SVKM'S NMIMS Deemed to be UNIVERSITY
NILKAMAL SCHOOL OF MATHEMATICS, APPLIED STATISTICS & ANALYTICS

LOAN DEFAULT PREDICTION

USING SUPERVISED LEARNING ALGORITHMS

Mentored by Prof. Swasti Desai

A010 Breema Alias
A016 Sharmeen Shaikh
A021 Heena Gagwani

FLOW OF THE PRESENTATION

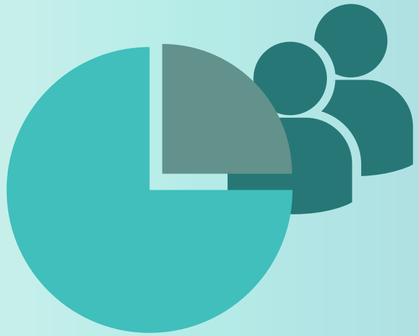
INTRODUCTION

Loan Default Prediction is the practice of using data analysis and predictive modeling to anticipate whether borrowers are likely to default on their loans. This enables institutions to

- make informed lending decisions
- reduce potential losses
- maintain a healthy loan portfolio.

PROBLEM STATEMENT

To identify high-risk borrowers and mitigate financial losses for institutions.



OBJECTIVE

To develop a smart system that predicts which borrowers may struggle to repay their loans. Using historical data and advanced techniques, the aim is to assist banks in making safer lending decisions, reducing the risk of loan defaults, and improving overall risk management..

RESEARCH PAPER

Nalawade, S., Andhe, S., Parab, S., & Sankhe, A. (2022).

Loan Approval Prediction

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

Loan Approval Prediction

Shubham Nalawade[1], Suraj Andhe[1], Siddhesh Parab[1], Prof. Amruta Sankhe[2]

[1] BE Student, Information Technology, Atharva College of Engineering, Mumbai
[2] Assistant Professor, Information Technology, Atharva College of Engineering, Mumbai

---

Abstract – *Today a lot of people/companies are applying for bank loans. The core business part of every bank is the distribution of loans. The main objective of the banking sector is to give their assets in safe hands. But the banks or the financial companies take a very long time for the verification and validation process and even after going through such a regress process there is no surety that whether the applicant chosen is deserving or not. To solve this problem, we have developed a system in which we can predict whether the applicant chosen will be a deserving applicant for approving the loan or not. The system predicts on the basis of the model that has been trained using machine learning algorithms. We have even compared the accuracy of different machine learning algorithms. We got a percentage of accuracy ranging from 75-85% but the best accuracy we got was from Logistic Regression i.e., 88.70% The system includes a user interface web application where the user can enter the details required for the model to predict. The drawback of this model is that it takes into consideration many attributes but in real life sometimes the loan application can also be approved on a single strong attribute, which will not be possible using this system.*

Key Words: Machine Learning, Loan Approval Prediction, Web Application, Bank, Algorithms, Random Forest, Naïve Bayes, Logistic Regression, K Nearest Neighbor, Decision Tree.

borrower is a defaulter or not. No doubt the manual process will be more accurate and effective, but this process cannot work when there are a large number of loan applications at the same time. If there occurs a time like this, then the decision-making process will take a very long time and also lots of manpower will be required. If we are able to do the loan prediction it will be very helpful for applicants and also for the employees of banks. So, the task is to classify the borrower as good or bad i.e., whether the borrower will be able to pay the debts back or not. This can be done with the help of machine learning algorithms.

2. LITERATURE SURVEY

In [1] they have used only one algorithm; there is no comparison of different algorithms. The algorithm used was Logistic Regression and the best accuracy they got was 81.11%. The final conclusion reached was only those who have a good credit score, high income and low loan amount requirement will get their loan approved. Comparison of two machine learning algorithms was made in [2]. The two algorithms used were two class decision jungle and two class decision and their accuracy were 77.00% and 81.00% respectively. Along with these they also calculated parameters such as Precision, recall, F1 score and AUC. The [3] shows a comparison of four

DATA SNAPSHOT

614 observations of bank users of a renowned bank

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome
LP001002	Male	No	0	Graduate	No	5849
LP001003	Male	Yes	1	Graduate	No	4583
LP001005	Male	Yes	0	Graduate	Yes	3000
LP001006	Male	Yes	0	Not Graduate	No	2583
LP001008	Male	No	0	Graduate	No	6000
LP001011	Male	Yes	2	Graduate	Yes	5417
LP001013	Male	Yes	0	Not Graduate	No	2333
LP001014	Male	Yes	3+	Graduate	No	3036

CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0		360	1	Urban	Y
1508	128	360	1	Rural	N
0	66	360	1	Urban	Y
2358	120	360	1	Urban	Y
0	141	360	1	Urban	Y
4196	267	360	1	Urban	Y
1516	95	360	1	Urban	Y
2504	158	360	0	Semiurban	N
1526	168	360	1	Urban	Y

Source: Kaggle

DATA

CATEGORICAL COLUMNS

1. GENDER

MALE, FEMALE

2. MARRIED

YES, NO

3. DEPENDENTS

0, 1, 2, 3+

4. EDUCATION

NON-GRADUATE, GRADUATE

5. SELF_EMPLOYED

YES, NO

6. CREDIT_HISTORY

1- GOOD, 0-BAD

7. PROPERTY_AREA

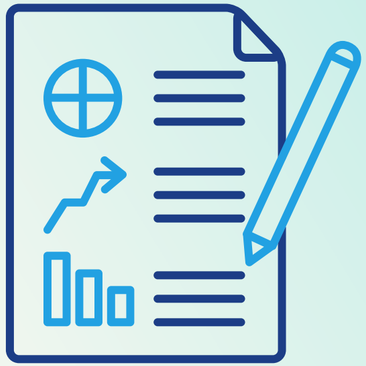
NUMERICAL COLUMNS

1. APPLICANT INCOME

2. COAPPLICANT INCOME

3. LOAN AMOUNT

4. LOAN_AMOUNT_TERM



MISSING VALUES

```
df.isnull().sum()
```

```
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
dtype: int64
```

Categorical columns for which missing values found are-

- Gender
- Married
- Dependents
- Self-Employed
- Credit-History

Numerical columns found are-

- Loan Amount
- Loan_Amount_Term

TREATING MISSING VALUES

Since 5 categorical columns described with 2 3 categories, we replace it with mode

113 total missing values in total of the 5 Categorical columns

MODE IMPUTATION

1. Finds the largest/highest ocurring value in each column.
2. Replaces it with the highest range .

TREATING MISSING VALUES

36 missing values together in the 2 Numerical columns

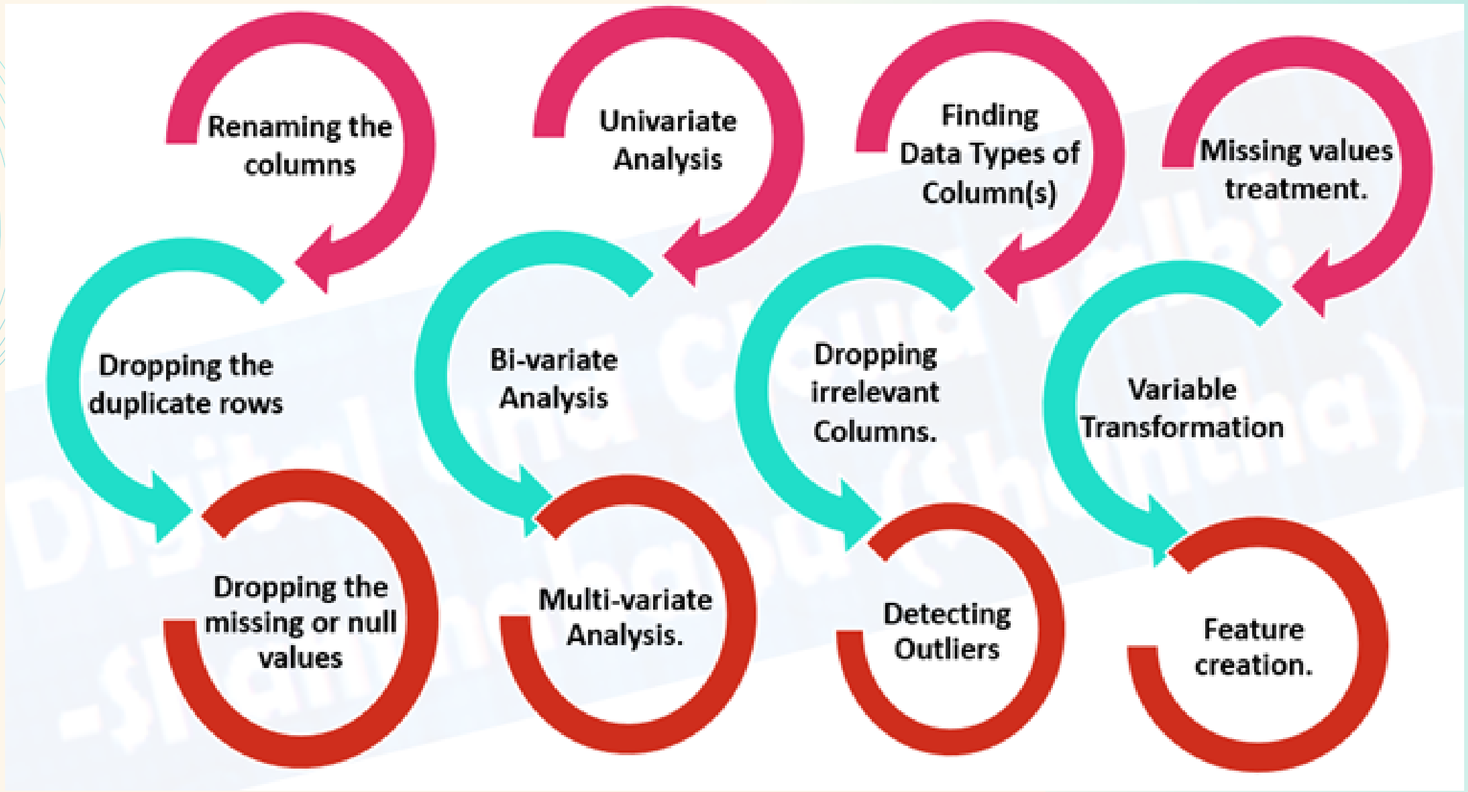
2 numerical columns left for which mean imputation can be done to replace values.

MEAN IMPUTATION

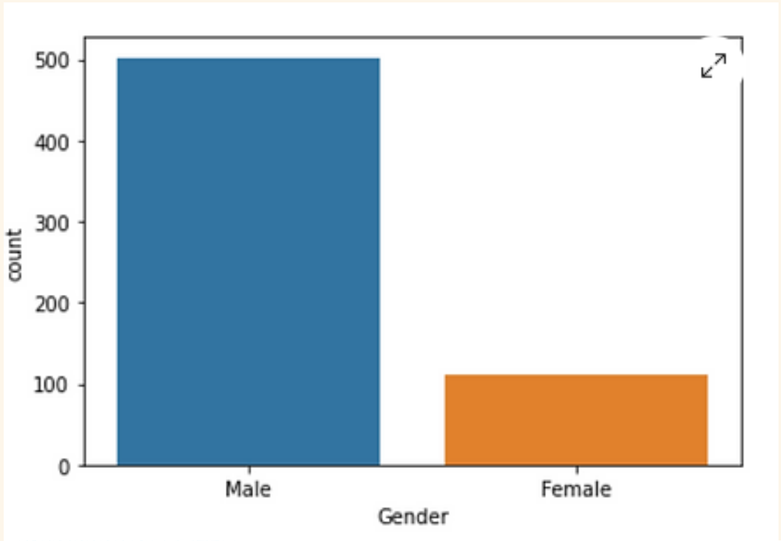
1. Finds the average of all the values of all cells together.
2. Replaces it with the mean of it for each missing value .

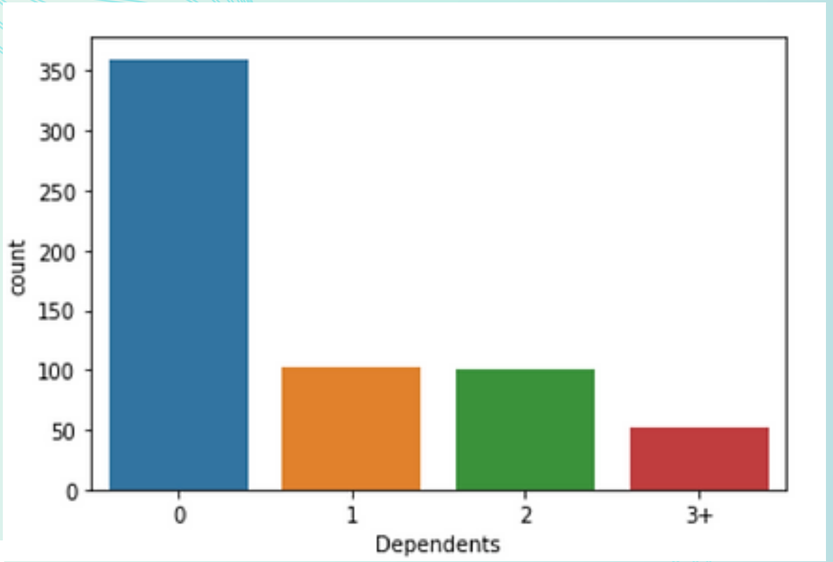
EXPLORATORY DATA ANALYSIS(EDA)

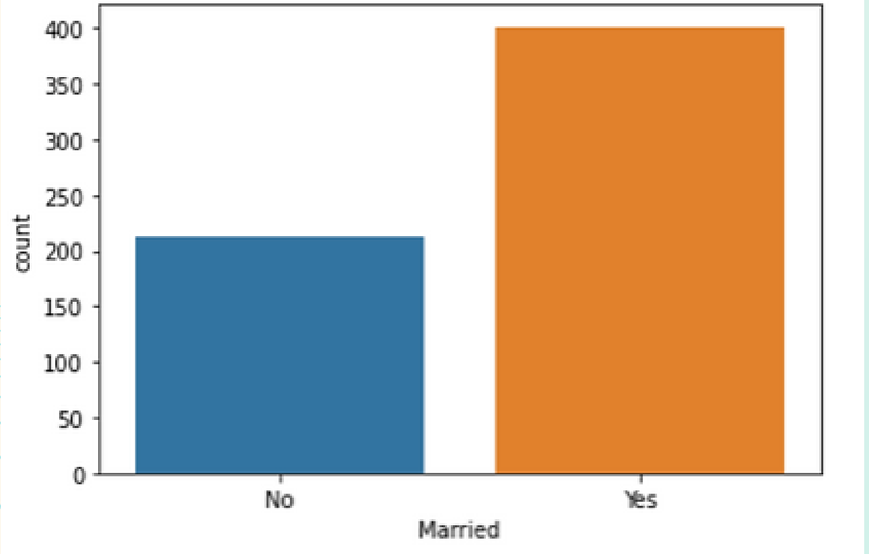
- Process of investigating a dataset to find patterns and anomalies (outliers)
- Entails producing summary statistics for the dataset's numerical data and developing various graphical representations to aid with data comprehension.



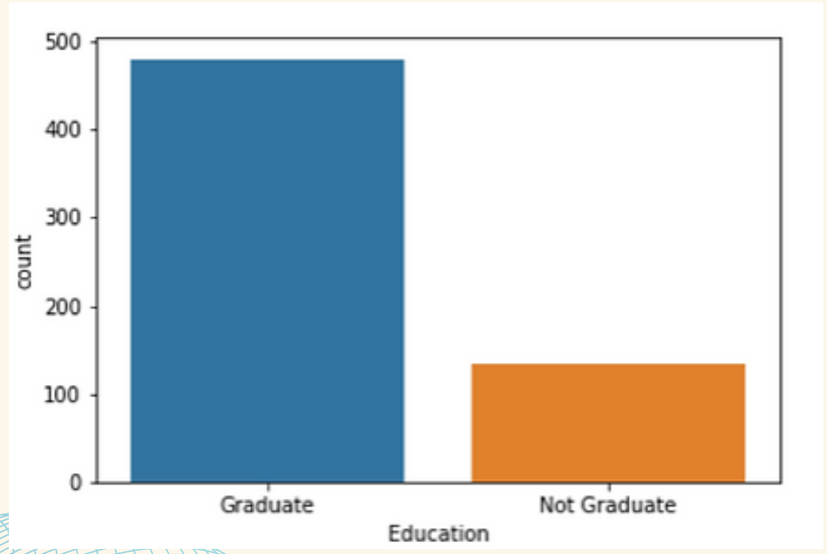
VISUALIZING THE DATA (CATEGORICAL ATTRIUTES)

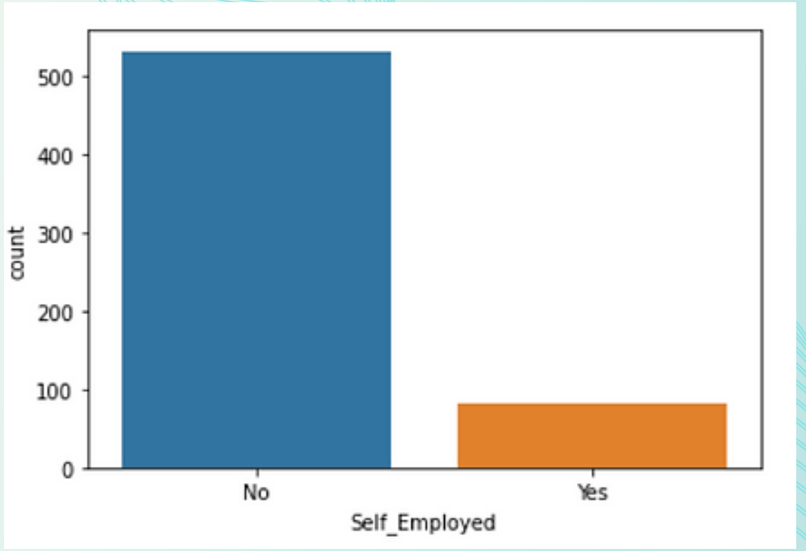


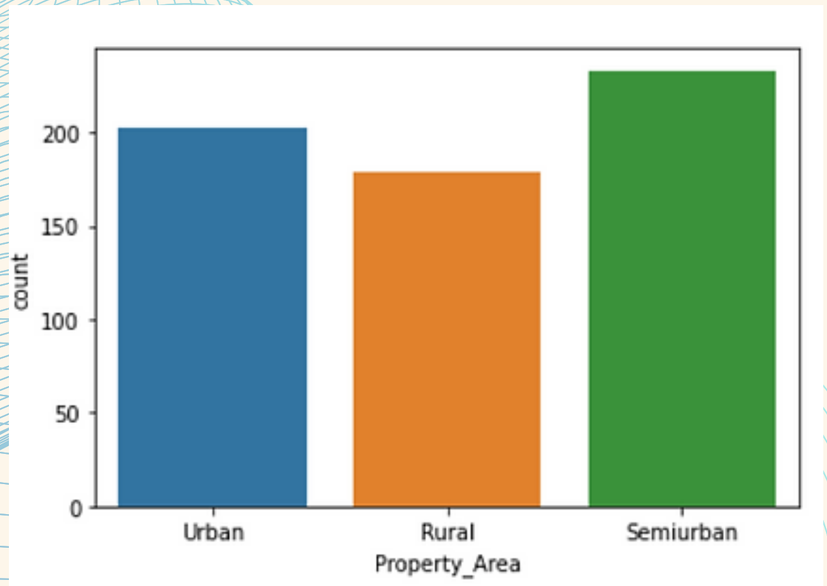


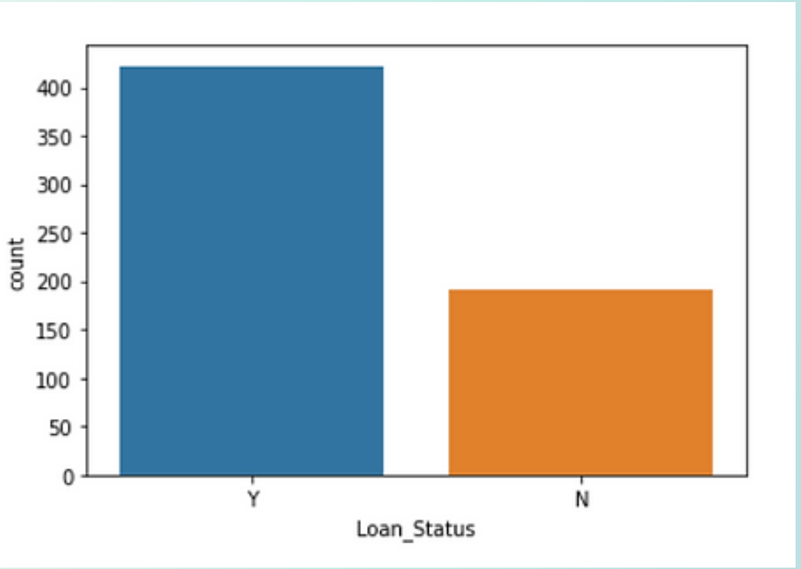


VISUALIZING THE DATA (CATEGORICAL ATTRIUTES)

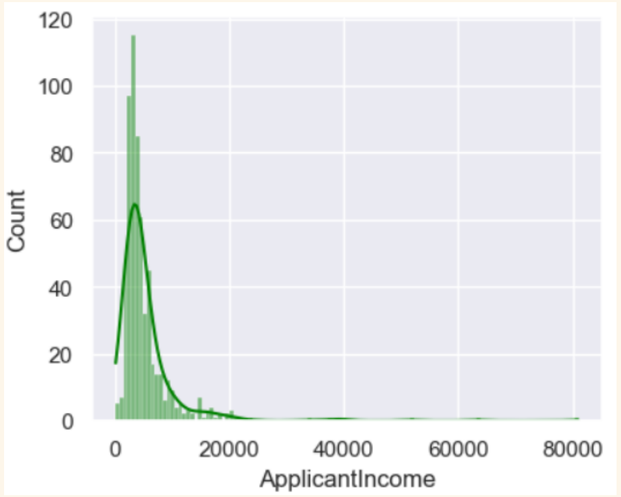


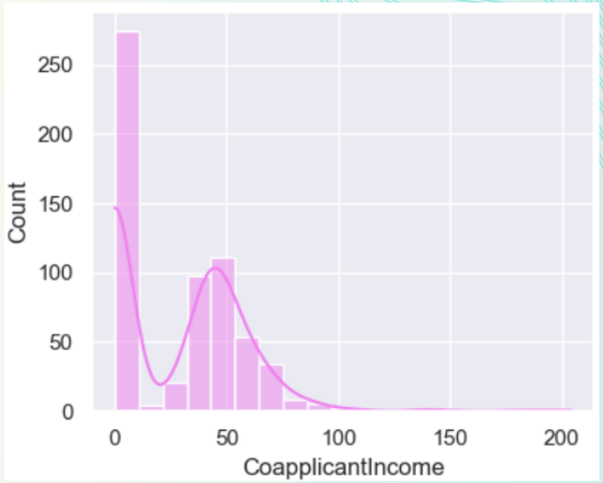


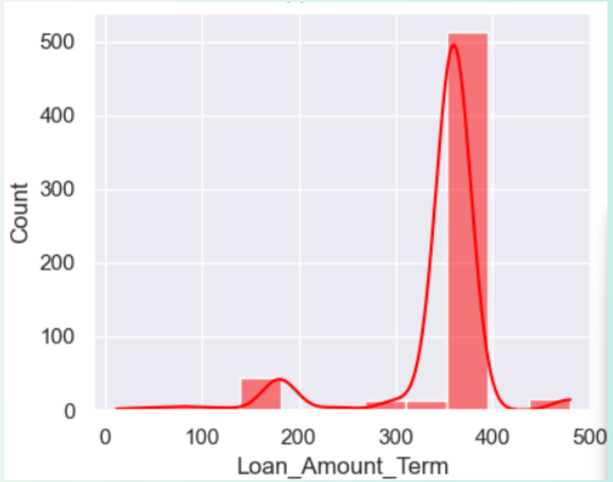




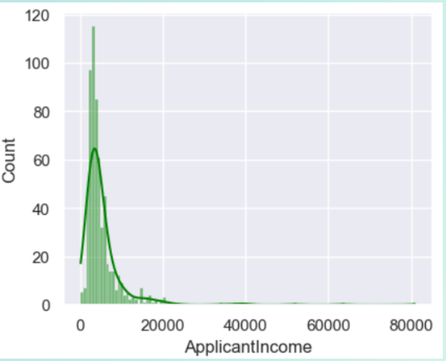
VISUALIZING THE DATA (NUMERICAL ATTRIUTES)

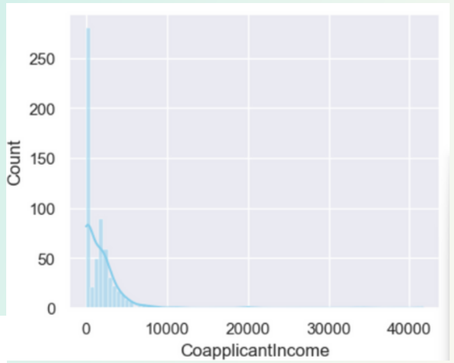


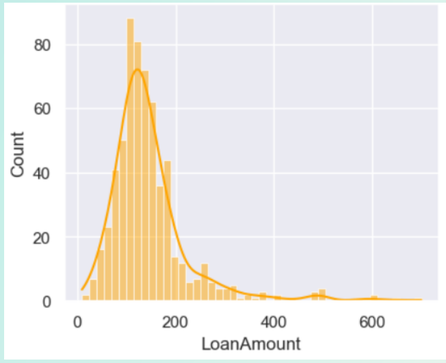


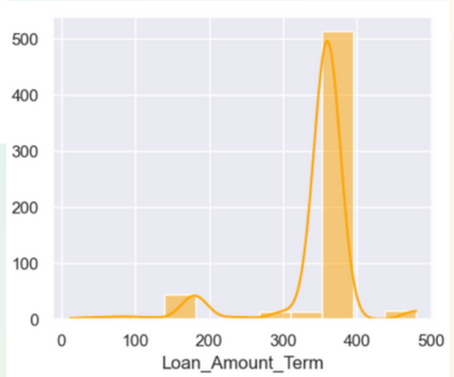


DATA TRANSFORMATION

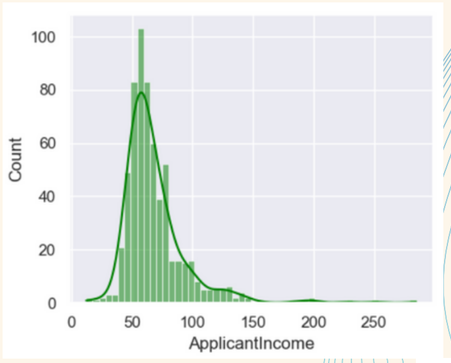


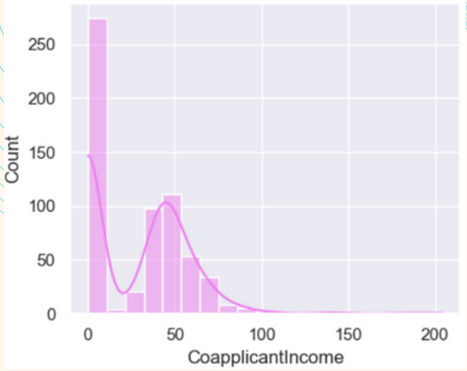


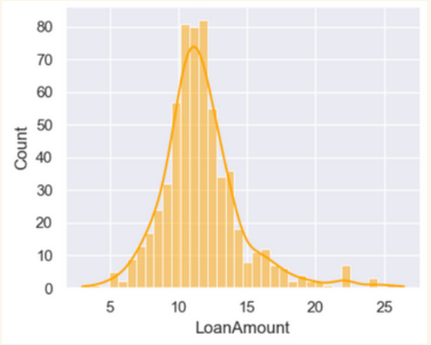


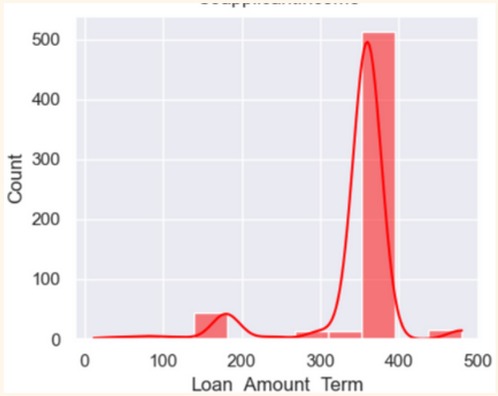


- There was a lot of skewness as there will always be a few people who have high income and loan amount.
- square root transformation reduces skewness and enhances the clarity and usefulness of the visualization.
- The accuracy of the machine learning models can improve with reduced skewness





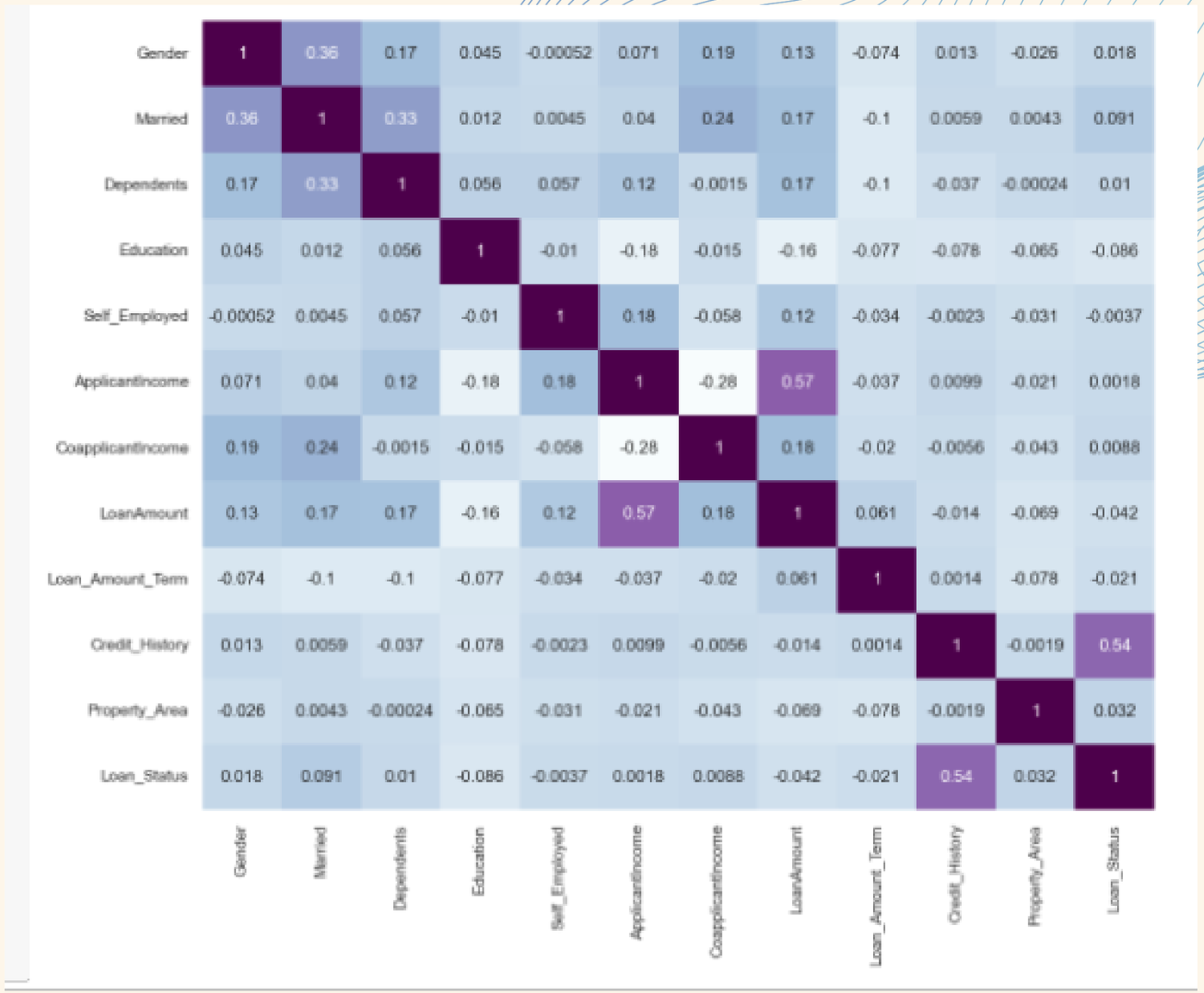




CORRELATION MATRIX

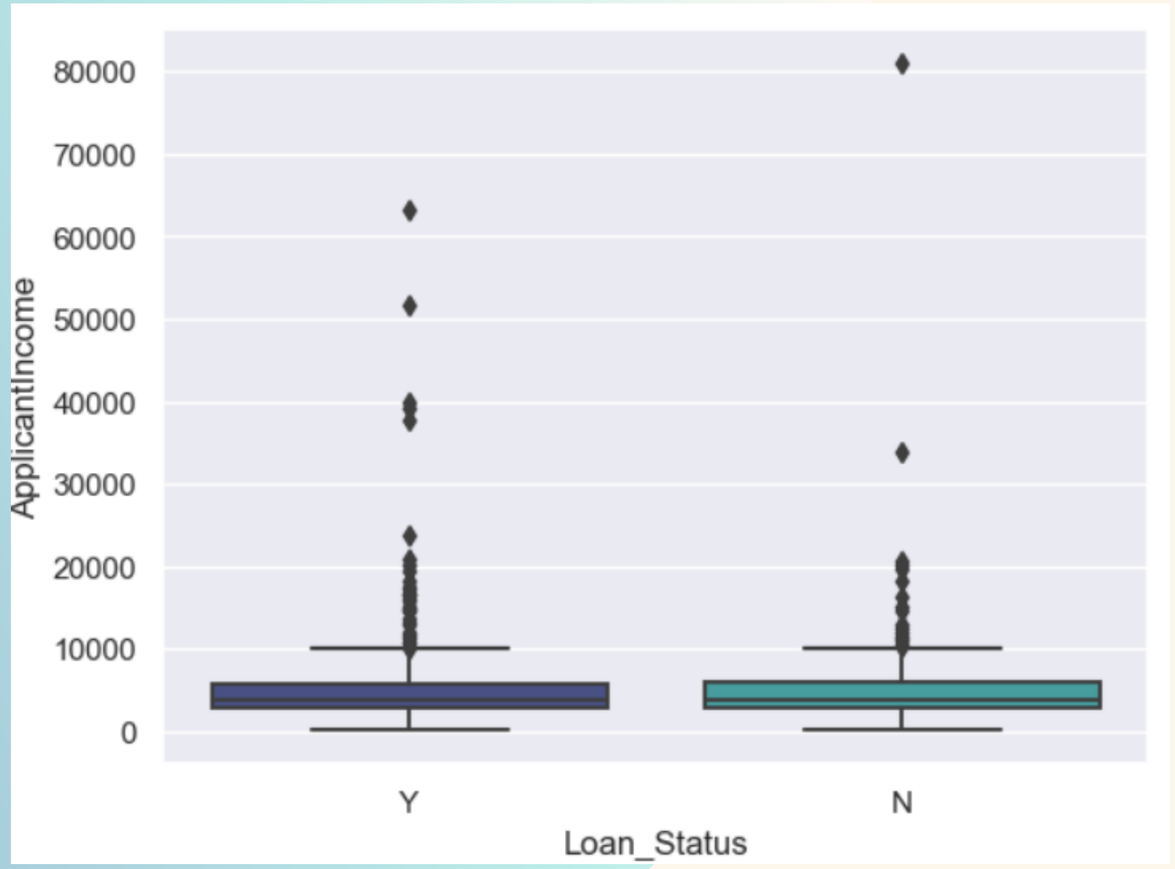
High correlation between following variables -

- Loan Amount and applicant income
- Credit history and Loan Status
- Married and dependents
- Gender and married



OUTLIERS DETECTION

Applicant income seem to have outliers with target variable loan status



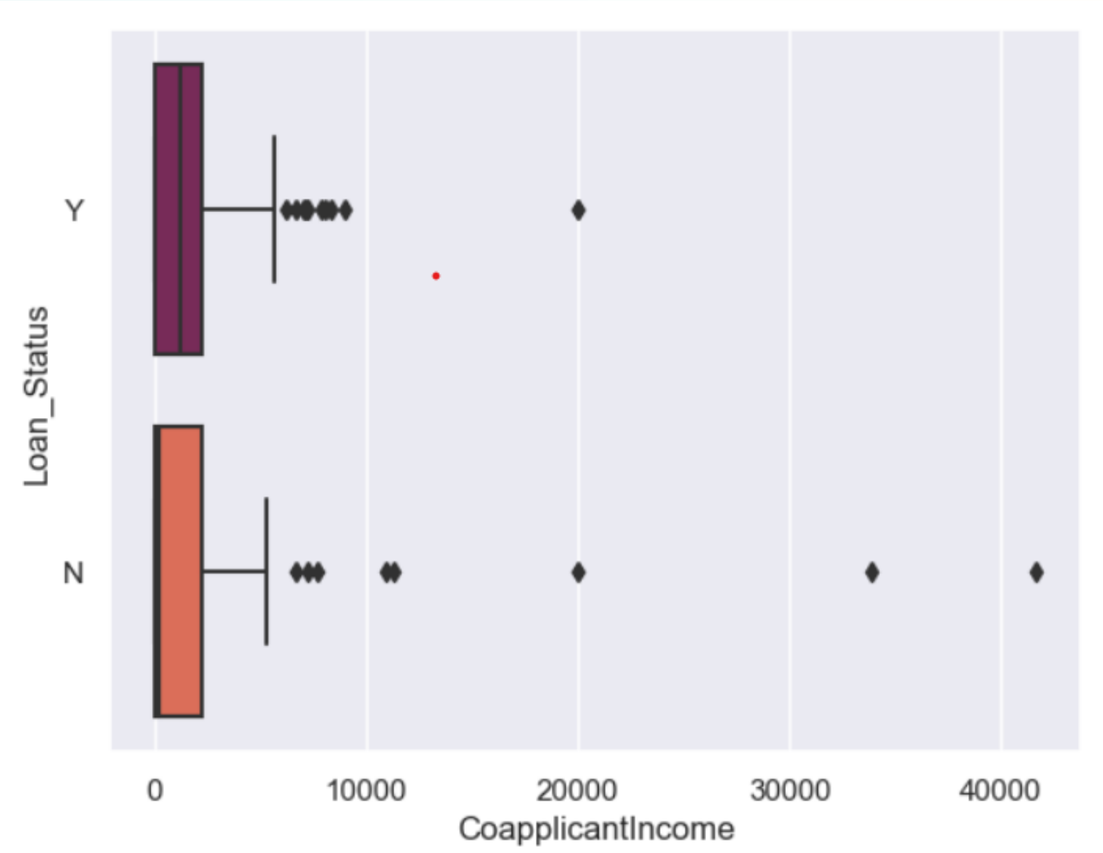
Unexpected outlier when applicant‘s income is near 80000 but still loan getting rejected

Key reason considered:

- Poor credit history

OUTLIERS DETECTION

Co-applicant income seem to have outliers with target variable loan status



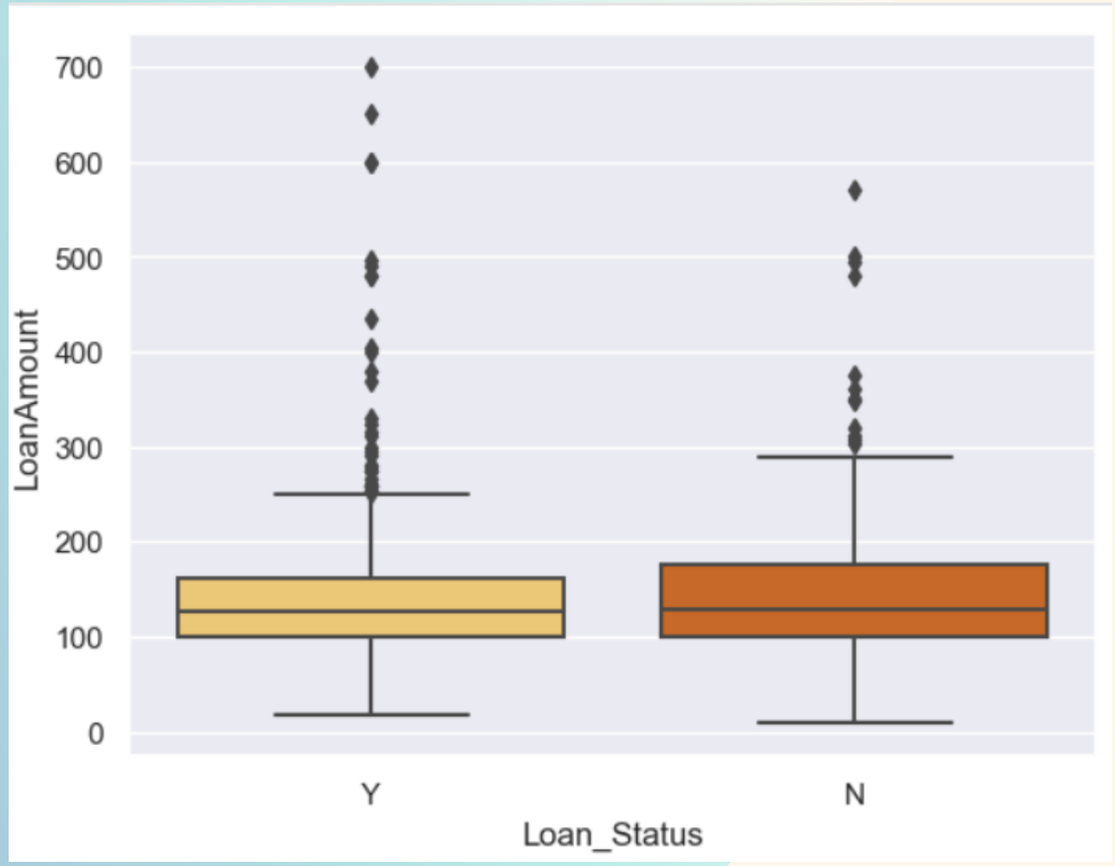
Unexpected outlier when applicant‘s income is near 30000-40000 but still loan getting rejected

Key reasons considered:

- Poor credit history
- Loan amount requested
- Self-employment

OUTLIERS DETECTION

Loan Amount seem to have outliers with target variable loan status



Unexpected outlier when loan amount near 600-700 and loan being approved

Key reasons considered:

- Lender's policy
- Local regulations
- Purpose of loan

ONE-HOT ENCODING
CATEGORICAL-NUMERICAL CONVERSION

One-hot encoding is a technique used to convert categorical data into a format that can be used by machine learning algorithms

WHY ONE-HOT?

models require numerical input, and one-hot encoding is crucial to efficiently handle categorical variables

PROCESS:

- Converts categorical variables into binary columns, by creating dummy variables.
- Drops one category column from each categorical variable to avoid multicollinearity.
- Renames columns as needed for clarity and consistency.

CONVERSION

Gender	Married	Dependents_0	Dependents_1	Dependents_2	Dependents_3+
1	0	1	0	0	0
1	1	0	1	0	0
1	1	1	0	0	0
1	1	1	0	0	0
1	0	1	0	0	0

Education	Self_Employed	Property_Area_Rural	Property_Area_Semiurban	Property_Area_Urban
1	0	0	0	1
1	0	1	0	0
1	1	0	0	1
0	0	0	0	1
1	0	0	0	1

FEATURE ENGINEERING

Feature engineering is the process of selecting, transforming, or creating input variables to improve machine learning model performance by enhancing data representation..

It is used in the loan prediction to defining the feature vector X (containing all input features) and the target vector Y (containing the output variable) for our machine learning model.

1 FEATURE VECTOR (X)

- comprised all the input features used for prediction.
- Were selected based on their relevance to the task:

2 TARGET VECTOR(Y)

- 'Loan_Status' is the target representing binary outcome for loan approval or rejection.

S M O T E

Synthetic Minority Oversampling Technique

- useful for imbalanced data
- Creates synthetic samples for the minority class.
- increases the representation of the minority class, making the dataset more balanced.

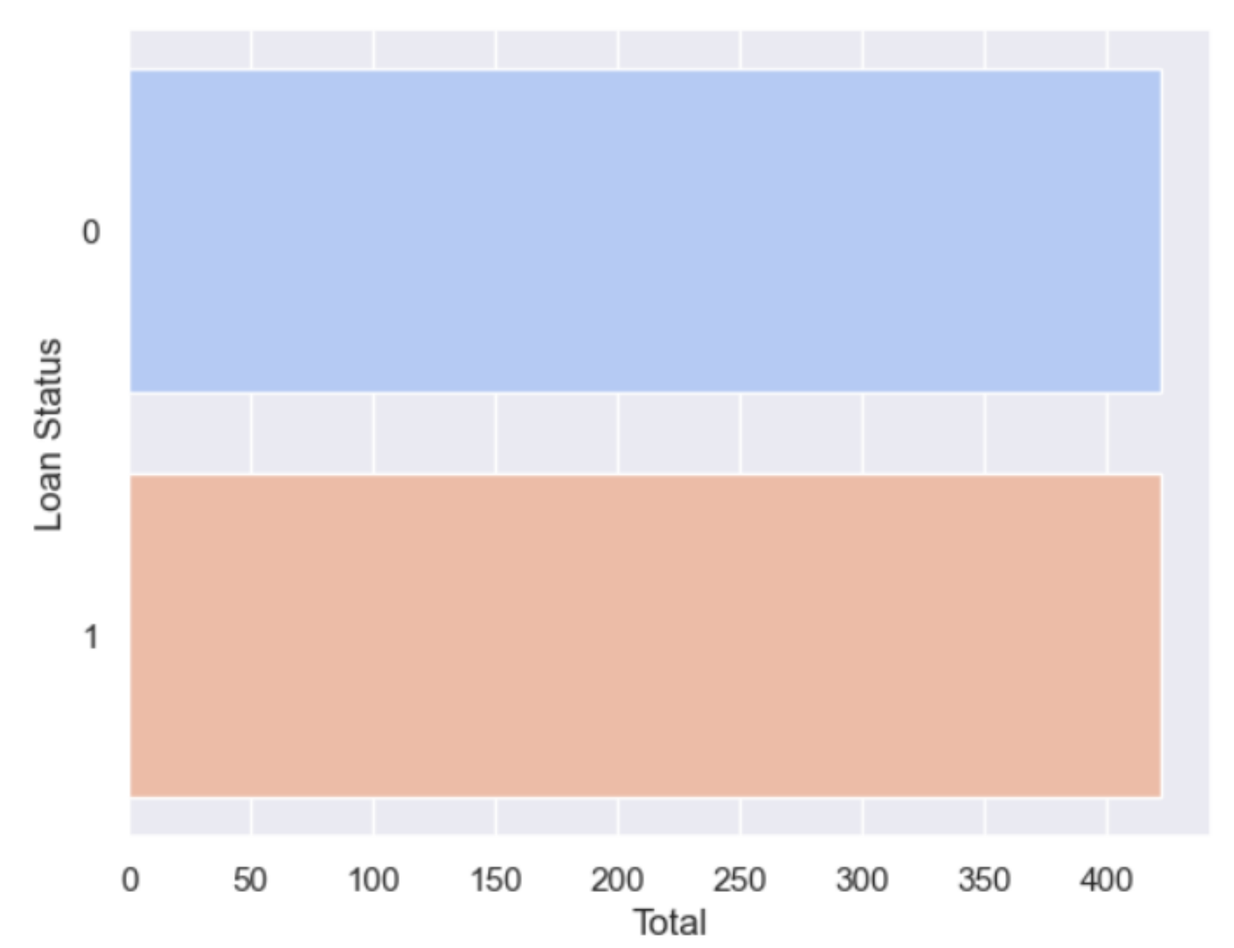
IMBALANCED DATA ISSUE

The original dataset exhibited a class imbalance, with a significantly higher number of 'Yes' (loan approved) instances and a lower number of 'No' (loan rejected) instances.

APPLICATION

SMOTE was applied to the 'Loan_Status' column to create synthetic samples for the 'No'(Loan rejected) class, effectively addressing the class imbalance.

VISUALIZATION



STANDARD-SCALING

It standardizes or normalizes the features of a dataset to have a mean of 0 and a standard deviation of 1.

Ensures that all features are on the same scale, preventing some features from dominating others in the modeling process.

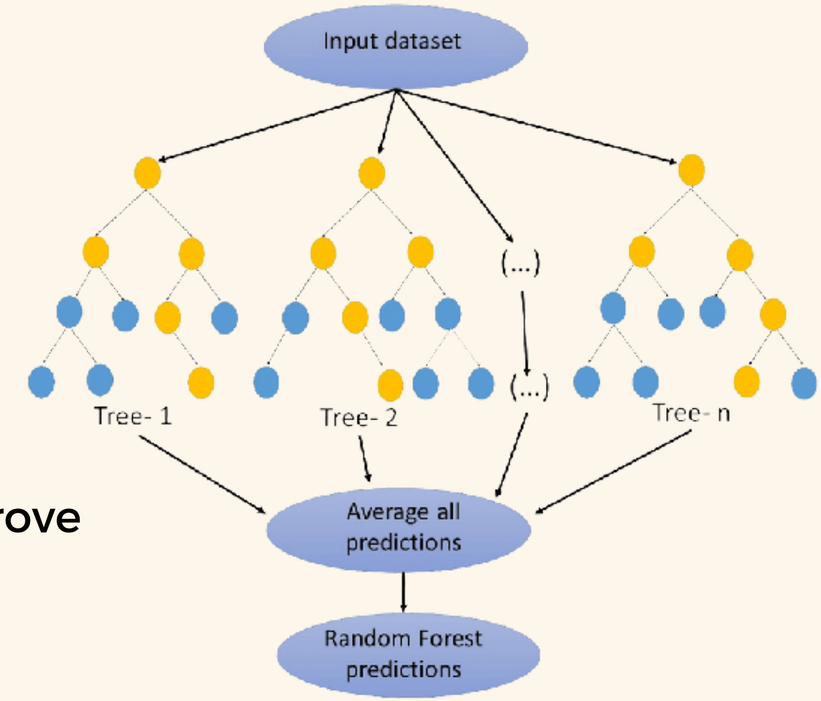
HOW'S IT DONE

Calculates the mean and standard deviation of each feature in the dataset

Subtracts the mean from each data point and then divides by the standard deviation for each feature.

The values are transformed linearly so that they are distributed around 0.

MODELING
RANDOM FOREST



- A technique that combines the power of multiple decision trees to improve predictive accuracy.

used 1900 decision trees also known as n_estimators

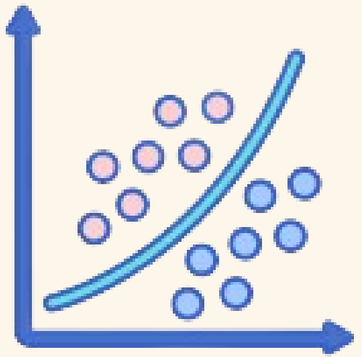
adjusted hyperparameters like max_depth and max_leaf_nodes to optimize performance.

	precision	recall	f1-score	support
0	0.97	0.76	0.85	94
1	0.76	0.97	0.85	75
accuracy			0.85	169
macro avg	0.87	0.86	0.85	169
weighted avg	0.88	0.85	0.85	169

```
[[71 23]
 [ 2 73]]
Random Forest Accuracy:  85.21%
```

MODELING

LOGISTIC REGRESSION



- helps determine whether a loan will be approved ('Yes' or 'No').

incorporated regularization techniques for improved generalization.

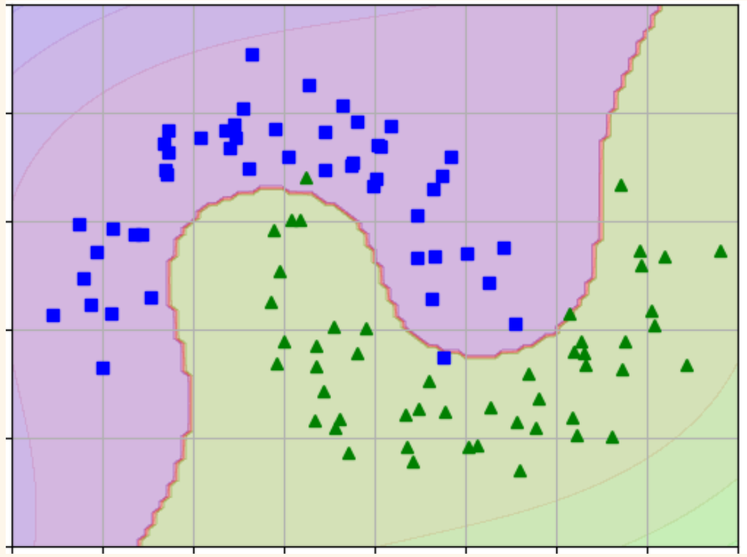
selected the 'saga' solver, set a maximum limit for convergence.

	precision	recall	f1-score	support
0	0.97	0.76	0.85	94
1	0.76	0.97	0.85	75
accuracy			0.85	169
macro avg	0.87	0.86	0.85	169
weighted avg	0.88	0.85	0.85	169

```
[[71 23]
 [ 2 73]]
LR accuracy: 85.21%
```

MODELING

SUPPORT VECTOR MACHINE(SVM)



- SVM can handle high-dimensional data and is highly effective in scenarios where clear margins of separation between classes are crucial.

fine-tuned with key hyperparameters, with the choice of kernel functions (rbf).

maximum iterations set to 500 during training to optimize performance.

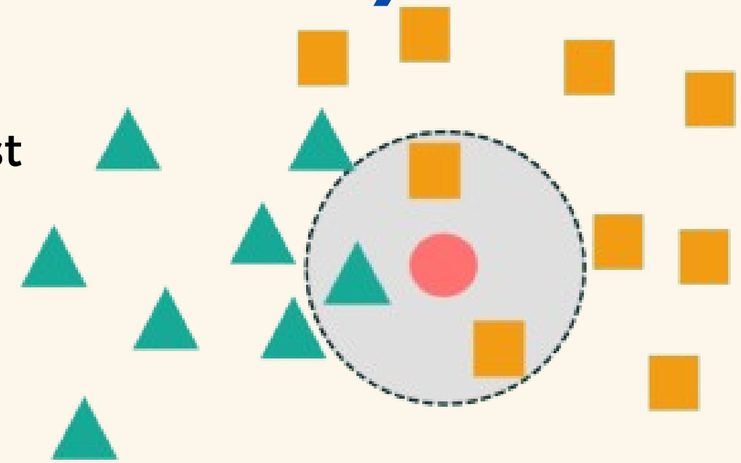
	precision	recall	f1-score	support
0	0.97	0.74	0.84	94
1	0.75	0.97	0.85	75
accuracy			0.85	169
macro avg	0.86	0.86	0.85	169
weighted avg	0.87	0.85	0.85	169

```
[[70 24]
 [ 2 73]]
SVC accuracy: 84.62%
```

MODELING

KNN(K-NEAREST NEIGHBORS)

makes predictions based on the majority class or nearest data points to a given input sample.

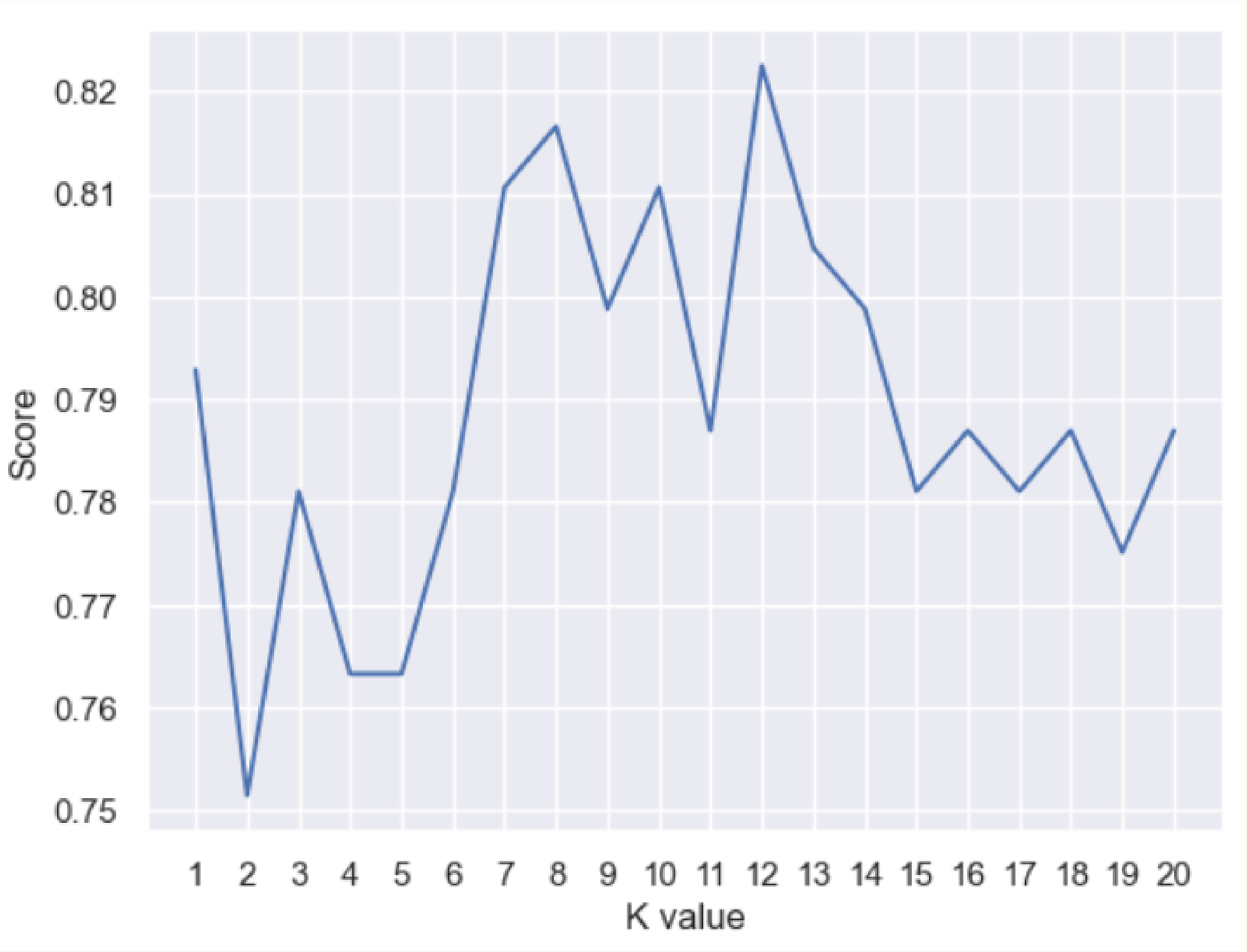


hyperparameter optimization involved testing a range of K values from 1 to 20.

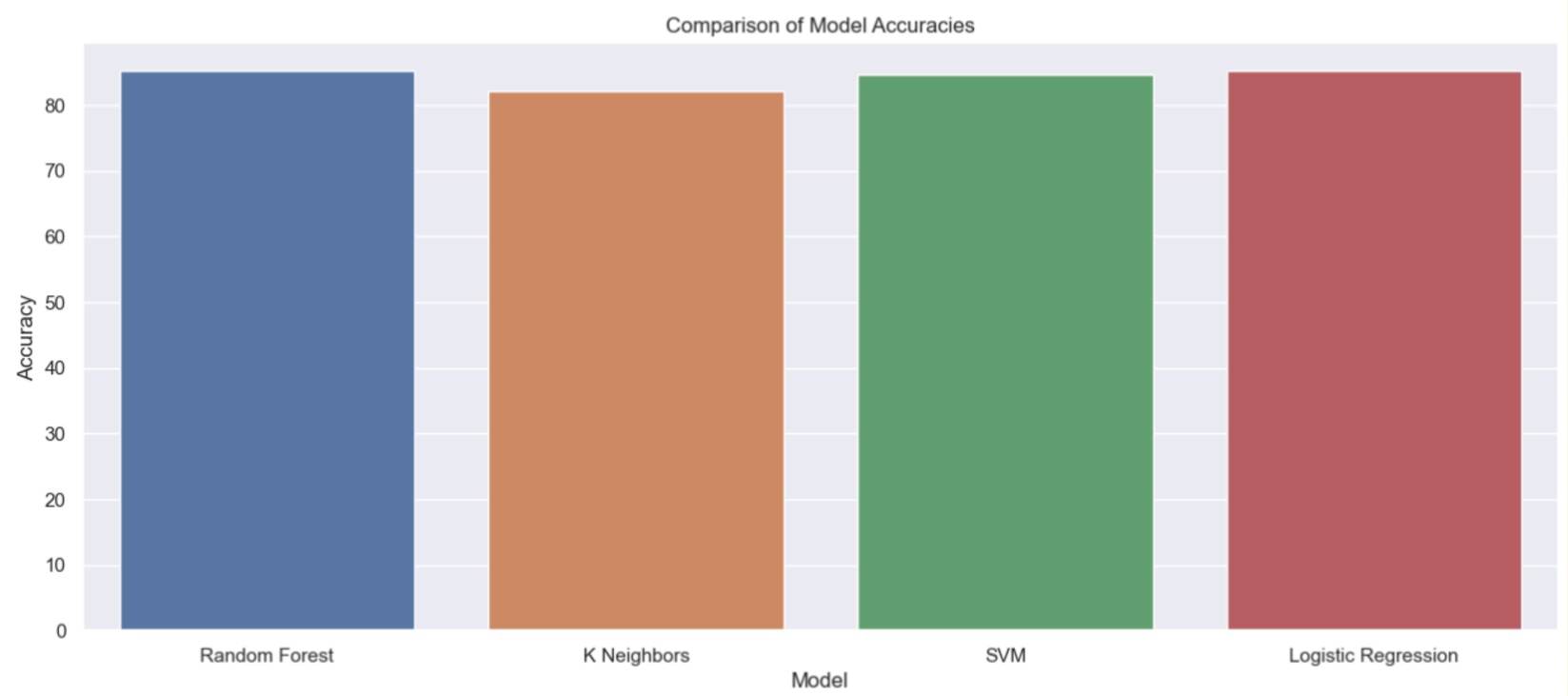
optimal value of K maximizes model's performance.

	precision	recall	f1-score	support
0	0.94	0.66	0.78	94
1	0.69	0.95	0.80	75
accuracy			0.79	169
macro avg	0.81	0.80	0.79	169
weighted avg	0.83	0.79	0.79	169

```
[[62 32]
 [ 4 71]]
KNN best accuracy: 82.25%
```


0.82
0.81
0.80
0.79
0.78
0.77
0.76
0.75
Score
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
K value

VISUALIZATION



	Model	Accuracy
0	Random Forest	85.207101
3	Logistic Regression	85.207101
2	SVM	84.615385
1	K Neighbors	82.248521

THANK YOU