Mini Project 4

By

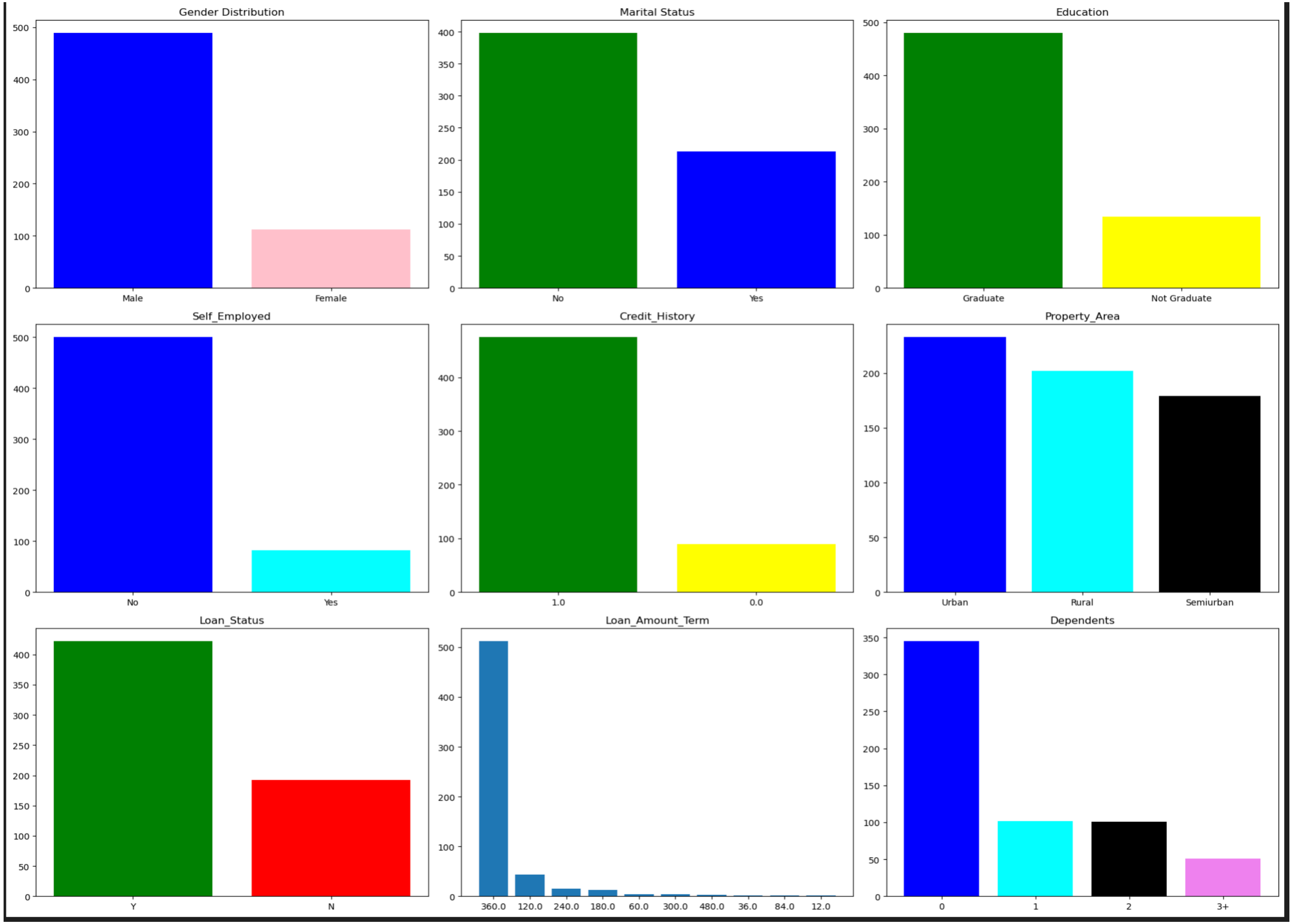
Ali Faisal Raza

Agenda

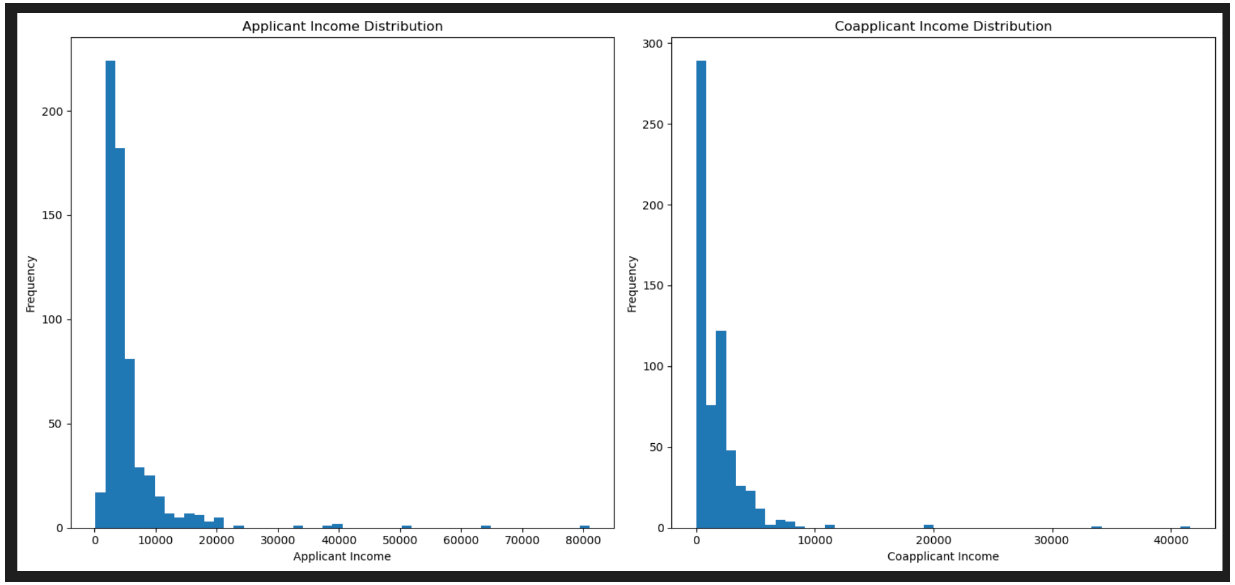
- Problem Formulation - Hypothesis
- Data Cleaning
- Final Feature Set
- Pipelining and Modeling – In progress
- Deployment of Model on Flask - upcoming
- Questions?

Notable Frequency Distributions

- Predominantly more males applicants than females.
- Majority are educated.
- Majority are salaried employees.
- Majority have no dependents.
- Majority possess credit history.
- The preference for property areas is evenly distributed.
- Majority loan terms are for 30 years.

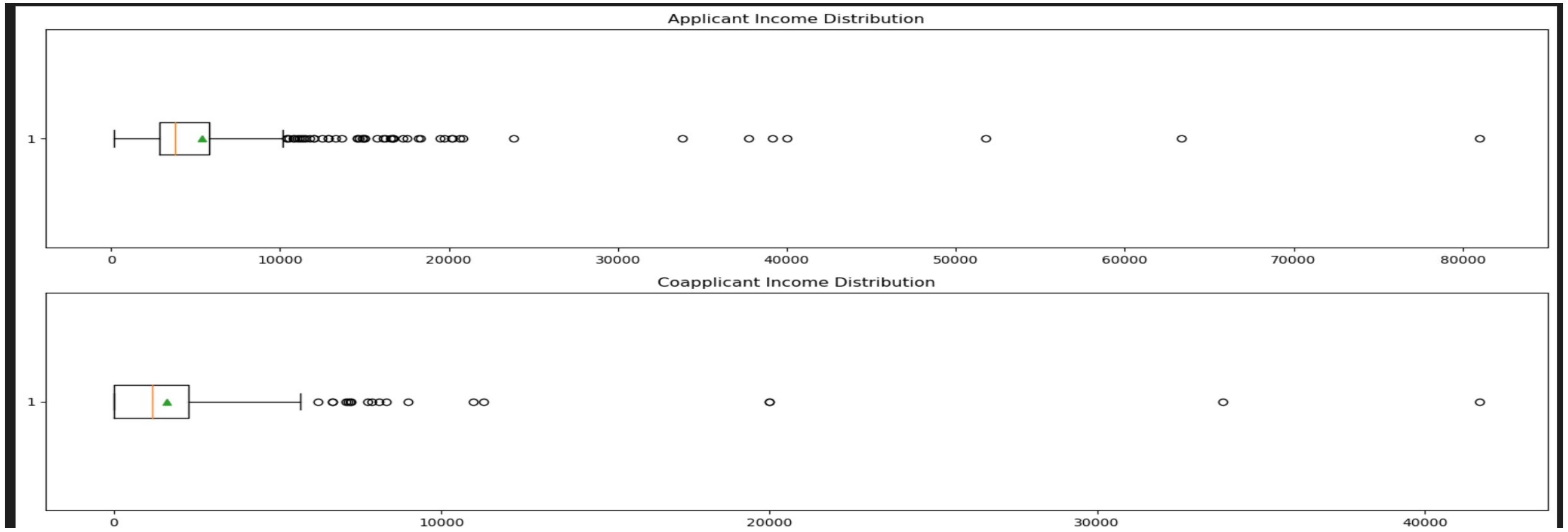


Nature of Income Distribution

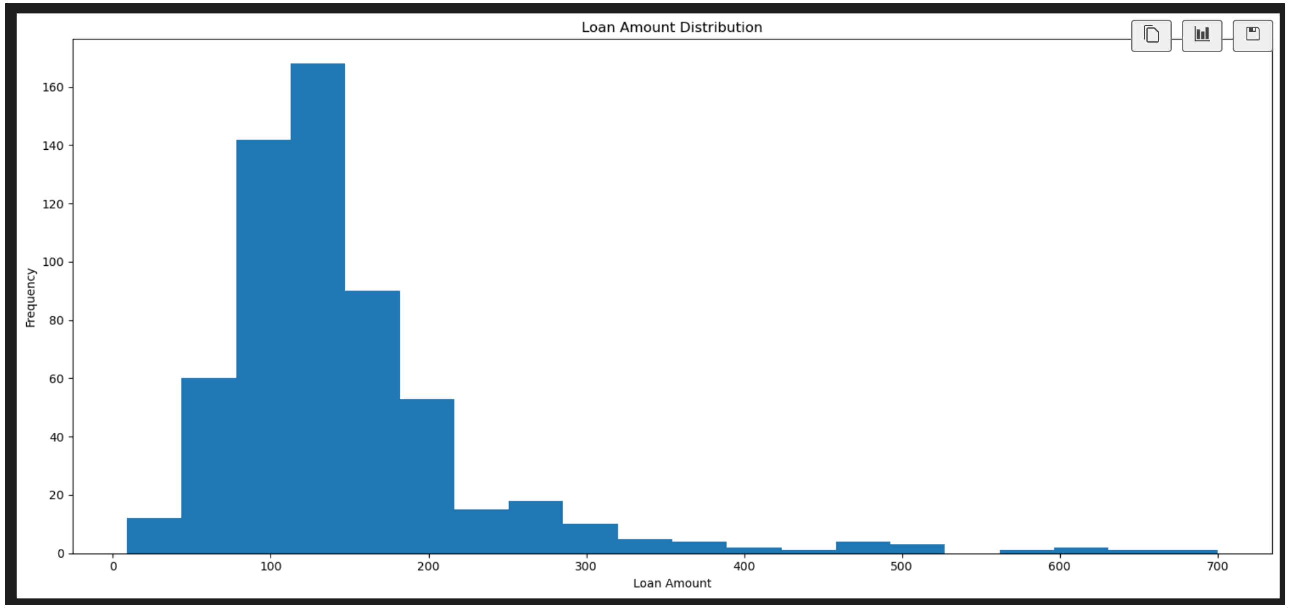


Skewed on left side.

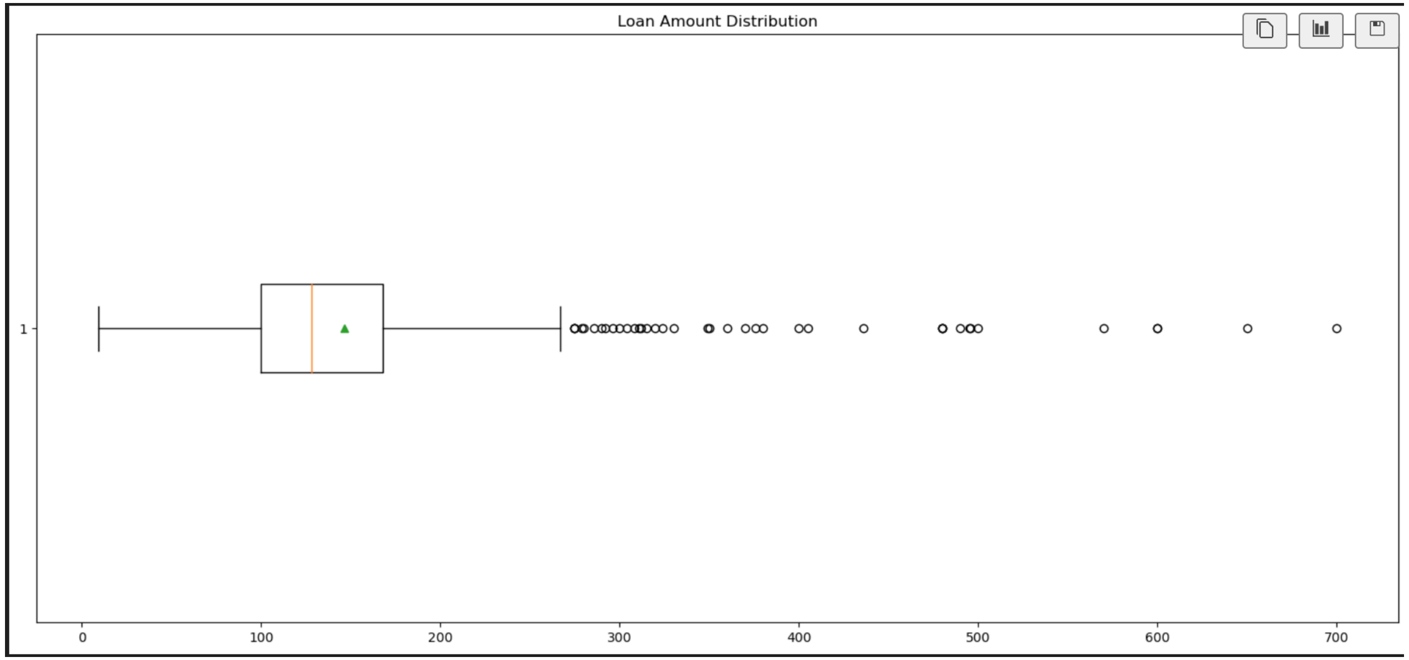
Too many outliers.



Loan Amount Distribution



Skewed on left side.



Too many outliers.

Problem Formulation and Hypothesis

To predict whether a customer is eligible for loan given his profile and requested loan amount and term.

Hypothesis:

a) Persons with high income have higher probability for loan approval.

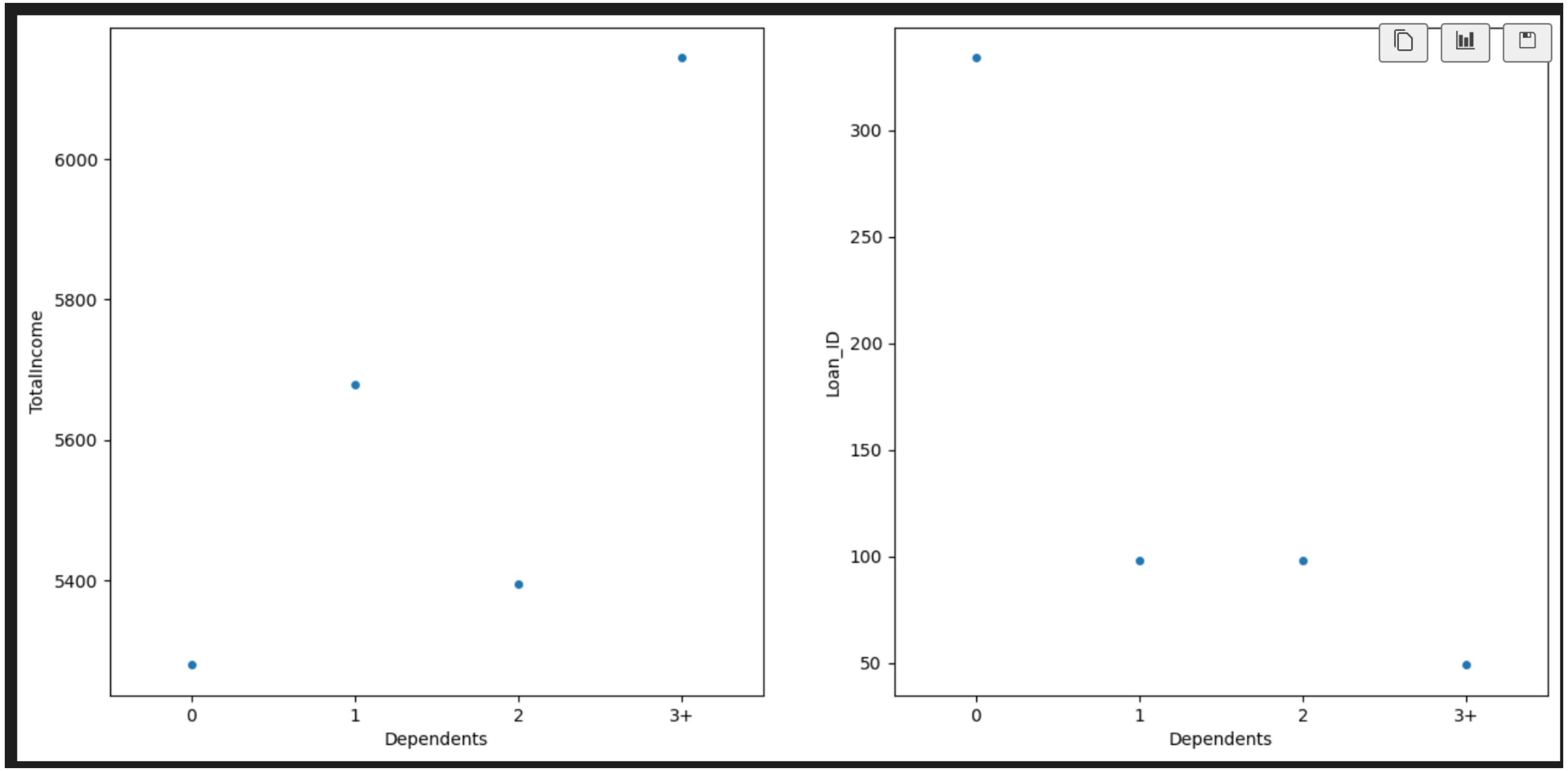
b) Person with existing credit history have higher probability for loan approval.

Data Cleaning – Imputing Missing Values

Technique:

- Correlate each affected category variable against income and infer pattern, if any.
- Impute most common value as default in case of no discernible trend.

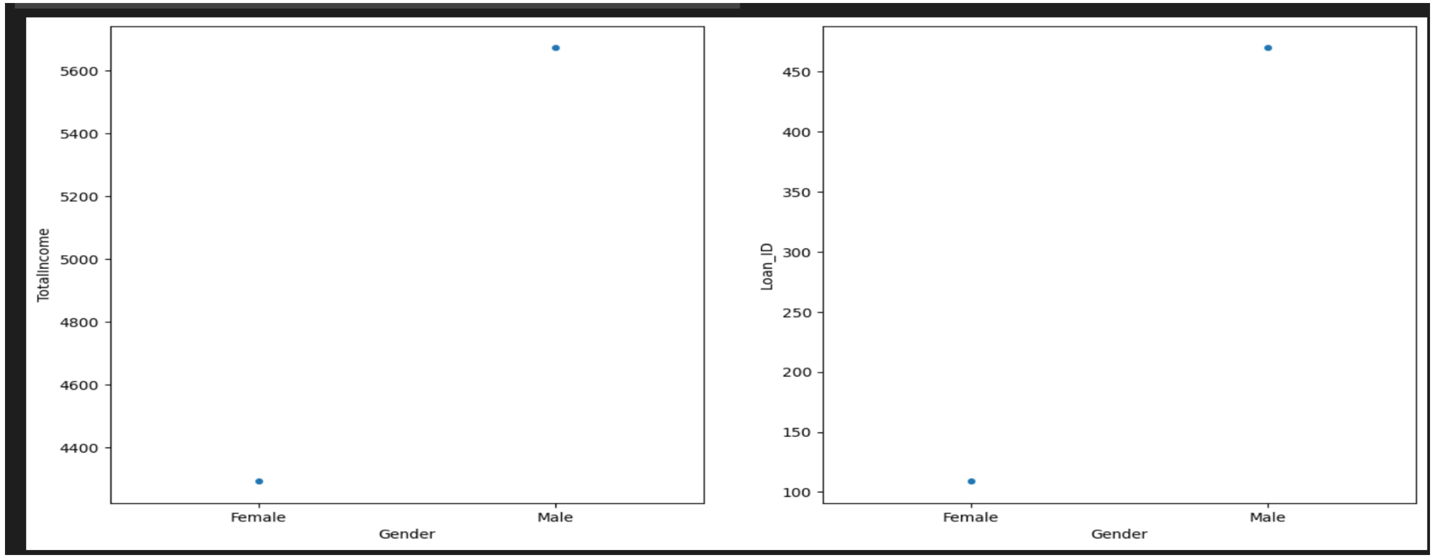
➢ Imputing missing ‘Dependents’ values



Assign all cases with total income higher than 6000 with 3+ dependents and rest with 0.

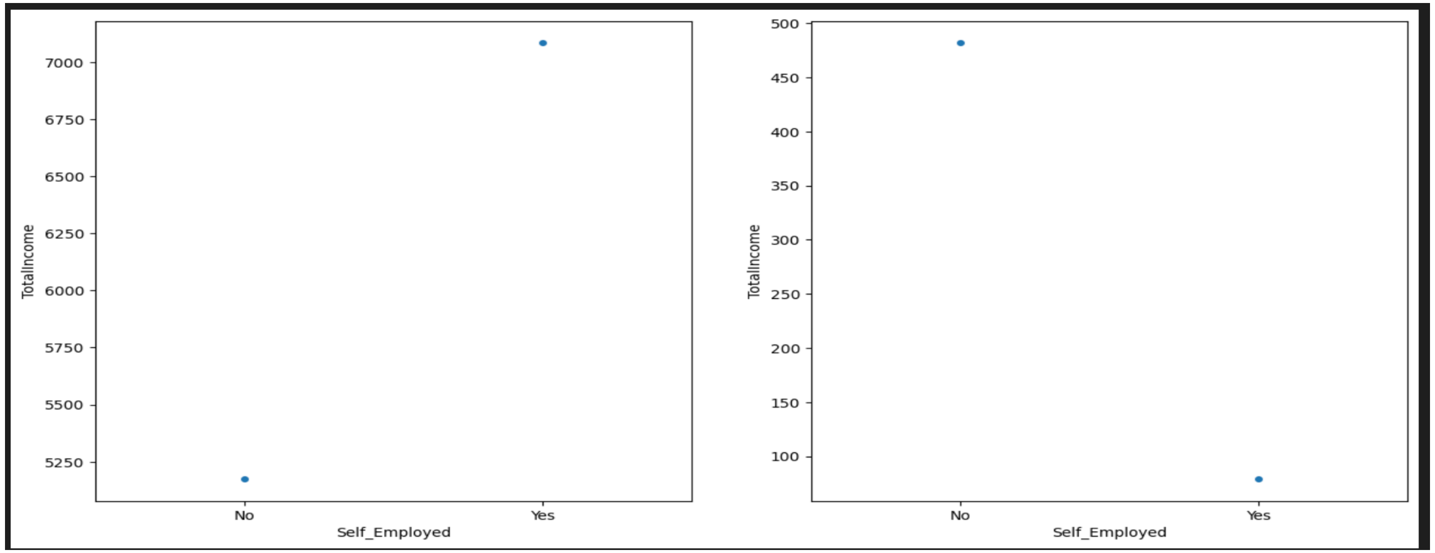
Data Cleaning – Imputing Missing Values

- Imputing ‘Gender’



Assign all cases with total income higher than 4600 under ‘Male’ bucket and rest under ‘Female’ bucket.

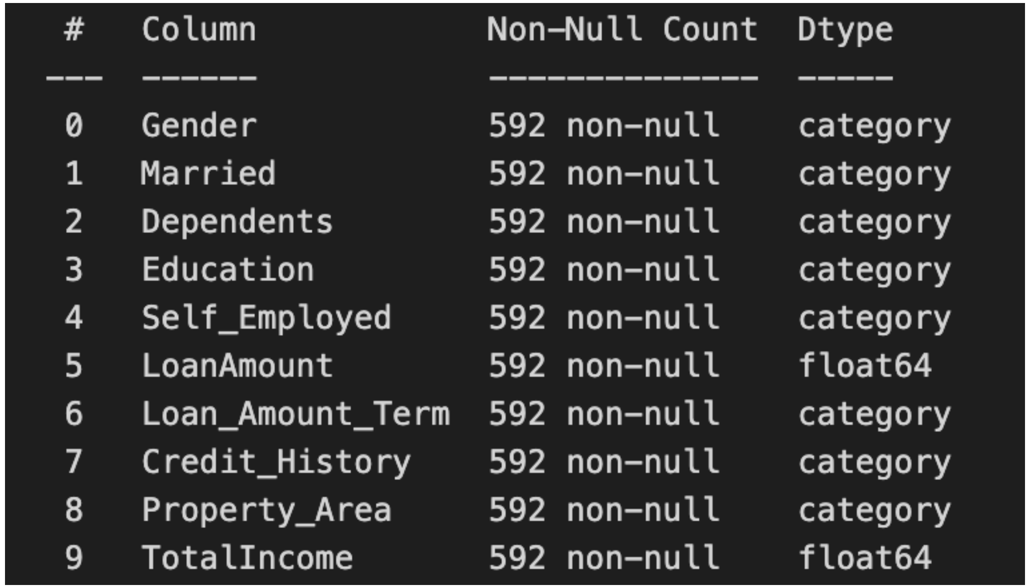
- Imputing ‘Self Employment’ factor



Assign all cases with total income higher than 6750 under ‘self-employed’ bucket and rest under ‘non self-employed’ bucket.

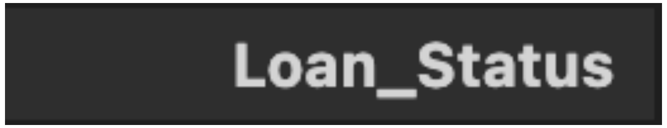
Final Feature Set

Input Features



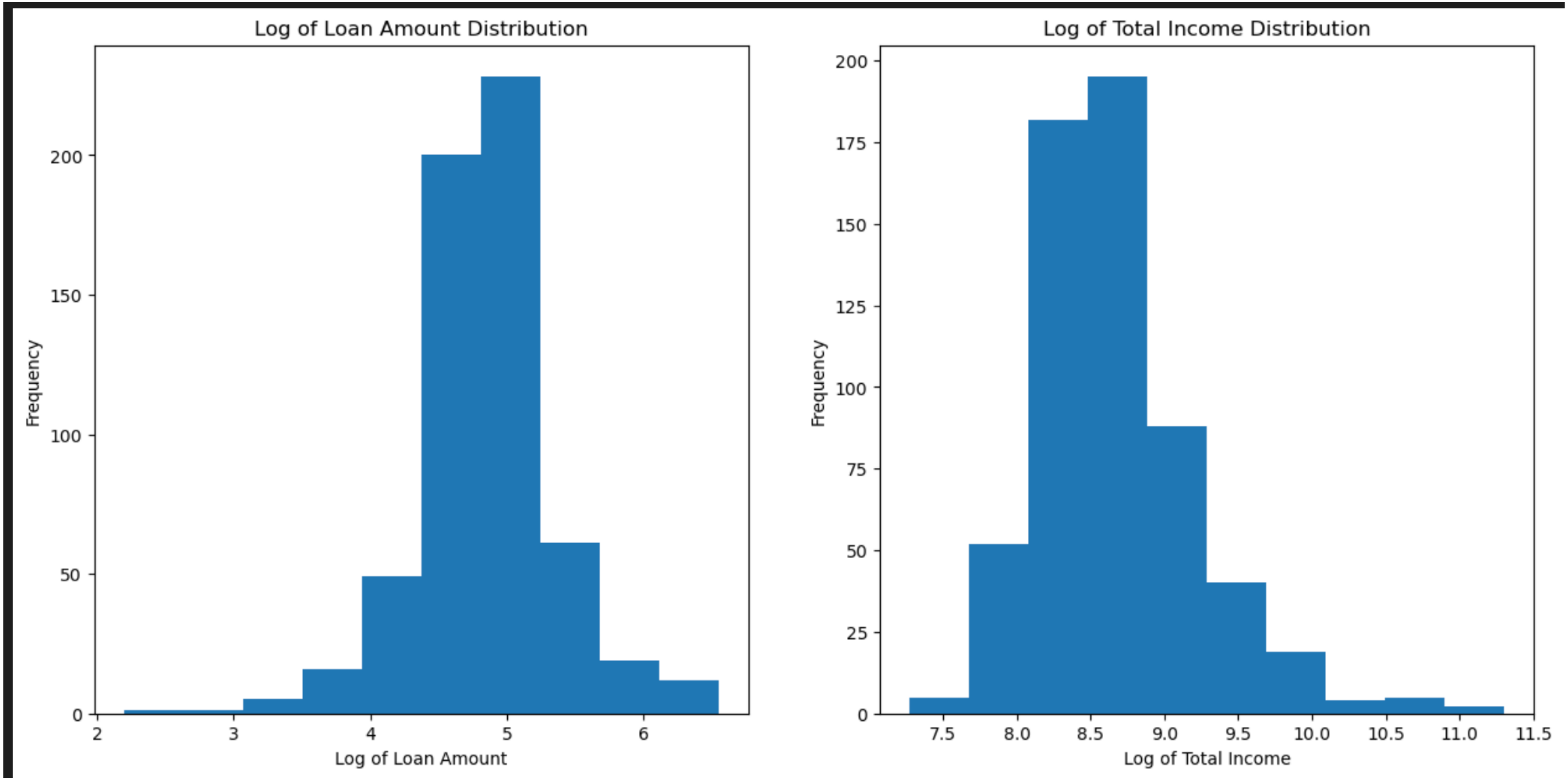
```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Gender             592 non-null    category
 1   Married            592 non-null    category
 2   Dependents         592 non-null    category
 3   Education          592 non-null    category
 4   Self_Employed      592 non-null    category
 5   LoanAmount         592 non-null    float64
 6   Loan_Amount_Term   592 non-null    category
 7   Credit_History     592 non-null    category
 8   Property_Area      592 non-null    category
 9   TotalIncome        592 non-null    float64
```

Target Variable



Loan_Status

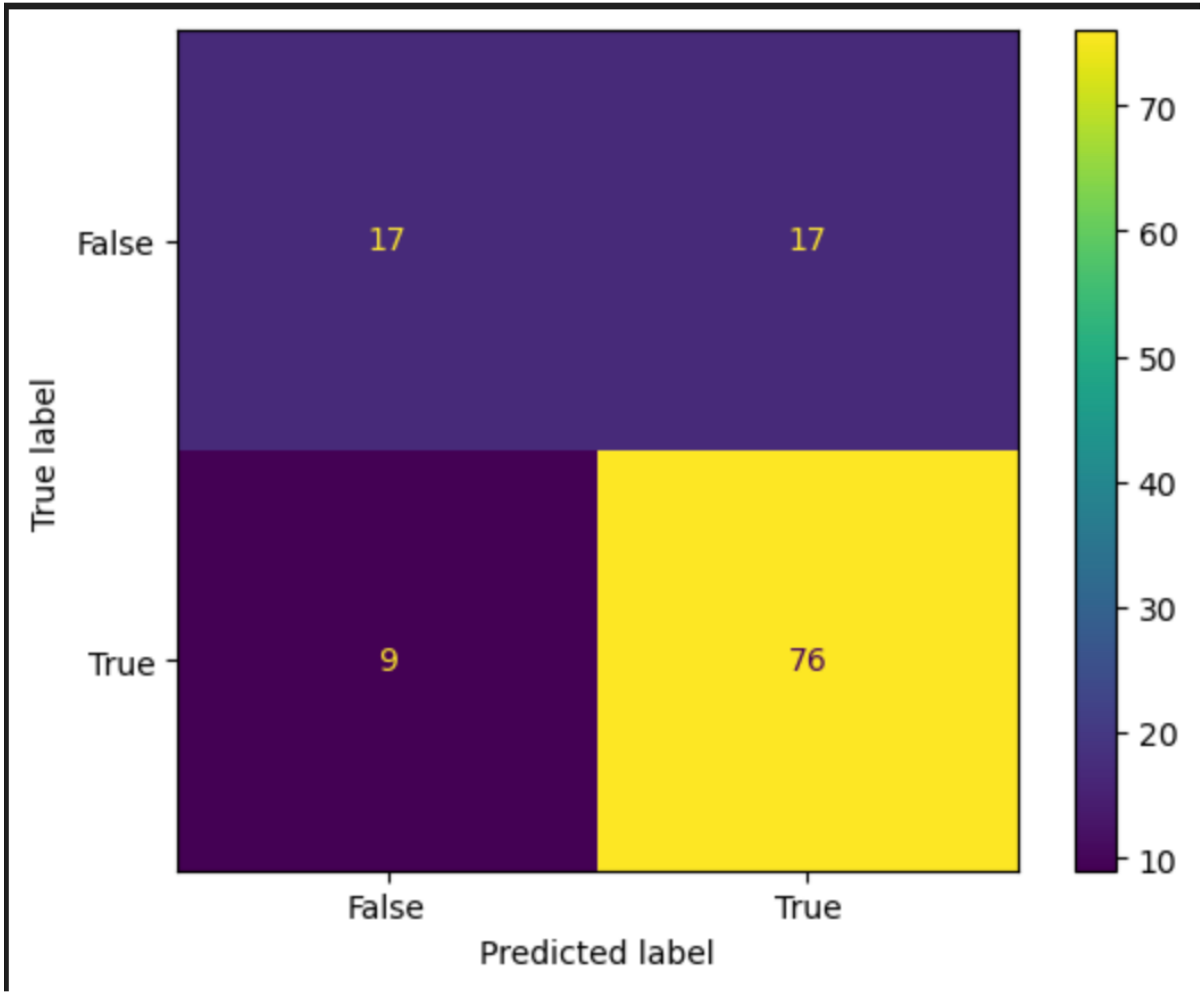
Applying Log Transformation on Income and Loan to approximate to normal distribution.



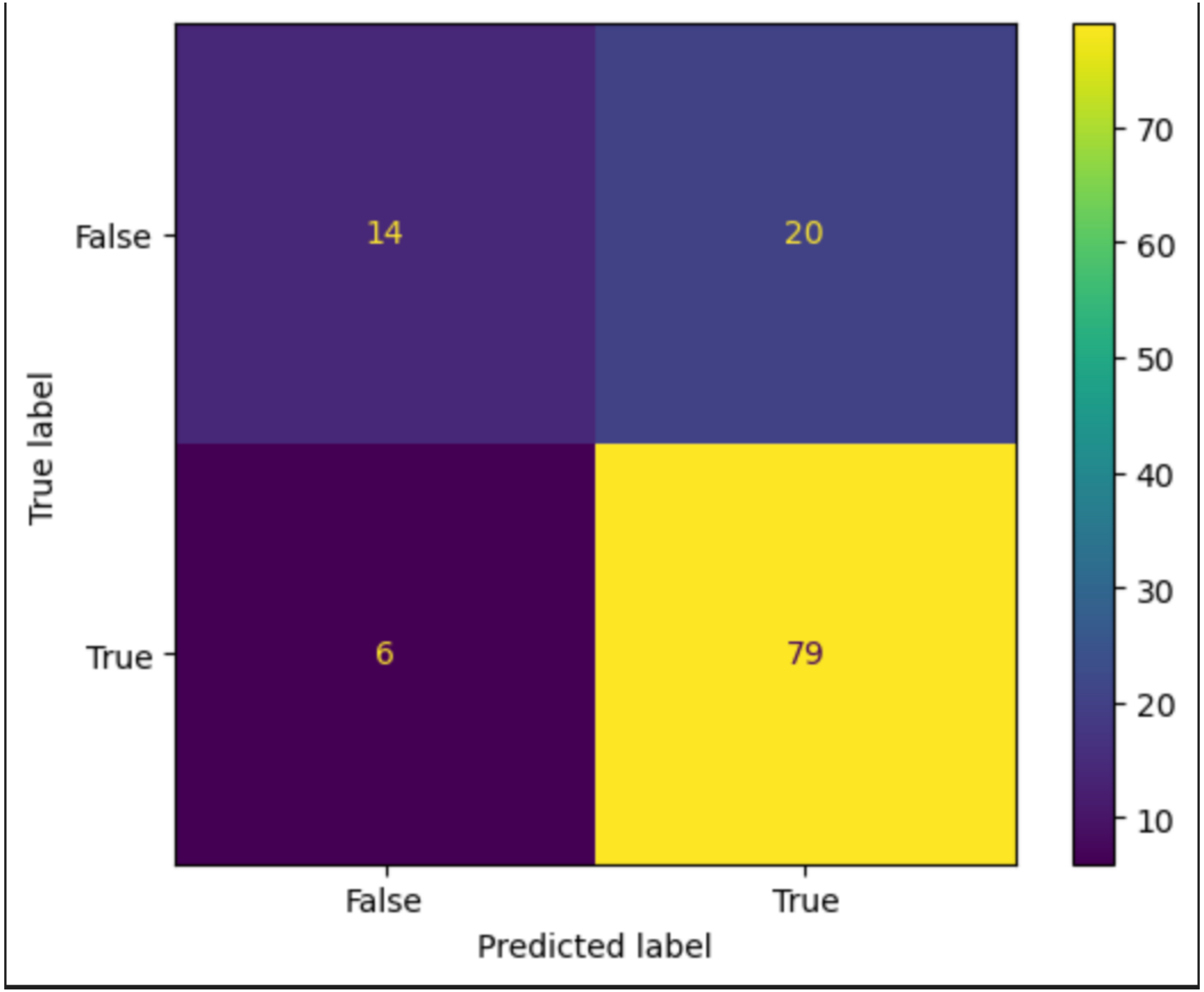
Applying Grid Search for Hyperparameter Tuning and Checking Model Accuracy

Model	Model Accuracy
Ridge Classifier	75.63%
Random Forest Classifier	78.15%
XGBoost Classifier	77.31%

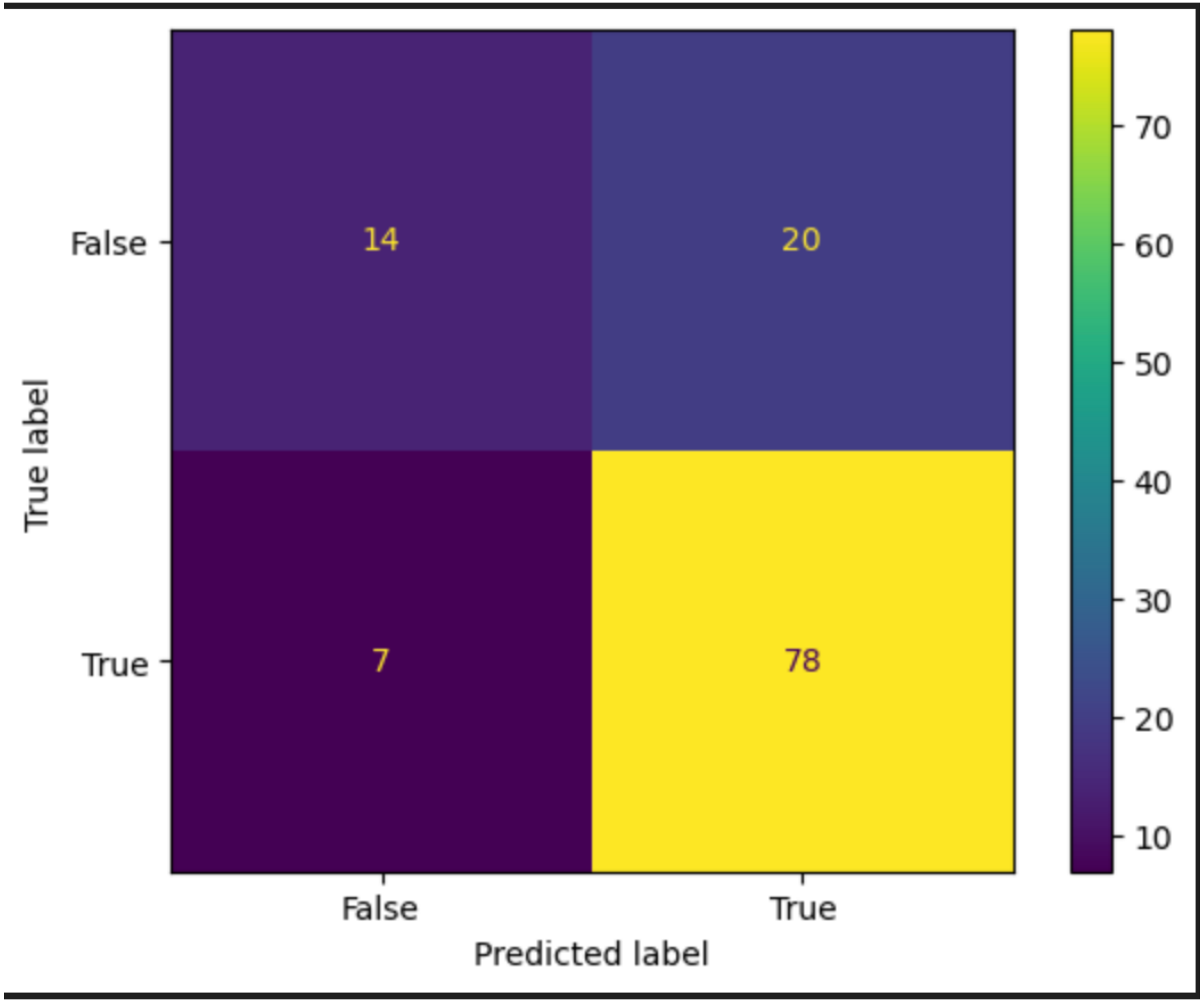
Confusion Matrix – Ridge Classifier



Confusion Matrix – Random Forest Classifier



Confusion Matrix – XGBoost Classifier



?