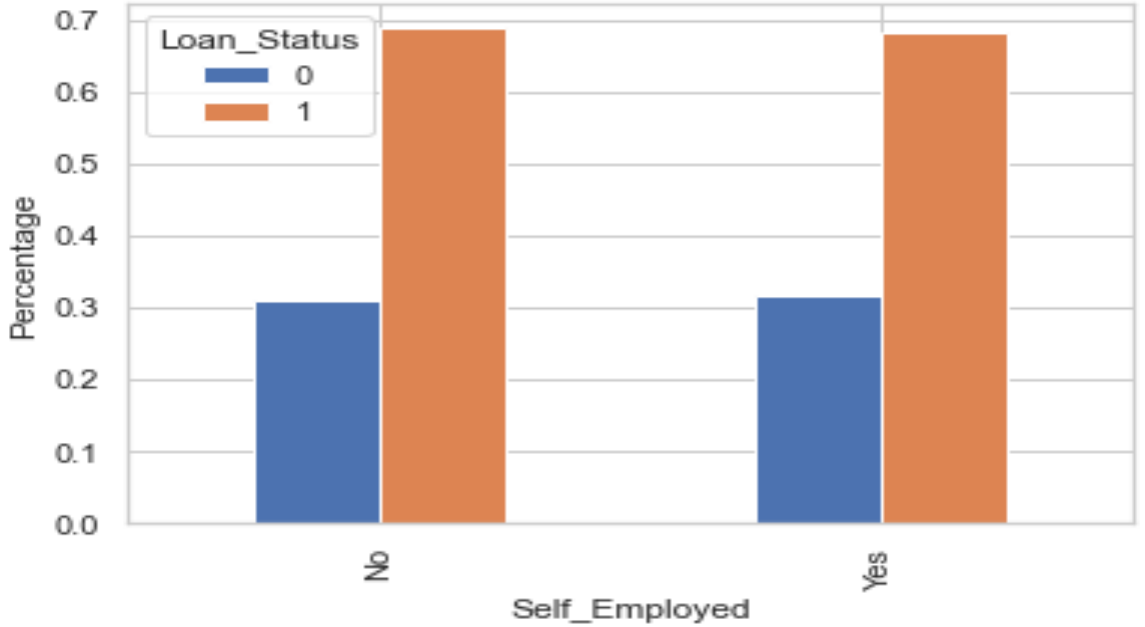

(2) Percentage of married applicant is more or less equal to unmarried applicants for approved loan.Married status is not significant deciding factor for loan status



(3) Percentage of graduate applicant is higher than non-graduate applicants for approved loan.

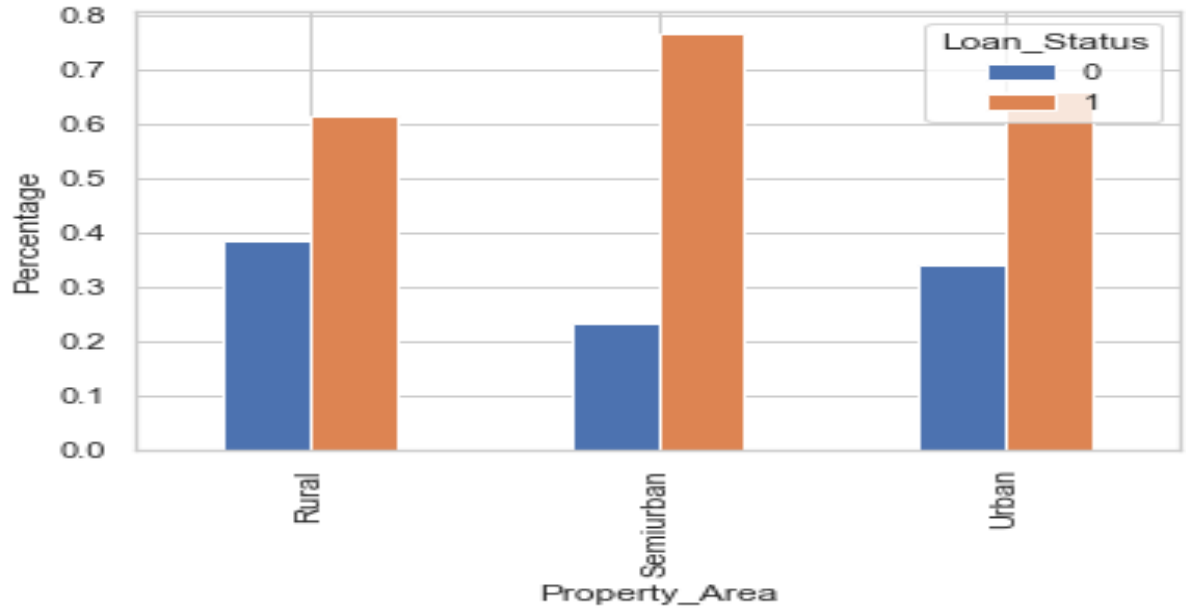



(4) Almost same percentage of self employed and non-self employed applicants have approved or disapproved loan. self employment is not significant deciding factor for loan status.



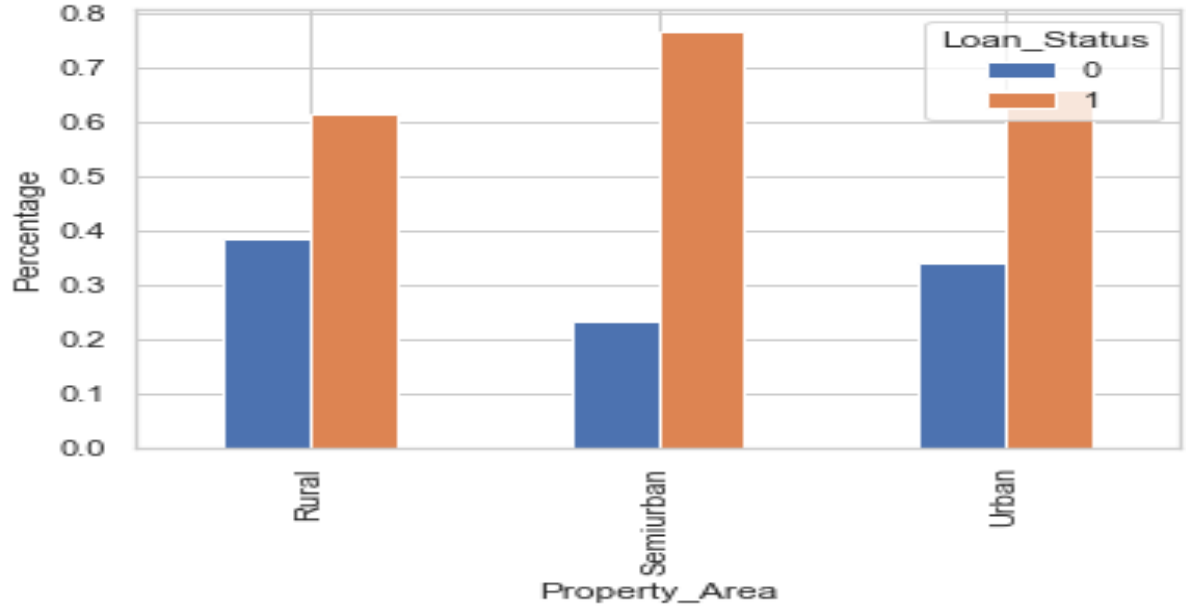
(5) People with a credit history as 1 are more likely to get their loans approved.


(6) The percentage of loans getting approved in the semi-urban area is higher as compared to that in rural or urban areas.



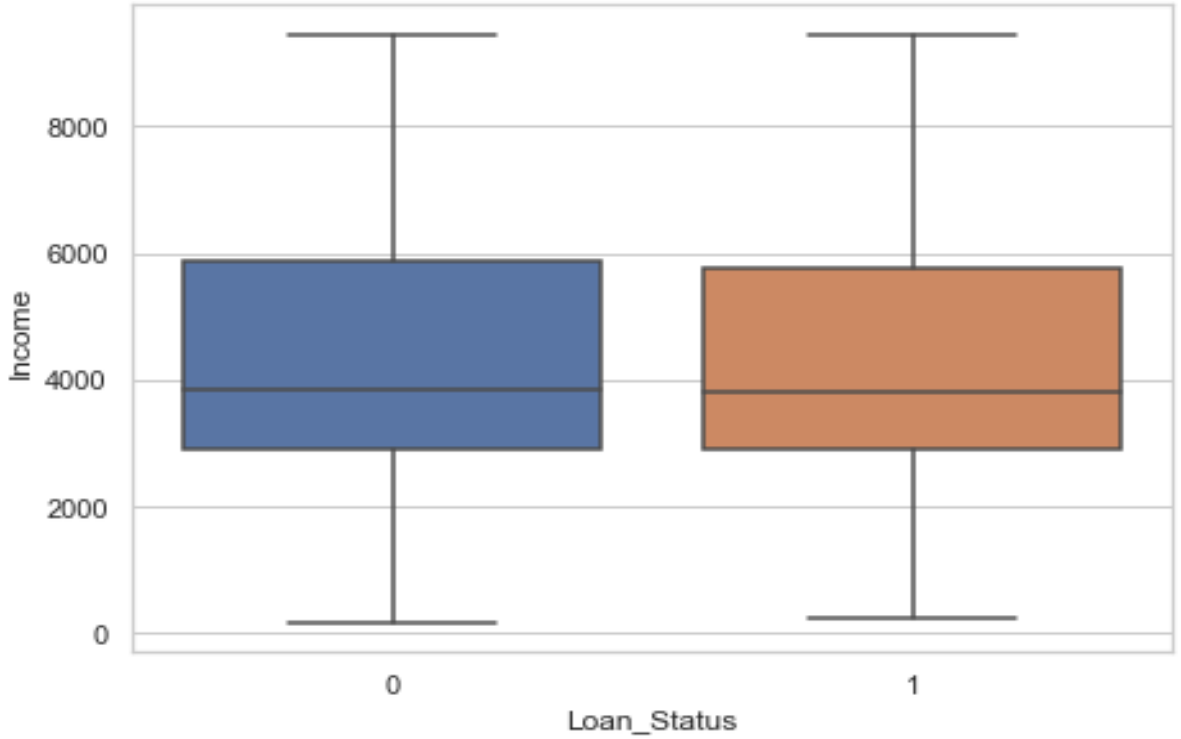
(5) People with a credit history as 1 are more likely to get their loans approved.


(6) The percentage of loans getting approved in the semi-urban area is higher as compared to that in rural or urban areas.

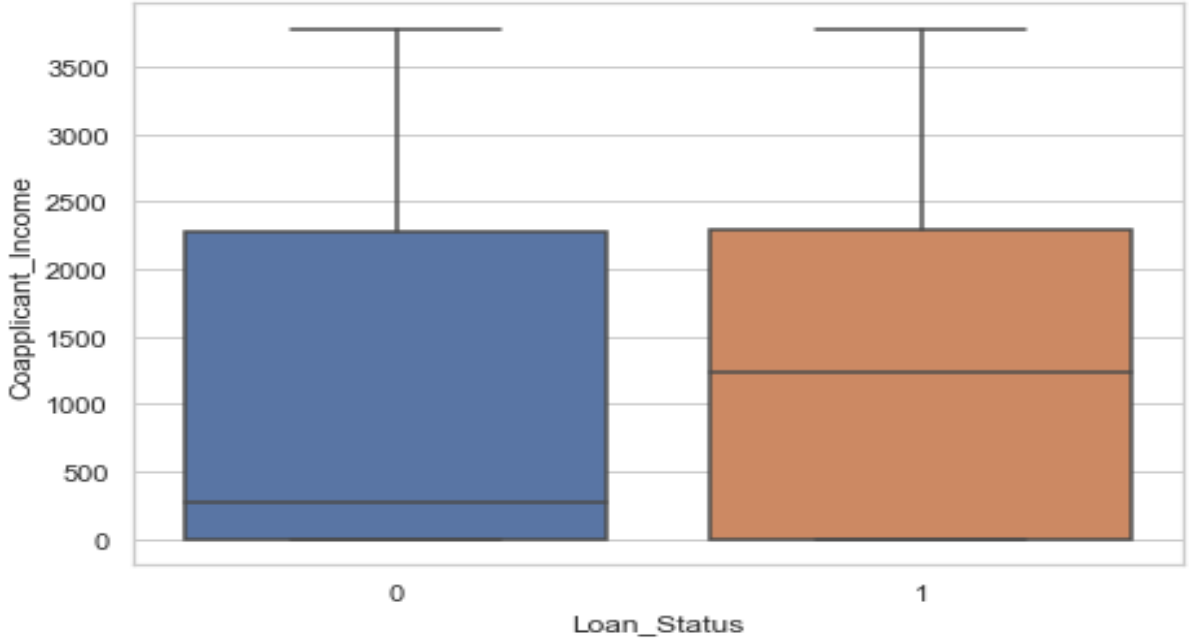


# Bivariate analysis loan status vs continuous variable

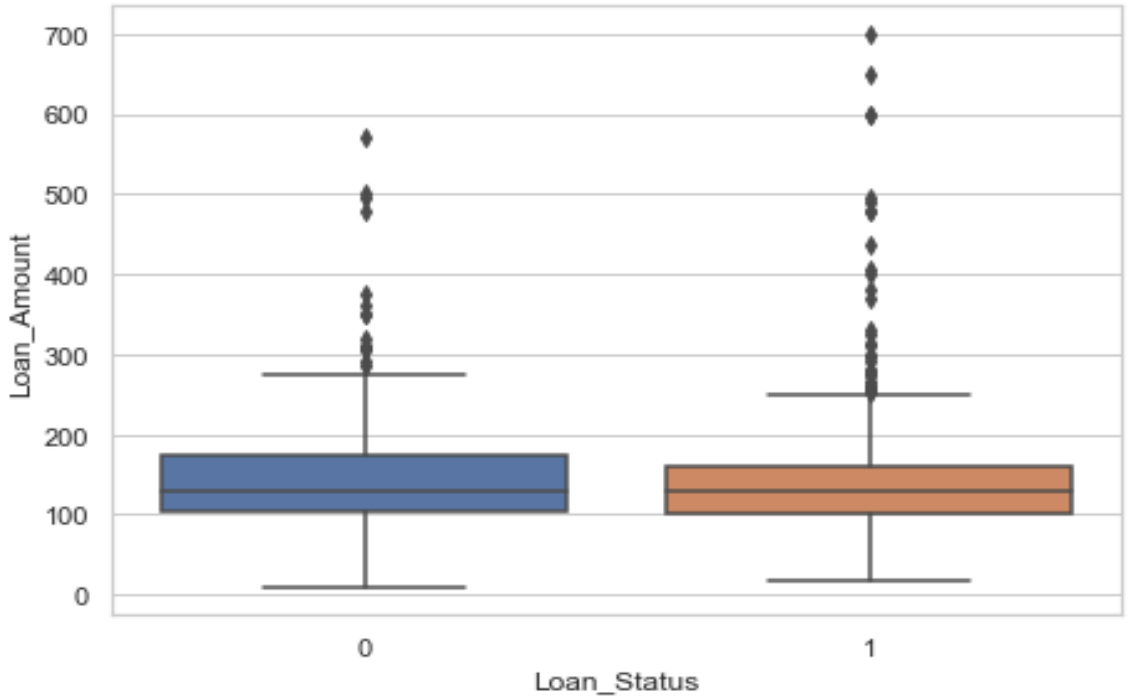
(1) Mean Applicant Income of approved loan(1) and non approved loan(0) are almost the same




(2) Mean of co-applicant Income of approved loan(1) is more than non approved loan(0).but The possible reason behind this may be that most of the applicants don't have any co-applicant so the co-applicant income for such applicants is 0 and hence the loan approval is not dependent on it. So, we can make a new variable in which we will combine the applicant's and co-applicants income to visualize the combined effect of income on loan approval. but also If co-applicant also have income then chances of getting loan approved increases

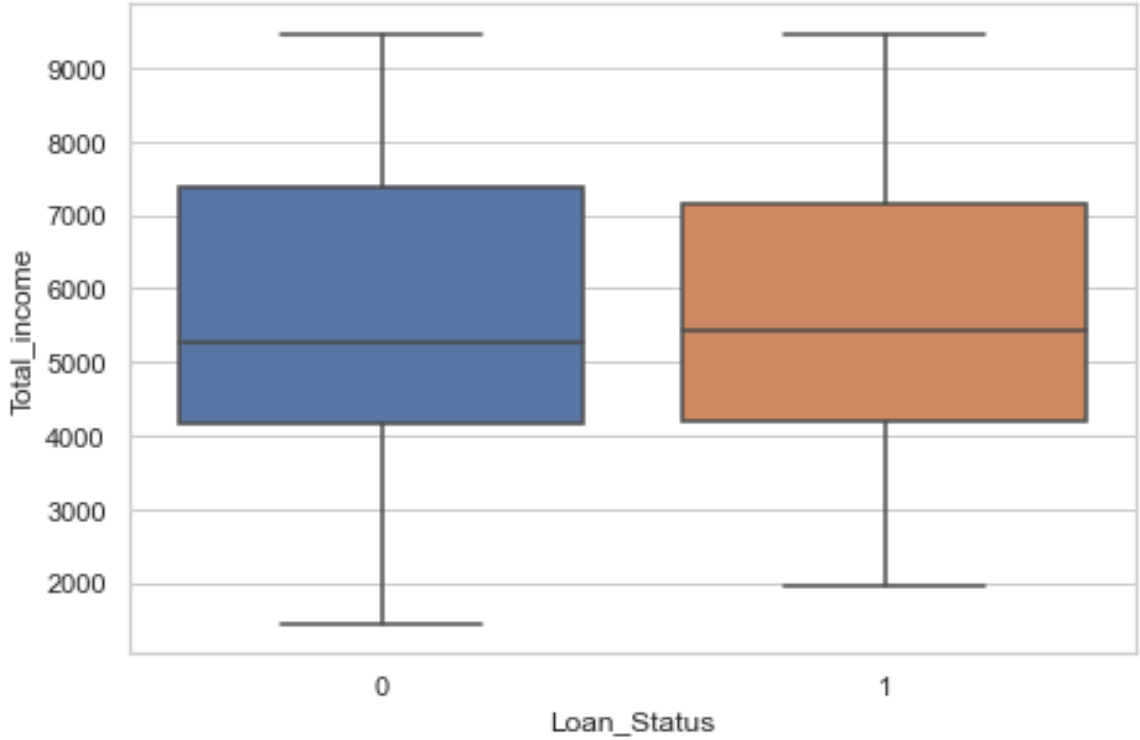


(3) Mean loan amount of disapproved loan is higher than approved loan. The odds of approved loans is higher for Low Loan Amount as compared to that of High Loan Amount.

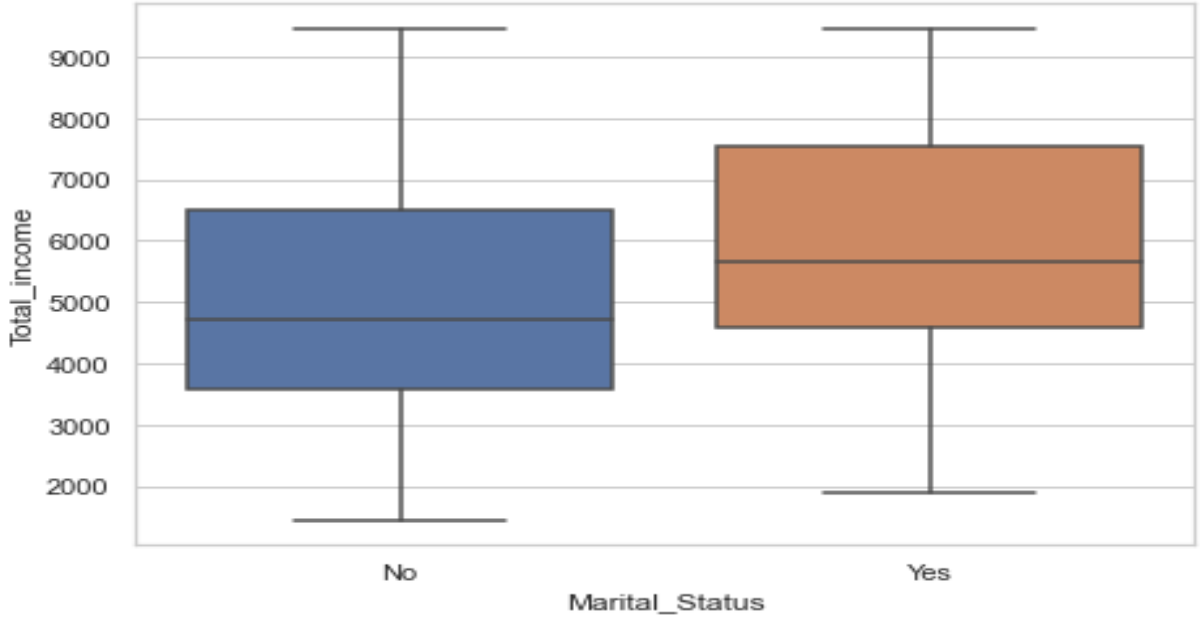


# Bivariate analysis categorical variable vs total income(new variable)

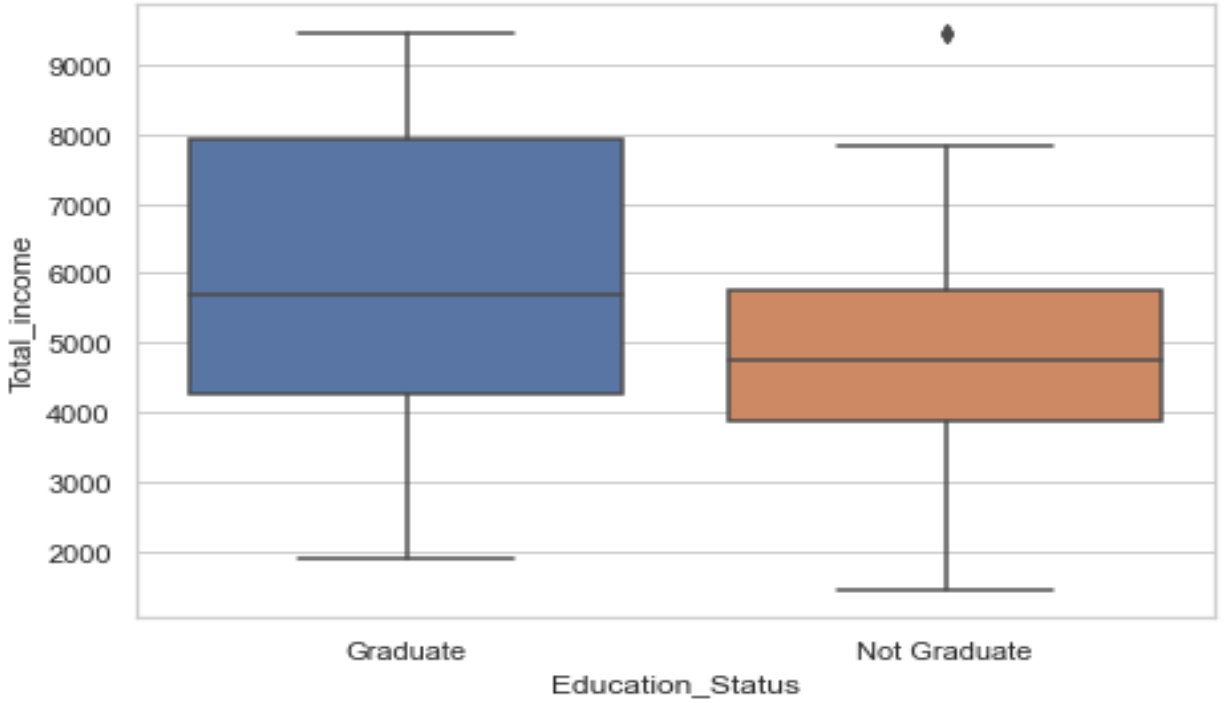
(1) Mean total Income of approved loan(1) and non approved loan(0) are almost the same



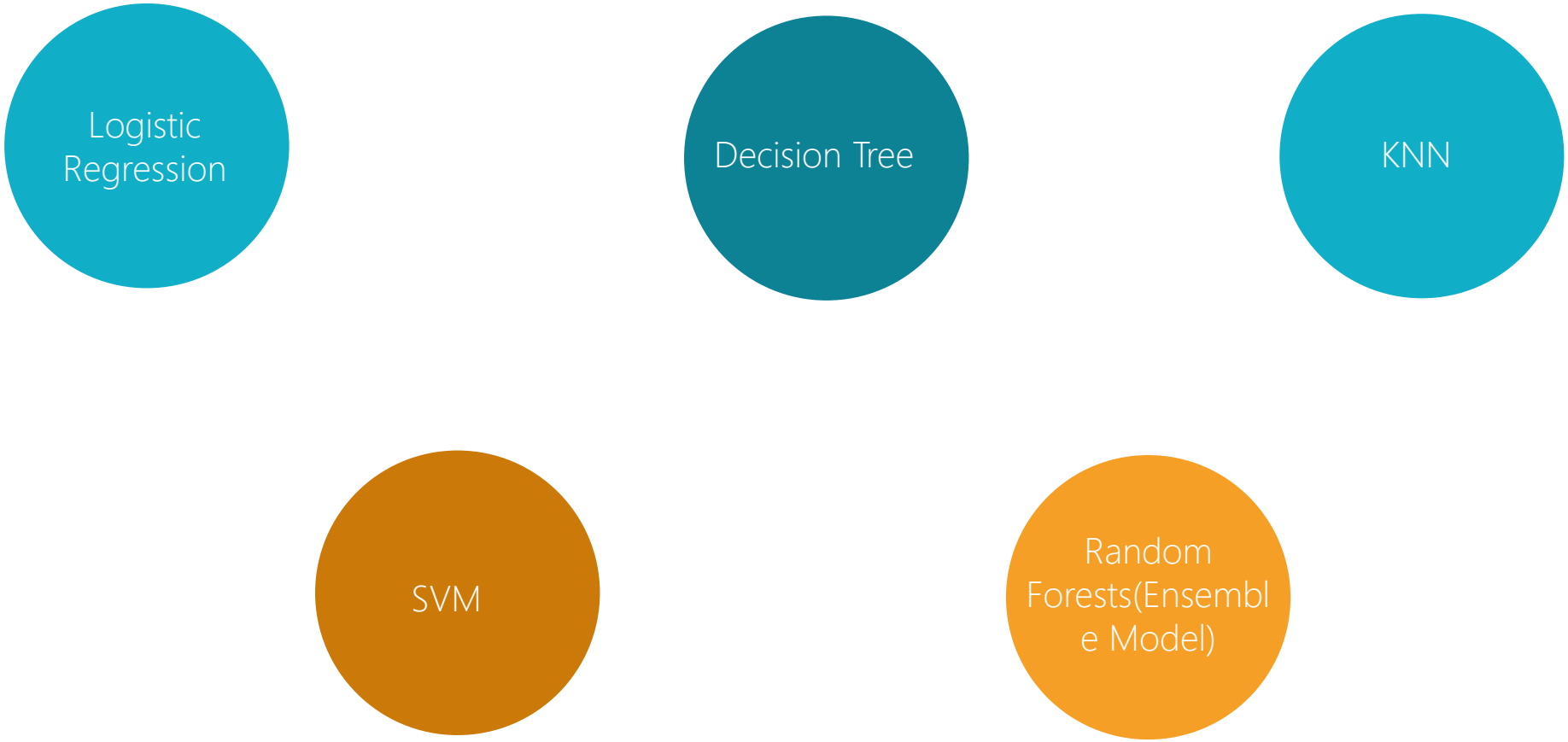

(2) Mean total income of married applicant is higher than non-married applicant



(3) Mean total Income of graduates is higher than non graduates.



# Classification Algorithms To Be Used



# Logistic Regression

78.37%

Accuracy score

| Actual \ Predicted | +ve | -ve |
|---|---|---|
| +ve | 26 | 38 |
| -ve | 2 | 119 |

Confusion matrix

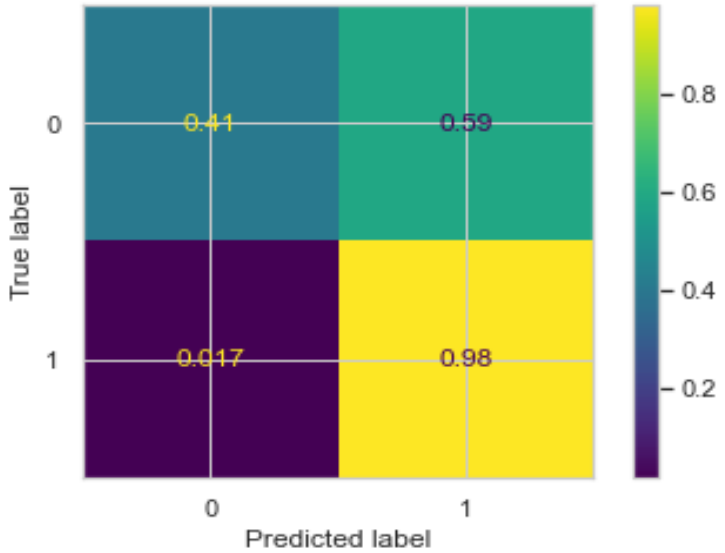
81.10%

Stratified k-Fold cross validation

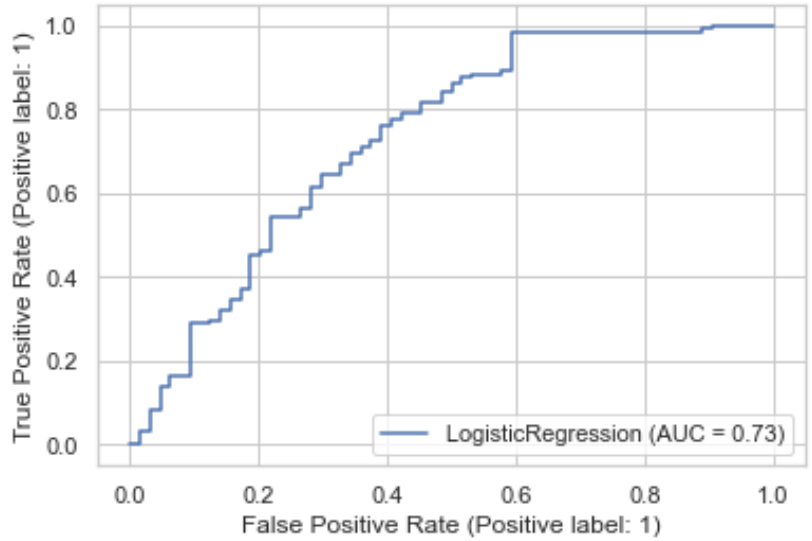
80.45%

k-Fold cross validation

## Confusion matrix




## ROC Curve




## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.41 | 0.57 | 64 |
| 1 | 0.76 | 0.98 | 0.86 | 121 |
| accuracy | | | 0.78 | 185 |
| macro avg | 0.84 | 0.69 | 0.71 | 185 |
| weighted avg | 0.82 | 0.78 | 0.76 | 185 |

# Decision tree

67.56%

Accuracy score

| Actual \ Predicted | +ve | -ve |
|---|---|---|
| +ve | 34 | 30 |
| -ve | 30 | 91 |

Confusion matrix

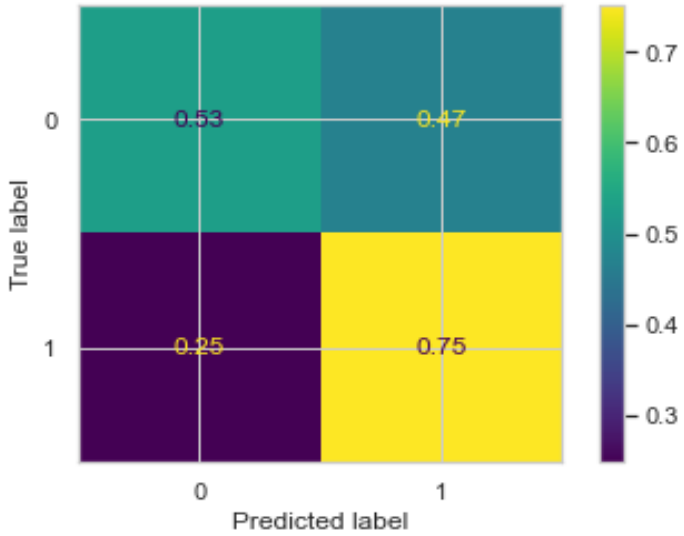
70.86%

Stratified k-Fold cross validation

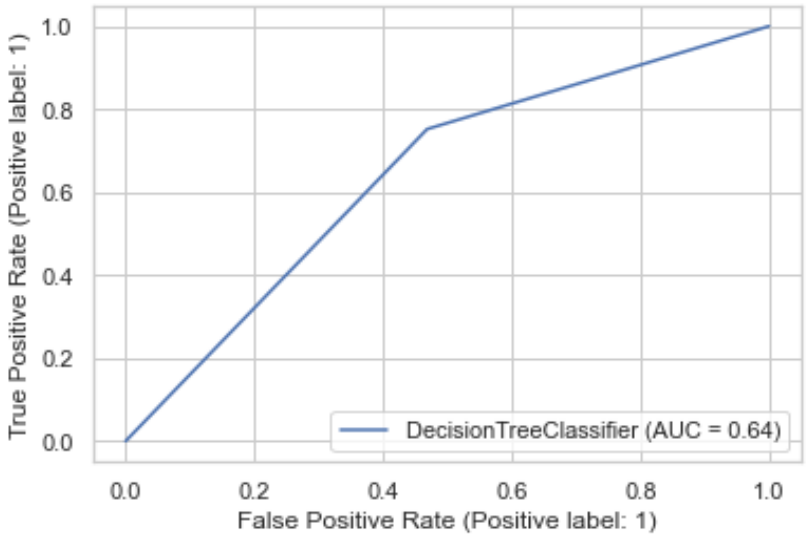
73.78%

k-Fold cross validation

## Confusion matrix




## ROC Curve




## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.53 | 0.53 | 0.53 | 64 |
| 1 | 0.75 | 0.75 | 0.75 | 121 |
| accuracy | | | 0.68 | 185 |
| macro avg | 0.64 | 0.64 | 0.64 | 185 |
| weighted avg | 0.68 | 0.68 | 0.68 | 185 |

# Random Forest Classification

77.29%

Accuracy score

| Actual \ Predicted | +ve | -ve |
|---|---|---|
| +ve | 31 | 33 |
| -ve | 9 | 112 |

Confusion matrix

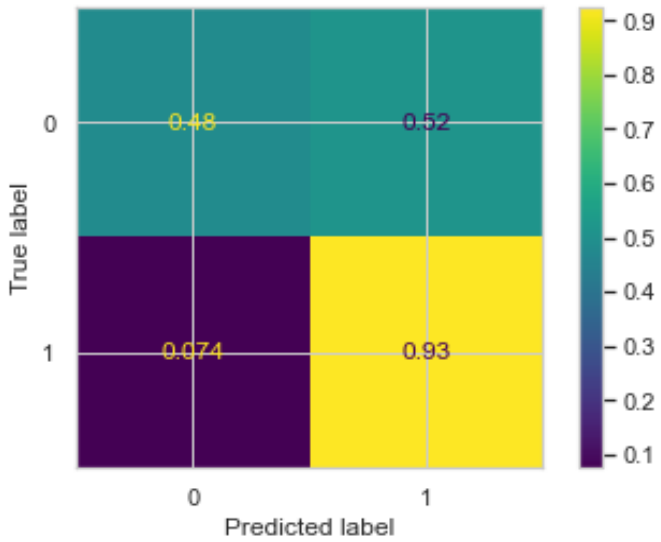
78.33%

Stratified k-Fold cross validation

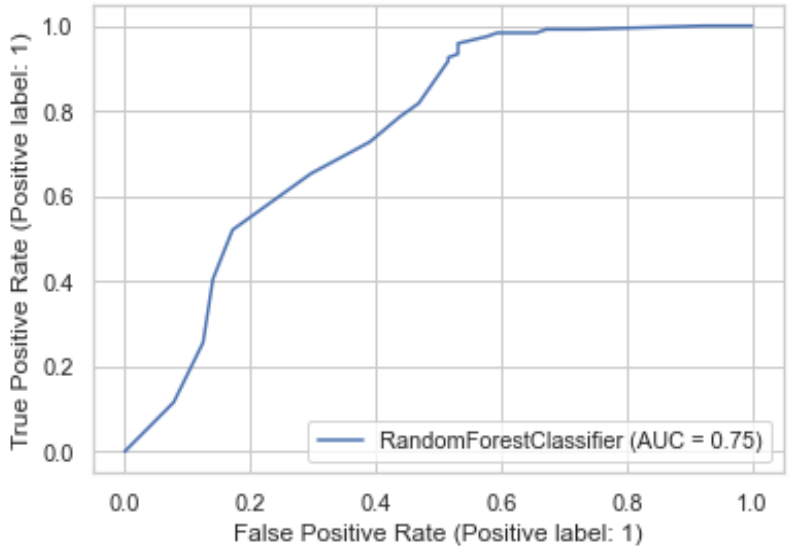
77.85%

k-Fold cross validation

## Confusion matrix




## ROC Curve




## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.48 | 0.60 | 64 |
| 1 | 0.77 | 0.93 | 0.84 | 121 |
| accuracy | | | 0.77 | 185 |
| macro avg | 0.77 | 0.70 | 0.72 | 185 |
| weighted avg | 0.77 | 0.77 | 0.76 | 185 |

# K-Neighbors Classifier

77.83%

Accuracy score

Predicted

| Actual | +ve | -ve |
|---|---|---|
| +ve | 30 | 34 |
| -ve | 7 | 114 |

Confusion matrix

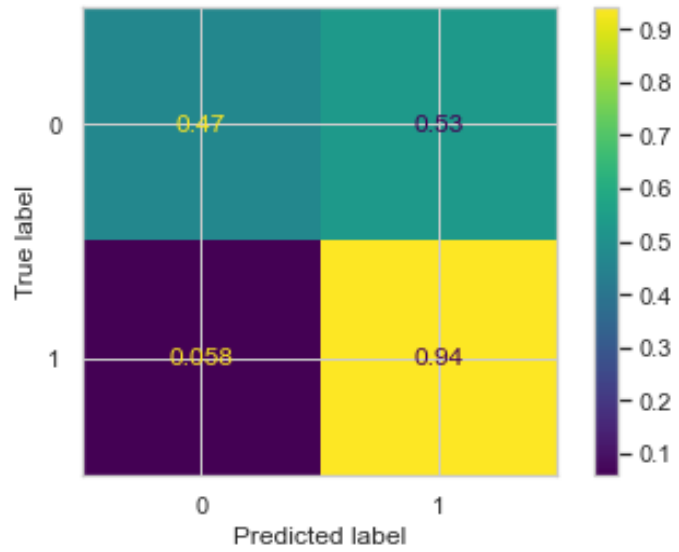
77.19%

Stratified k-Fold cross validation

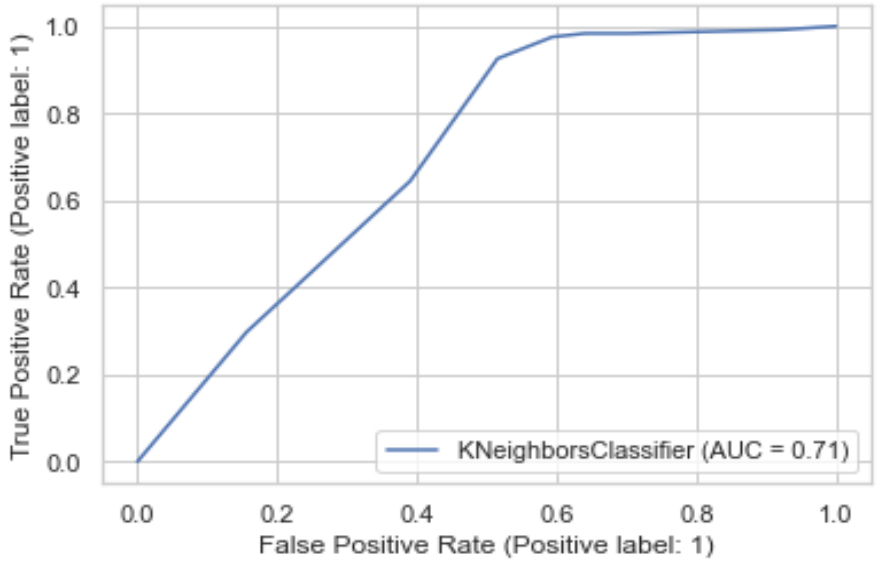
76.71%

k-Fold cross validation

## Confusion matrix




## ROC Curve




## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.81 | 0.47 | 0.59 | 64 |
| 1 | 0.77 | 0.94 | 0.85 | 121 |
| accuracy | | | 0.78 | 185 |
| macro avg | 0.79 | 0.71 | 0.72 | 185 |
| weighted avg | 0.78 | 0.78 | 0.76 | 185 |

# SVM Classifier

78.37%

Accuracy score

| Actual \ Predicted | +ve | -ve |
|---|---|---|
| +ve | 26 | 38 |
| -ve | 2 | 119 |

Confusion matrix

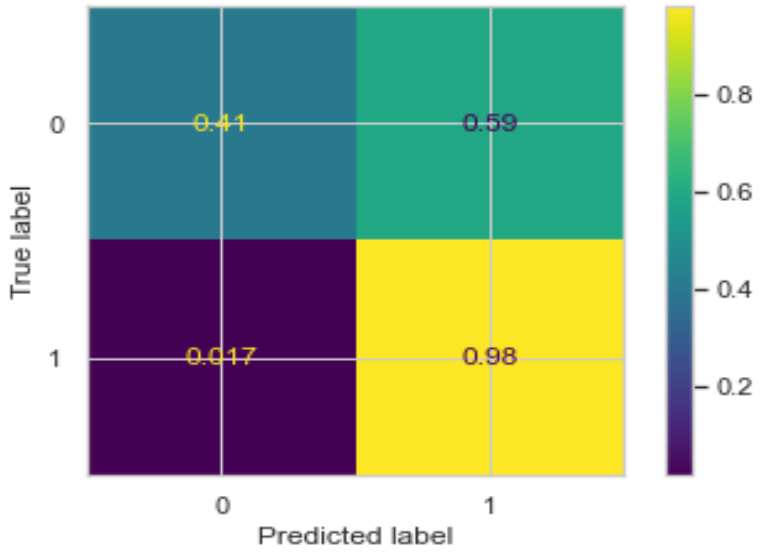
80.94%

Stratified k-Fold cross validation

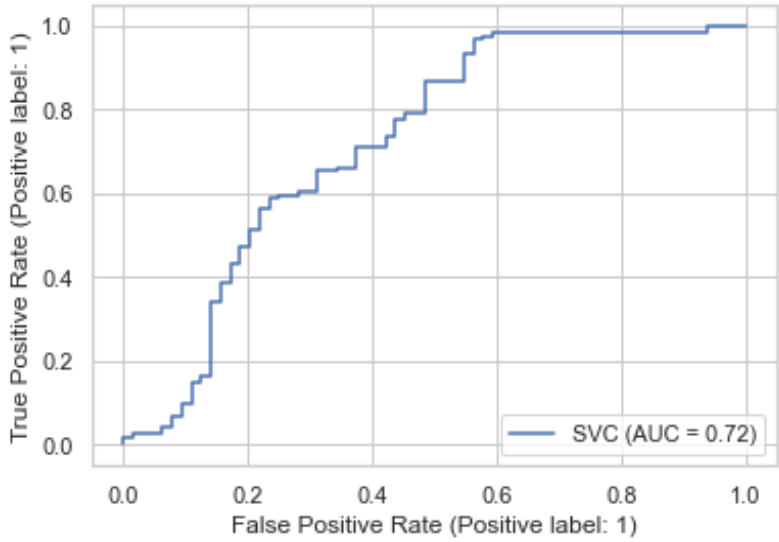
80.94%

k-Fold cross validation

## Confusion matrix




## ROC Curve




## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.41 | 0.57 | 64 |
| 1 | 0.76 | 0.98 | 0.86 | 121 |
| accuracy | | | 0.78 | 185 |
| macro avg | 0.84 | 0.69 | 0.71 | 185 |
| weighted avg | 0.82 | 0.78 | 0.76 | 185 |

# Comparing Models

| Sr. no. | Algorithm | Accuracy % | k-Fold cross validation % | Stratified K-Fold Cross Validation % | AUC |
|---|---|---|---|---|---|
| 1 | Logistic Regression | 78.37 | 80.45 | 81.10 | 0.73 |
| 2 | Decision Tree | 67.56 | 73.78 | 70.86 | 0.64 |
| 3 | Random Forest | 77.29 | 77.85 | 78.65 | 0.73 |
| 4 | KNN | 77.83 | 76.71 | 77.19 | 0.71 |
| 5 | SVM | 78.37 | 80.94 | 80.94 | 0.72 |

# Comparing Models

| No. | ML algorithm | Model name | Condition | Accuracy | K-fold accuracy | Stratified cv | AUC |
|---|---|---|---|---|---|---|---|
| 1 | logistic reg | log_model1 | | 77.29 | 79.48 | 79.96 | 0.73 |
| 2 | logistic reg | log_model2 | Scaled X | 78.37 | 80.45 | 81.10 | 0.73 |
| 3 | decision tree | tree_model1 | | 67.56 | 73.78 | 70.86 | 0.64 |
| 4 | Random forest | forest_model_1 | n=10 | 74.59 | | | |
| 5 | Random forest | forest_model_2 | n=7 | 77.83 | 73.89 | 77.04 | 0.75 |
| 6 | Random forest | forest_model_3 | n_estimators=23,bootstrap=True,max_features=6, oob_score=True | 77.29 | 77.85 | 78.33 | 0.75 |
| 7 | knn | knn_model1 | n=1 | 71.35 | | | |
| 8 | knn | knn_model2 | n=9 | 77.83 | | | |
| 9 | knn | knn_model3 | n=6 | 77.83 | 76.71 | 77.19 | 0.71 |
| 10 | SVM | svm_model1 | kernel = 'linear' | 78.37 | 80.94 | 80.94 | 0.72 |
| 11 | SVM | svm_model2 | kernel = 'rbf' | 78.37 | 80.78 | 80.78 | 0.69 |

# Model selection

- Model selection refers to the process of choosing the model that best generalizes.
- Model selection criteria used in this project are Complexity of model, Accuracy, Hyper parameters. Even though accuracy of model based on Logistic Regression & SVM is highest. We have other criteria's to judge. Based on all this kNN is chosen as Predictive model for this project.

# Conclusion

- The main purpose of the project is to classify and analyze the nature of the loan applicants. We analyzed each variable to check if data is cleaned. Null values are treated. With Exploratory Data Analysis of Features of we saw how each feature is distributed. In this different graphs were visualized & many conclusions have been made.

- The predictive models based on Logistic Regression, Decision Tree, Random Forest, KNN and SVM give the accuracy as 78.37%, 67.56% ,77.29%,77.83 and 78.37 whereas the cross-validation is found to be 80.45%, 73.78%, 77.85%, 76.71% and 80.94% respectively. This shows that for the given dataset, the accuracy of model based on Logistic Regression & SVM is highest. kNN is chosen as Predictive model for this project based on Complexity of model, Accuracy, Hyper parameters

# Future Scope

- Primary objective of bank is to provide their wealth in the safe hand. This project automates the whole loan eligibility process of customers and is useful in reducing the time and manpower required to approve loans and select right candidate for loan.

- Also project work can be extended to higher level in future. Data available in this project might not be enough to give best generalized and highest accuracy model. With more data we can expect better Predictive model for loans that uses machine learning algorithms

# Thank You