Loan Default Prediction

TABLE OF CONTENTS

INTRODUCTION

Introduction

An assurance company is having a critical issue with loan approval process as the current process requires a lot of time and labor resources.

The company seeks to automate (in real time) the loan qualifying procedure based on information given by borrowers.

A machine learning model can help accelerate decision-making process with higher accuracy in approving a loan for a new borrower.

DATASET

Variable description in the original dataset

- Dependent Variable:

Variable Name	Description	Sample Data	Type
Loan_Status	Status of loan (Y = accepted, N = not accepted)	Y; N	Categorical

- Independent Variables:

Variable Name	Description	Sample Data	Type
Loan_ID	Loan reference number (unique ID)	LP001002; LP001003; ...	Categorical
Gender	Applicant gender (Male or Female)	Male; Female	Categorical
Married	Applicant marital status (Married or not married)	Married; Not Married	Categorical
Dependents	Number of family members	0; 1; 2; 3+	Categorical
Education	Applicant education/qualification (graduate or not graduate)	Graduate; Under Graduate	Categorical
Self_Employed	Applicant employment status (yes = self-employed, no = employed/others)	Yes; No	Categorical
ApplicantIncome	Applicant's monthly salary/income	5849; 4583; ...	Numerical
CoapplicantIncome	Additional applicant's monthly salary/income	1508; 2358; ...	Numerical
LoanAmount	Loan amount	128; 66; ...	Numerical
Loan_Amount_Term	The loan's repayment period (in days)	360; 120; ...	Categorical
Credit_History	Records of previous credit history (0: bad credit history, 1: good credit history)	0; 1	Categorical
Property_Area	The location of property (Rural/Semiurban/Urban)	Rural; Semiurban; Urban	Categorical

UNIVARIATE AND BIVARIATE ANALYSIS

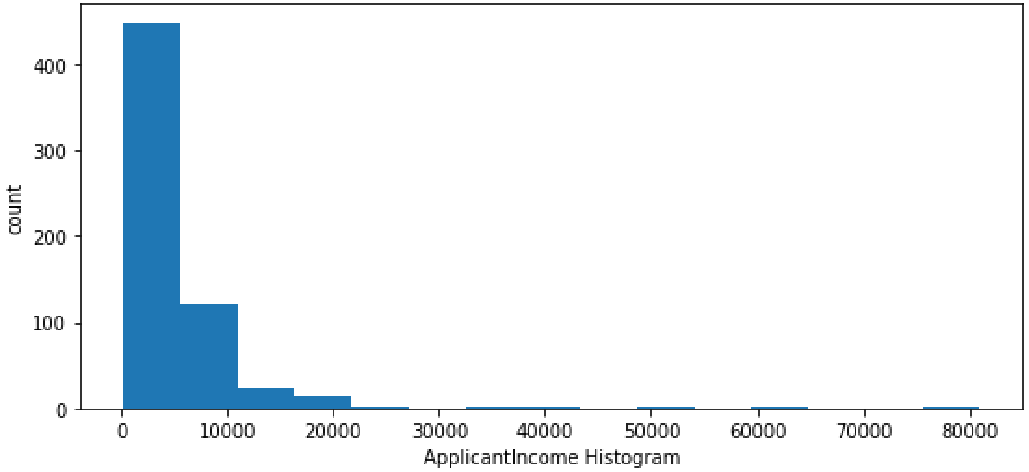
Univariate Analysis: Numerical Variables

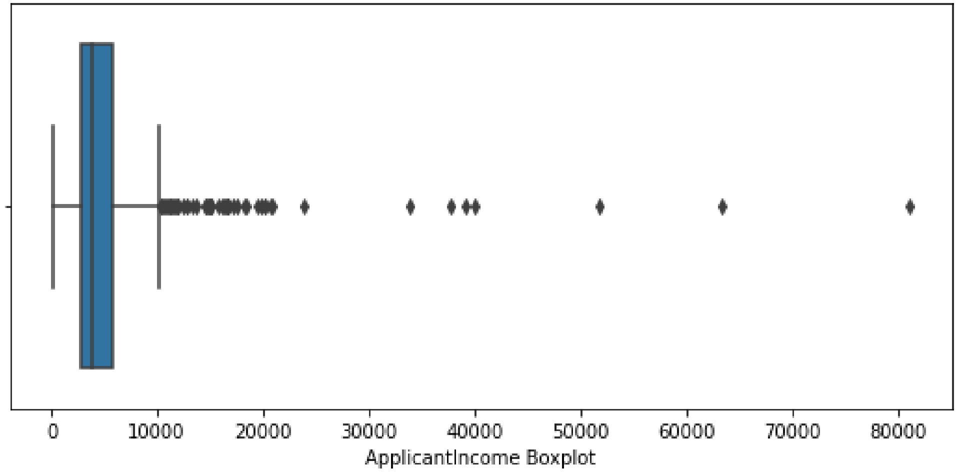
- Descriptive Statistics

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term
count	614.0	614.0	592.0	600.0
mean	5403.5	1621.2	146.4	342.0
std	6109.0	2926.2	85.6	65.1
min	150.0	0.0	9.0	12.0
25%	2877.5	0.0	100.0	360.0
50%	3812.5	1188.5	128.0	360.0
75%	5795.0	2297.2	168.0	360.0
max	81000.0	41667.0	700.0	480.0

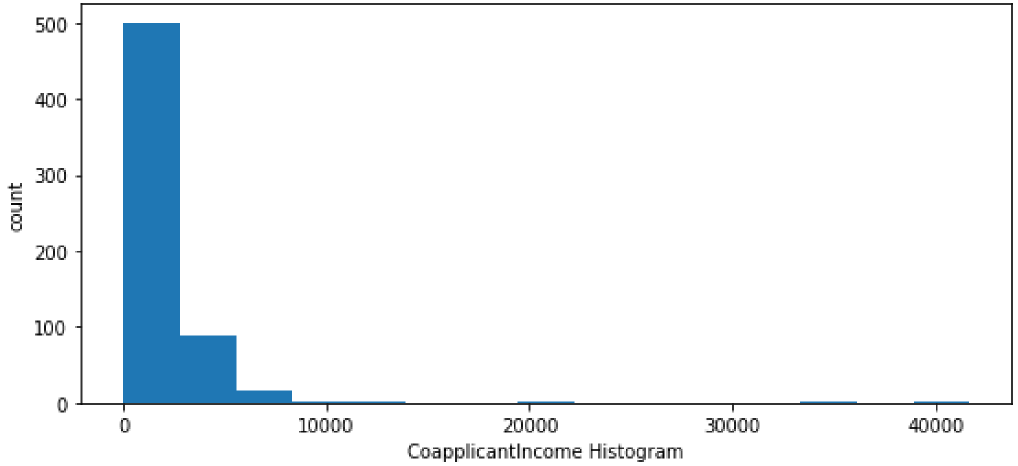
Univariate Analysis: Numerical Variables

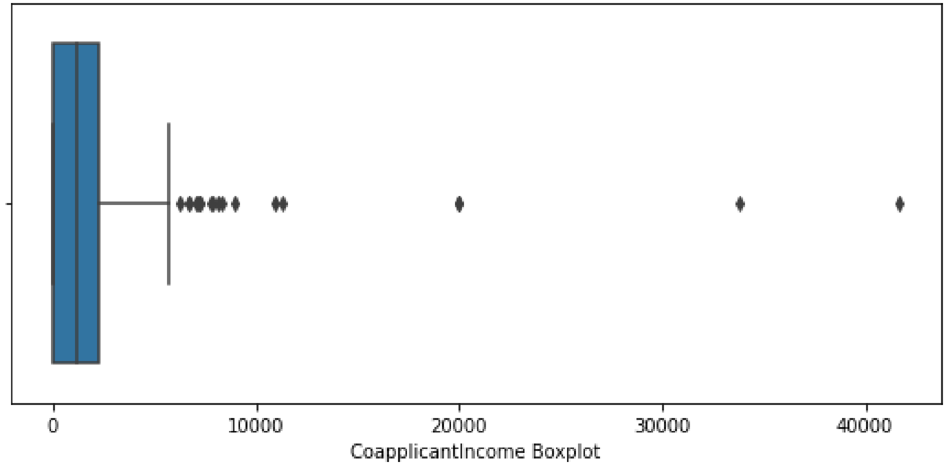
- Distribution of Applicant Income:





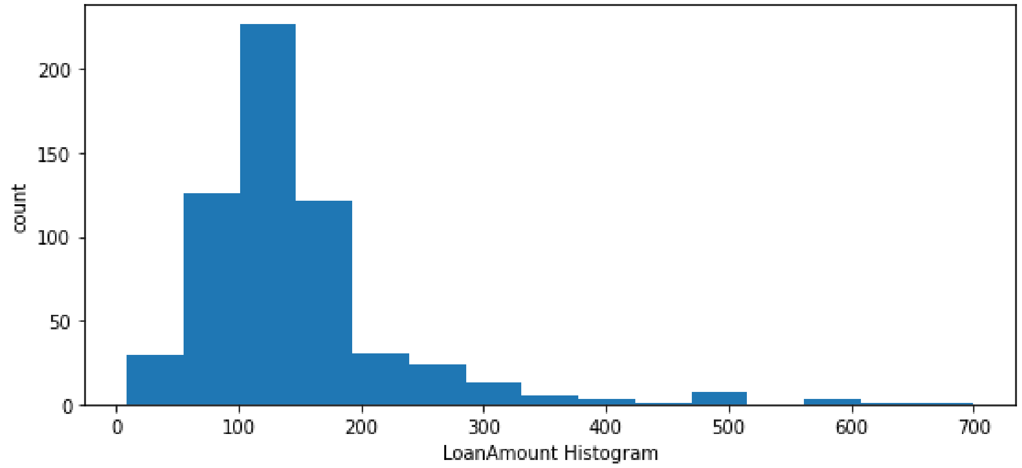
- Distribution of Co-applicant Income:

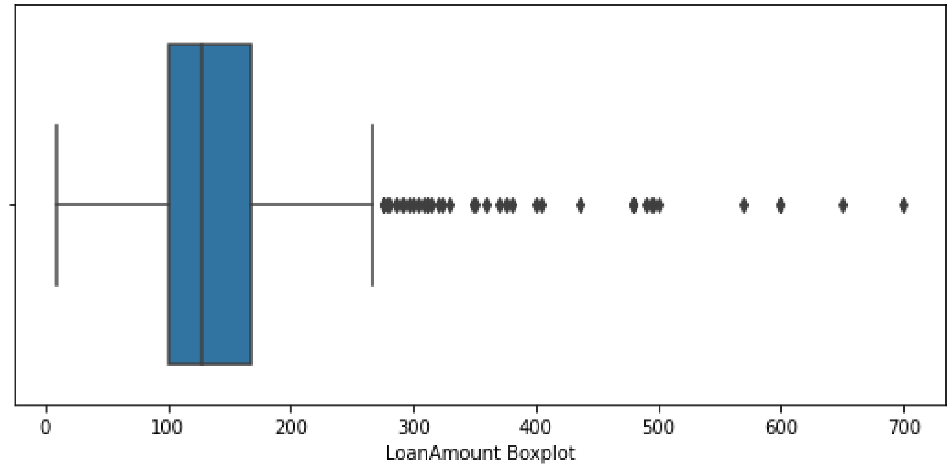




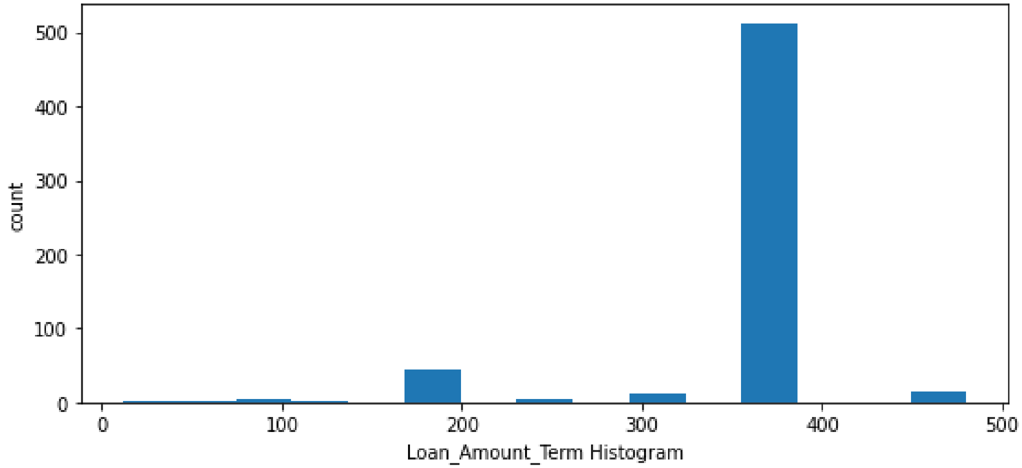
Univariate Analysis: Numerical Variables

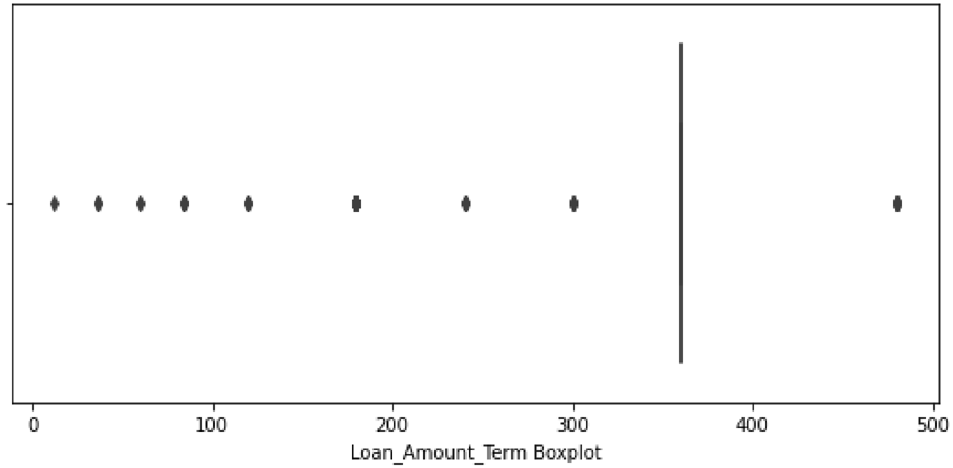
- Distribution of Loan Amount:





- Distribution of Loan Amount Term

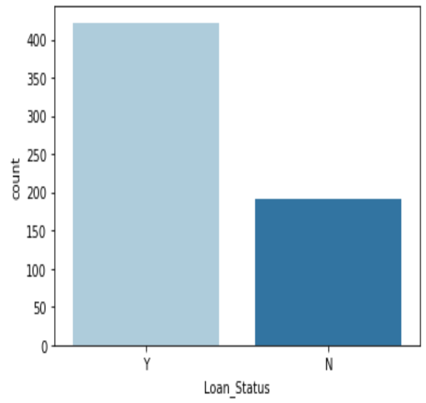




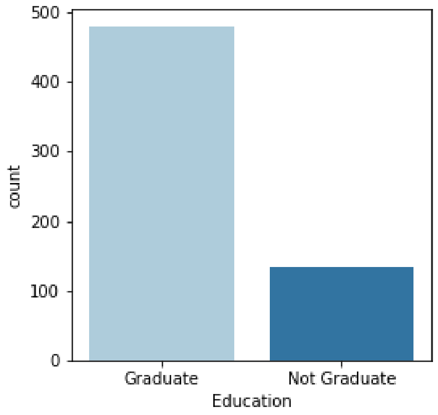
Dataset & Descriptive Analysis

Univariate Analysis: Categorical Variables

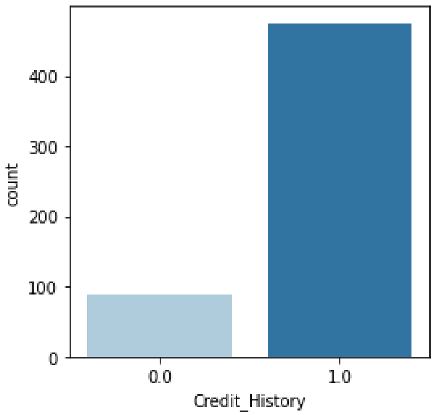
Loan Status (*)



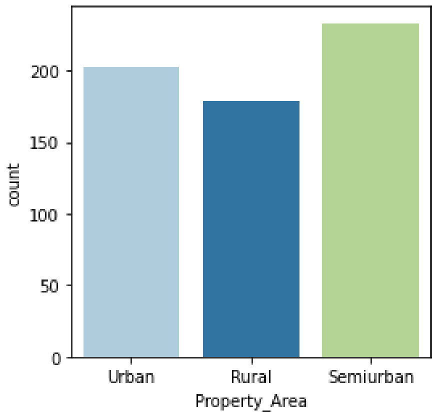
Education



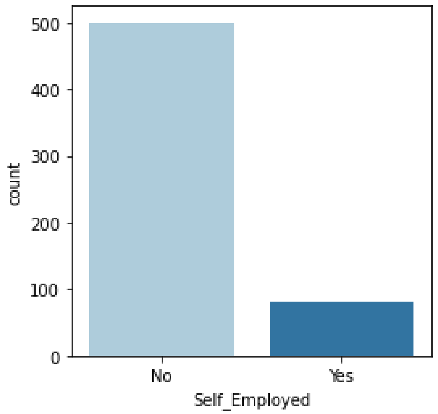
Credit History



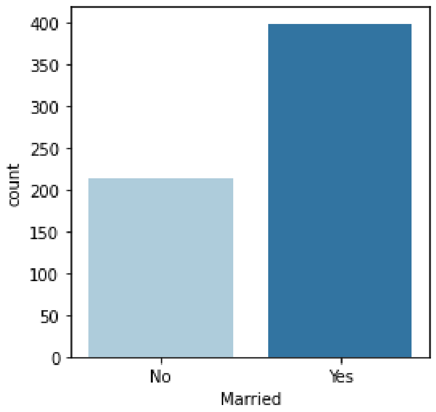
Property Area



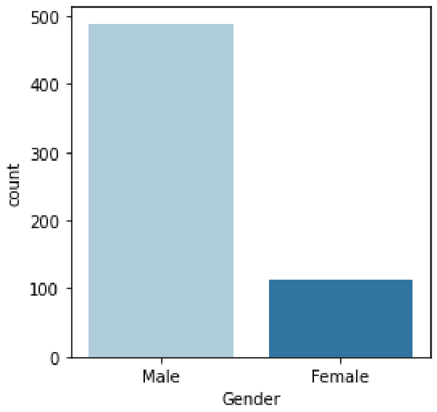
Self Employed



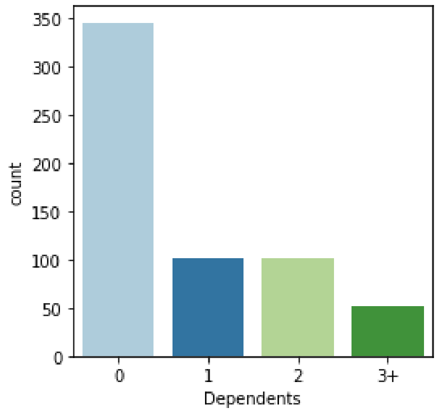
Married



Gender



Dependents



(*): Dependent Variable

Univariate Analysis: Insights

#1

The distribution of numerical variables seems to be right-skewed but should not be a problem since we have over 600 observations.

#2

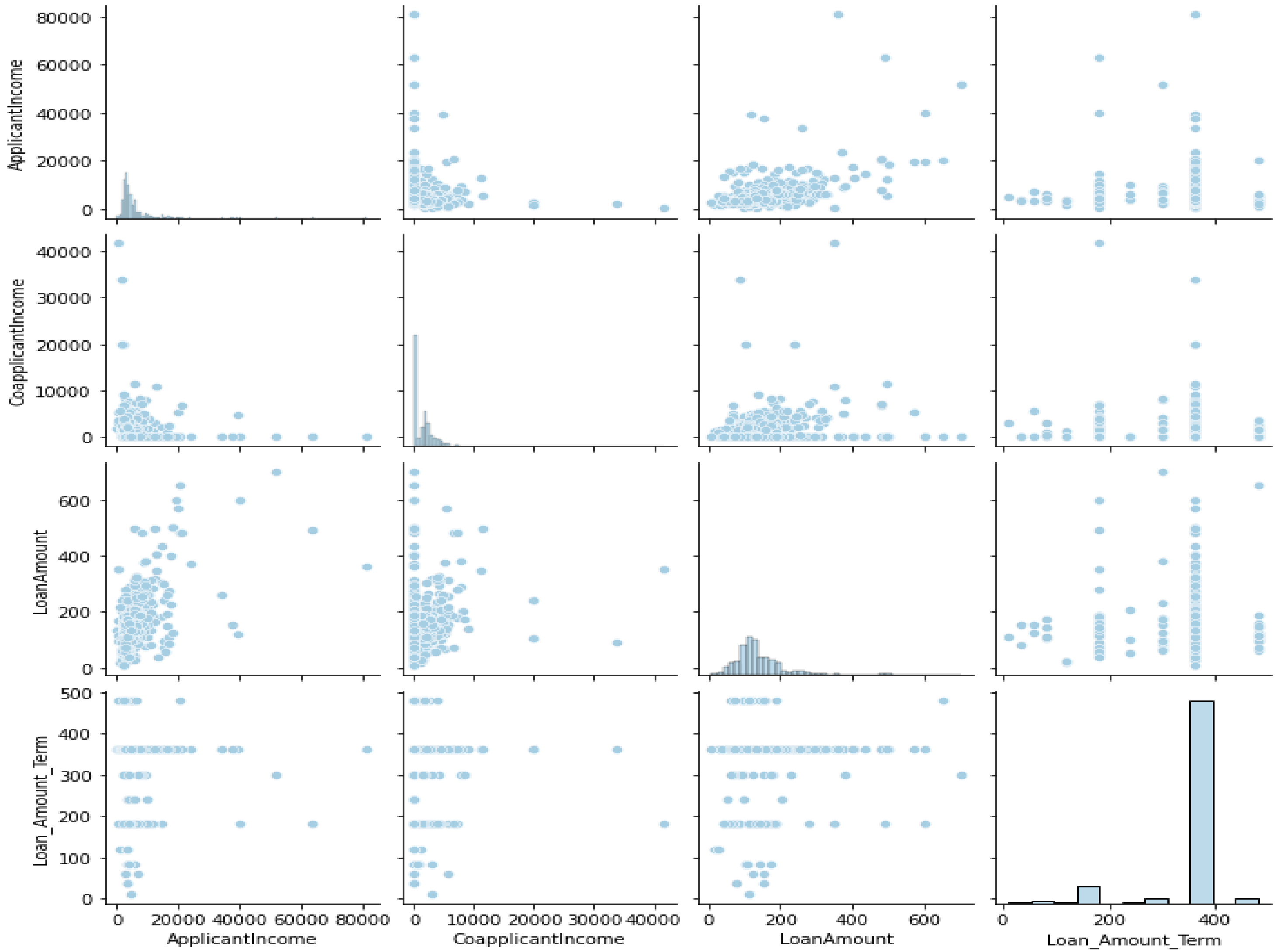
There is a different in variation between numerical variables, for example, maximum loan amount is ~ $700, and maximum applicant income is ~ $81,000. Therefore, standardization might be applicable.

#3

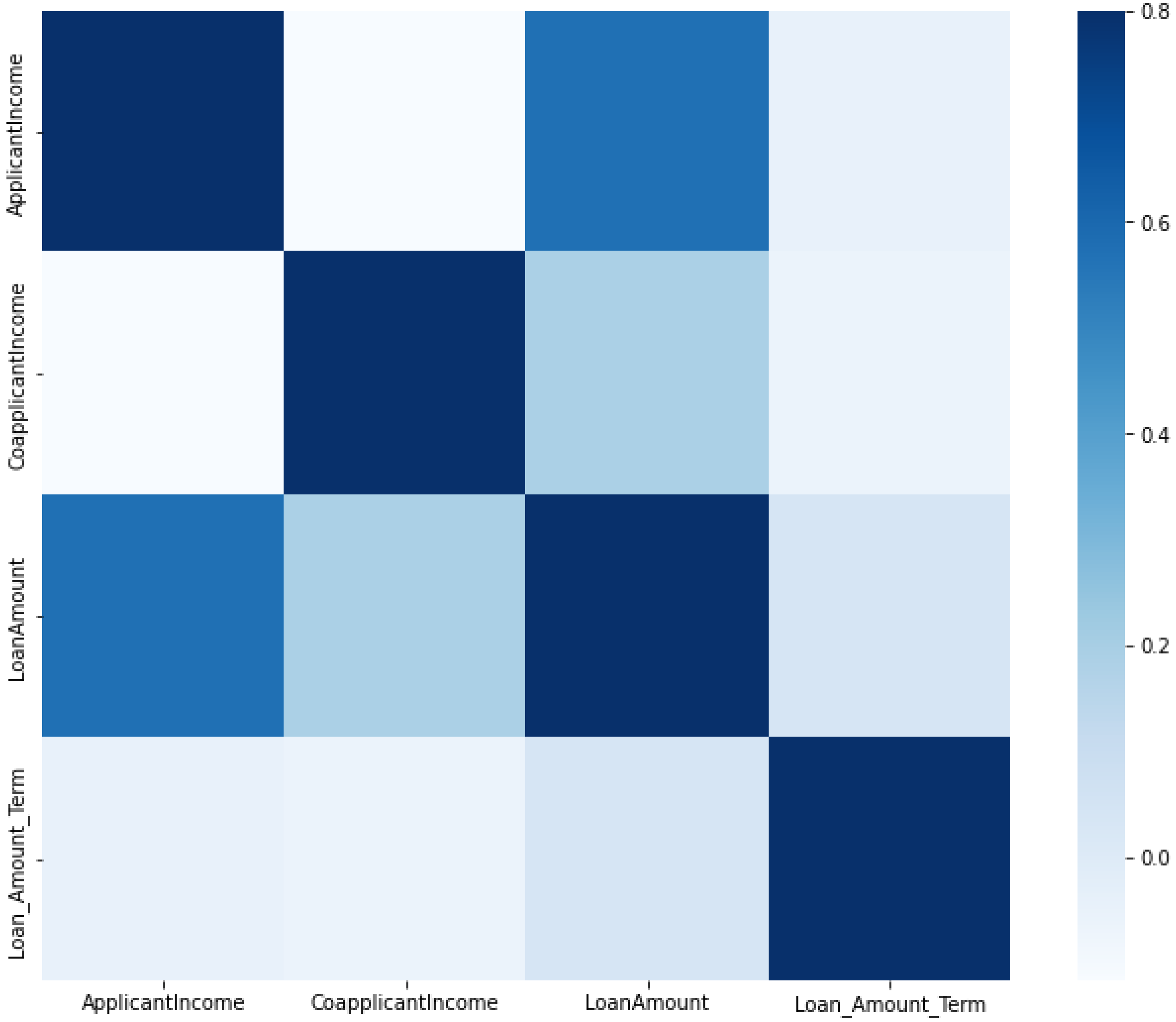
There is a large discrepancy in outcomes for the dependent variable in the data set, which might indicate oversampling.

Dataset & Descriptive Analysis

Bivariate Analysis: Scatterplots



Bivariate Analysis: Correlation Matrix Heatmap



DATA PROCESSING

Dropping Variables:

- Drop “Loan_ID” as it is irrelevant to our analysis.

Imputing Data:

- There are a few missing values in our data set (figure 1). Since the number of missing values is relatively small, we decided to fill the missing value with the mean (numerical variables) and median (categorical variables).
- We created dummy variables using onehot encoding for categorical variables for better prediction.
- Standardization applied for numerical variables.

Encoding:

- Encoding Loan Status as 0 and 1 instead of Y and N.
- To fix oversampling in our dependent variable, we used the smote techniques to balance out the data.

Figure 1: Number of missing values

```
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
```

MODELS

Choosing algorithms for our analysis:

Logistic Regression

- Since our dependent variable is categorical with 2 possible outcomes (0 and 1), it is better to use a logistic regression then a linear regression for our analysis

Decision Tree

- Decision Tree can be used to handle non-linear dataset and is proven to be applicable to financial dataset.

Random Forest

- Same as Decision Tree, Random Forest is also a good option for our dataset.
- Also, Random Forest is less affected by outliers, which are presented in our numerical variables.

KNN

- KNN algorithm also can produce highly accurate predictions, thus we would want to apply to our dataset.

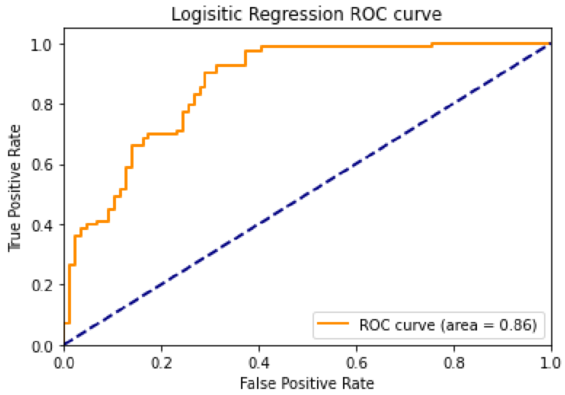
PERFORMANCE

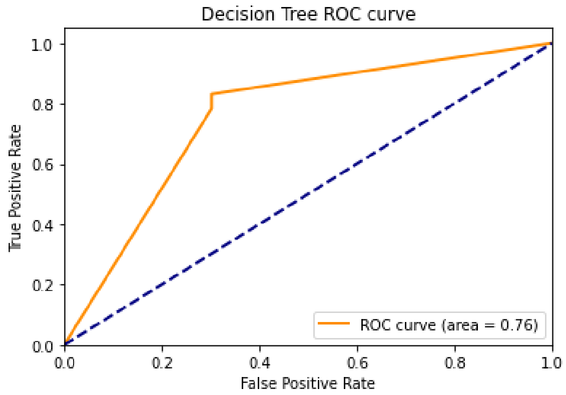
Performance

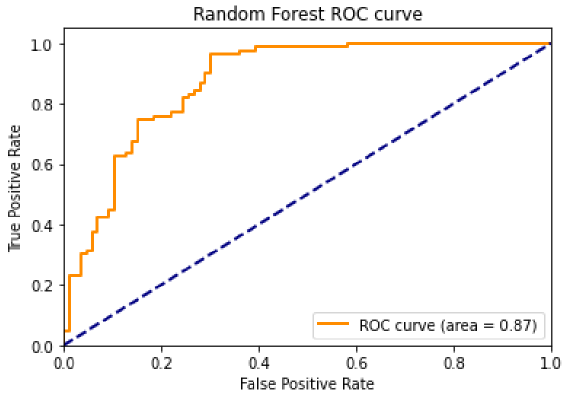
Initial results:

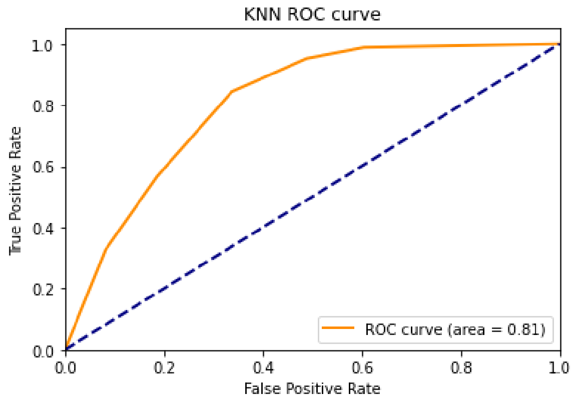
	Accuracy	Sensitivity	Specificity
Logistic Regression	0.799	0.709	0.892
Decision Trees	0.763	0.698	0.831
Random Forest	0.805	0.698	0.916
K-NN	0.751	0.663	0.843

ROC curve:









Parameters:

We use mostly default parameters to set the baseline. Details below:

- Logistic Regression: solver = 'lbfgs', multi_class = 'ovr'.
- Decision Tree: criterion = 'gini', splitter='best', max _depth=15.
- Random Forest: n_estimators=100, max_depth=5, r andom_state=0.
- KNN: n_neighbors = 5

Assessment:

- Based on our initial model assessment, **Random Forest** is the model with highest accuracy.
- However, based on sensitivity, **Logistic Regression** performs the best, which serves the purposes to be more accurately in determining default outcomes.

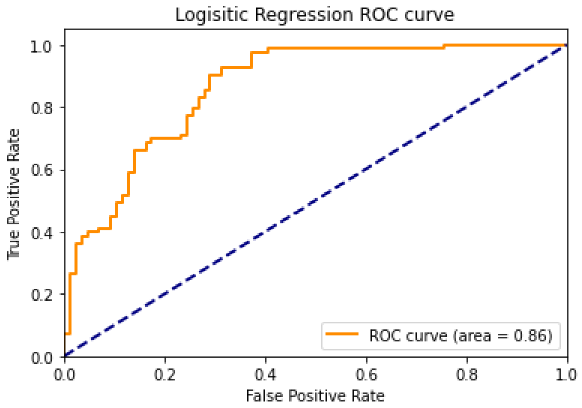
Performance

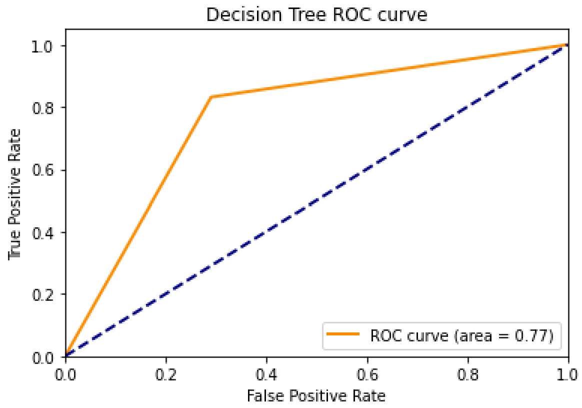
Initial results:

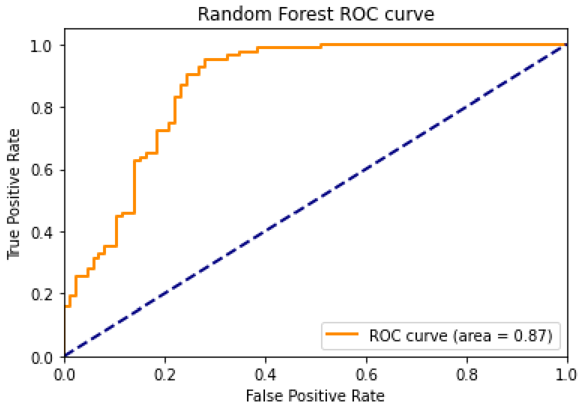
	Accuracy	Sensitivity	Specificity
Logistic Regression	0.799	0.709	0.892
Decision Trees	0.769	0.709	0.831
Random Forest	0.828	0.733	0.928
K-NN	0.751	0.663	0.843

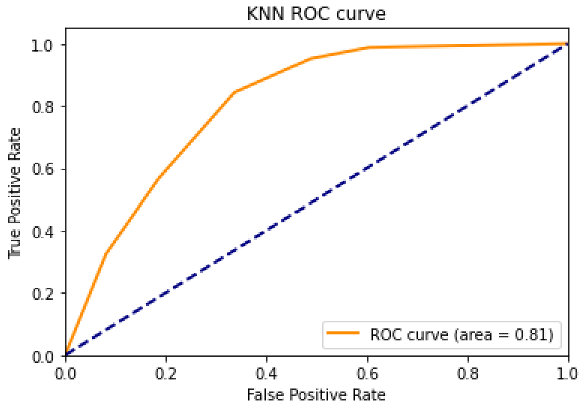
: Improved after parameters tuning

ROC curve:









Parameters Tuning:

- Logistic Regression: change in parameters does not affect prediction.
- Decision Tree: change max_depth from 15 to 21.
- Random Forest: change n_estimator from 100 to 1000, max_depth from 5 to 15.
- KNN: change in parameters does not affect prediction.

Assessment:

- Based on our improved model assessment, **Random Forest** is still the model with highest accuracy.
- Also, **Random Forest** has outperformed Logistic Regression in sensitivity, becoming the best model in terms of all sensitivity, specificity and accuracy.