Loan Prediction

Yifan Chen, Xupeng Tang, Chris Yang, Jiajie Yao, Hongyan Zhao

Background

In the realm of financial services, the accurate prediction of loan approval stands as a cornerstone for responsible lending.

This project delves into the realm of machine learning, specifically leveraging the Loan Prediction Problem Dataset. By scrutinizing applicant data encompassing demographics, financial metrics, and credit history, our goal is to construct a predictive model, to determine whether to approve the loan for the applicant.

This model aims to not only streamline decision-making within financial institutions but also foster a more nuanced understanding of risk factors, contributing to the cultivation of responsible and informed lending practices.

Data Description

Our dataset comprises several key variables:

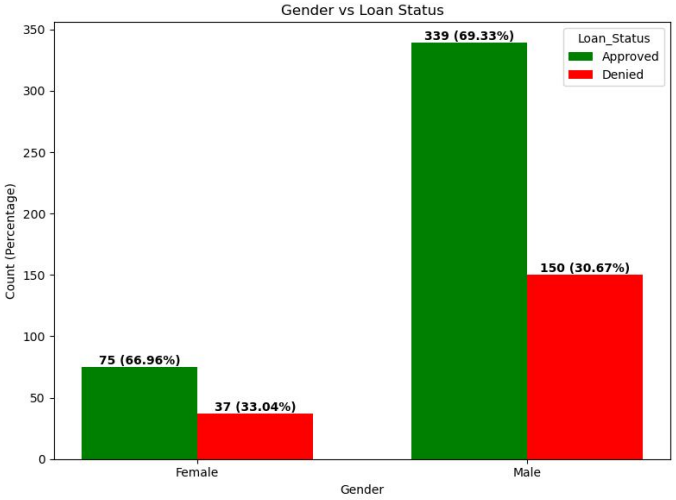
- `Loan_ID`: Unique identifier for each loan application.
- `Gender`: Gender of the applicant (Male, Female).
- `Married`: Marital status of the applicant (Married, Not Married).
- `Dependents`: Number of dependents (0, 1, 2, 3+).
- `Education`: Educational background of the applicant (Graduate, Not Graduate).
- `Self_Employed`: Whether the applicant is self-employed (Self Employed, Not Self Employed).
- `ApplicantIncome`: Income of the applicant (in dollars).
- `CoapplicantIncome`: Income of the co-applicant (in dollars).
- `LoanAmount`: Amount of the loan requested (in thousands of dollars).
- `Loan_Amount_Term`: Term of the loan (in months).
- `Credit_History`: Credit history of the applicant (1 for good credit history, 0 for otherwise).
- `Property_Area`: Area where the property is located (Rural, Urban, Semi-Urban).
- `Loan_Status`: The target variable indicating loan approval status (Y for approved, N for denied).

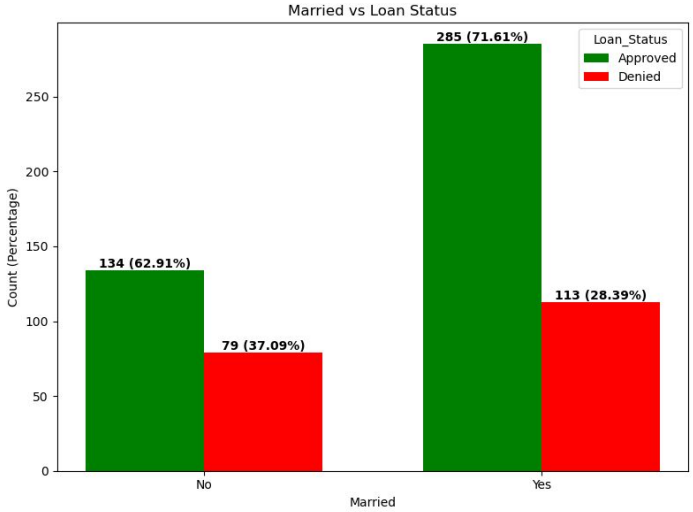
Dataset

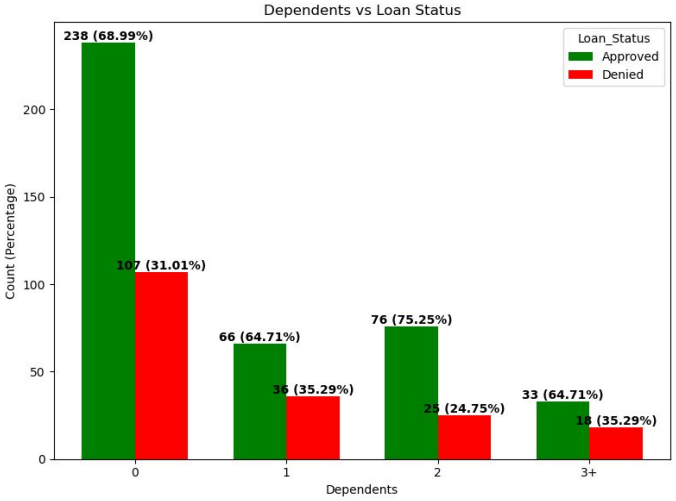
The training data has 614 rows and 13 columns.
Training Data:

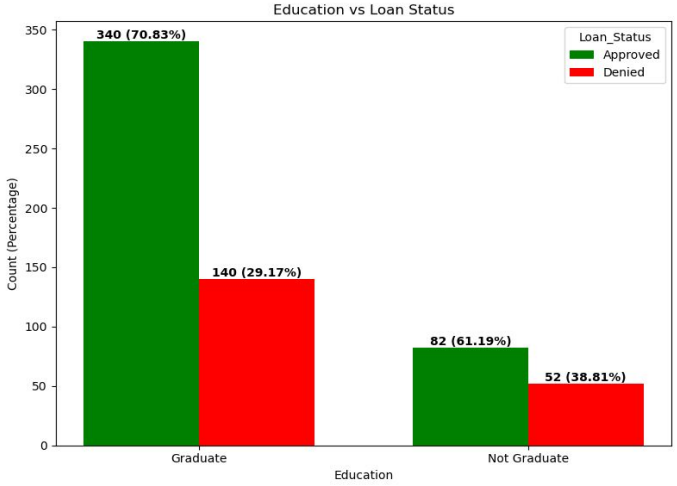
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y

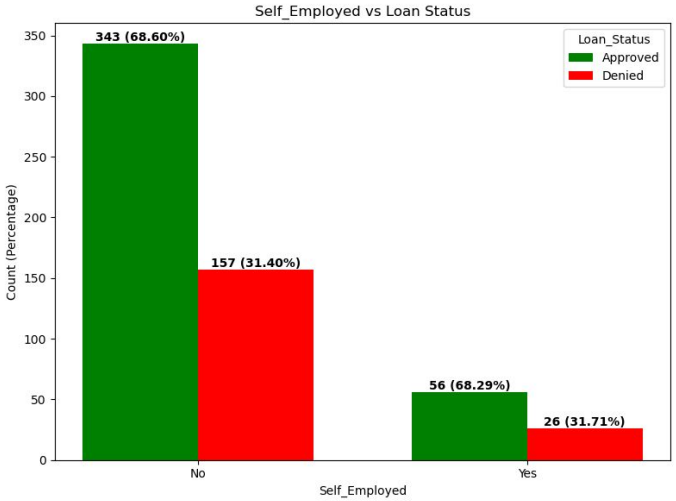
Exploratory Data Analysis

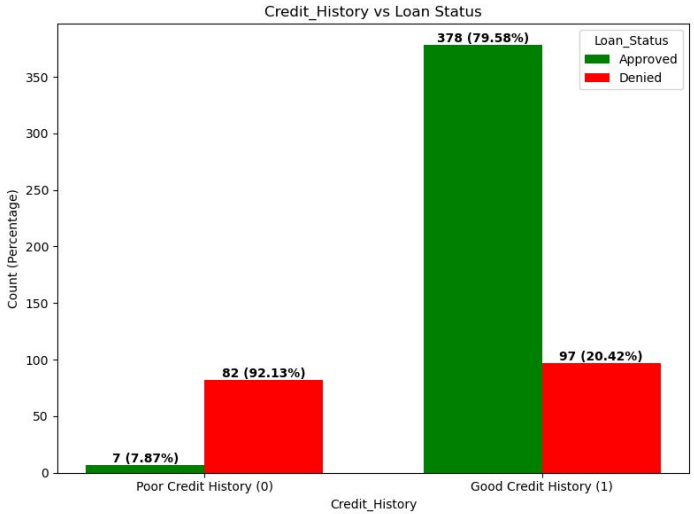

Gender vs Loan Status
Loan_Status
Approved
Denied
Count (Percentage)
Gender
Female
Male
75 (66.96%)
37 (33.04%)
339 (69.33%)
150 (30.67%)

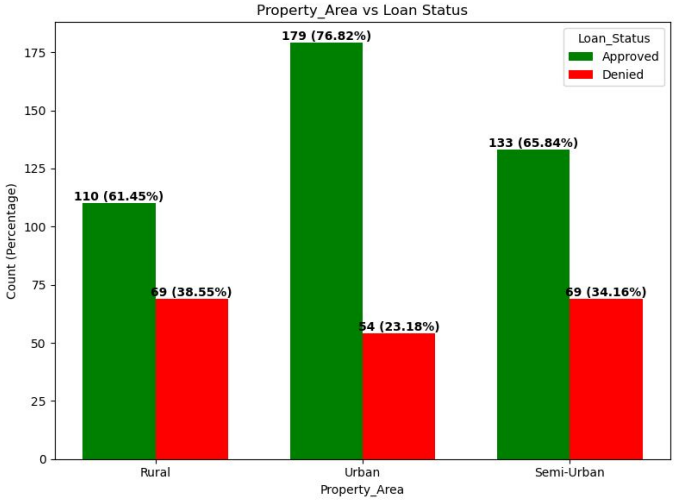

Married vs Loan Status
Loan_Status
Approved
Denied
Count (Percentage)
Married
No
Yes
134 (62.91%)
79 (37.09%)
285 (71.61%)
113 (28.39%)

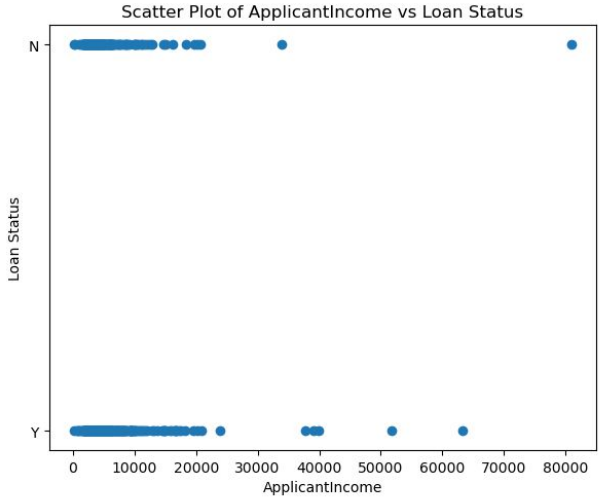

Dependents vs Loan Status
Loan_Status
Approved
Denied
Count (Percentage)
Dependents
238 (68.99%)
107 (31.01%)
66 (64.71%)
36 (35.29%)
76 (75.25%)
25 (24.75%)
33 (64.71%)
18 (35.29%)

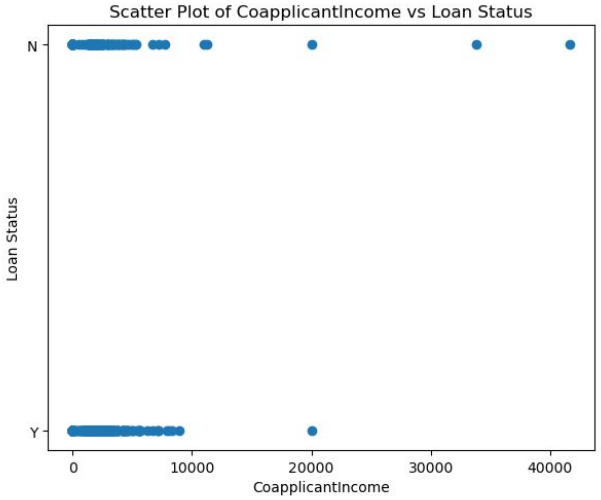

Education vs Loan Status
Loan_Status
Approved
Denied
Count (Percentage)
Education
Graduate
Not Graduate
340 (70.83%)
140 (29.17%)
82 (61.19%)
52 (38.81%)

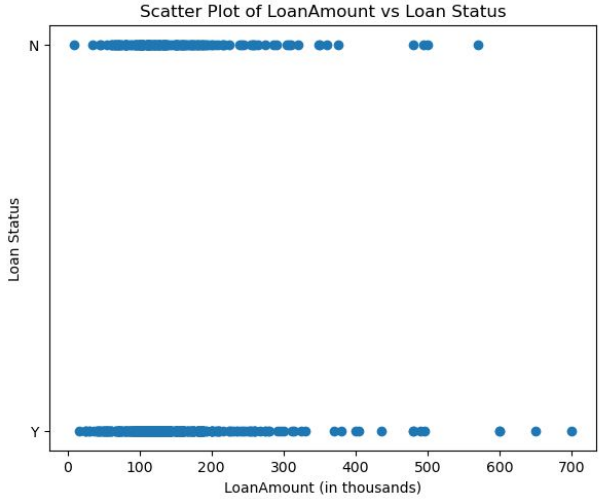

Self_Employed vs Loan Status
Loan_Status
Approved
Denied
343 (68.60%)
157 (31.40%)
56 (68.29%)
26 (31.71%)
Count (Percentage)
No
Yes
Self_Employed

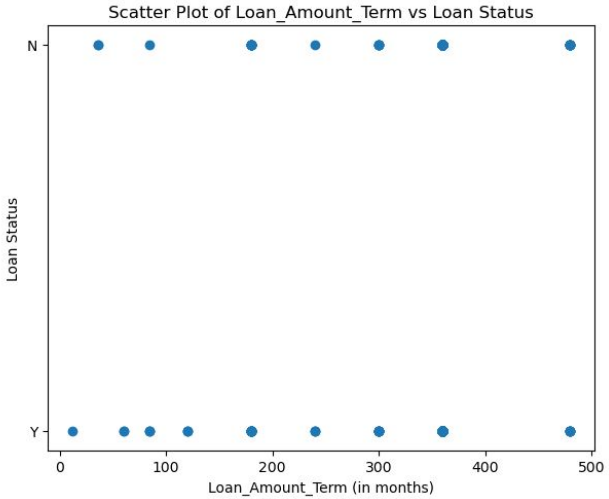

Credit_History vs Loan Status
Loan_Status
Approved
Denied
7 (7.87%)
82 (92.13%)
378 (79.58%)
97 (20.42%)
Count (Percentage)
Poor Credit History (0)
Good Credit History (1)
Credit_History

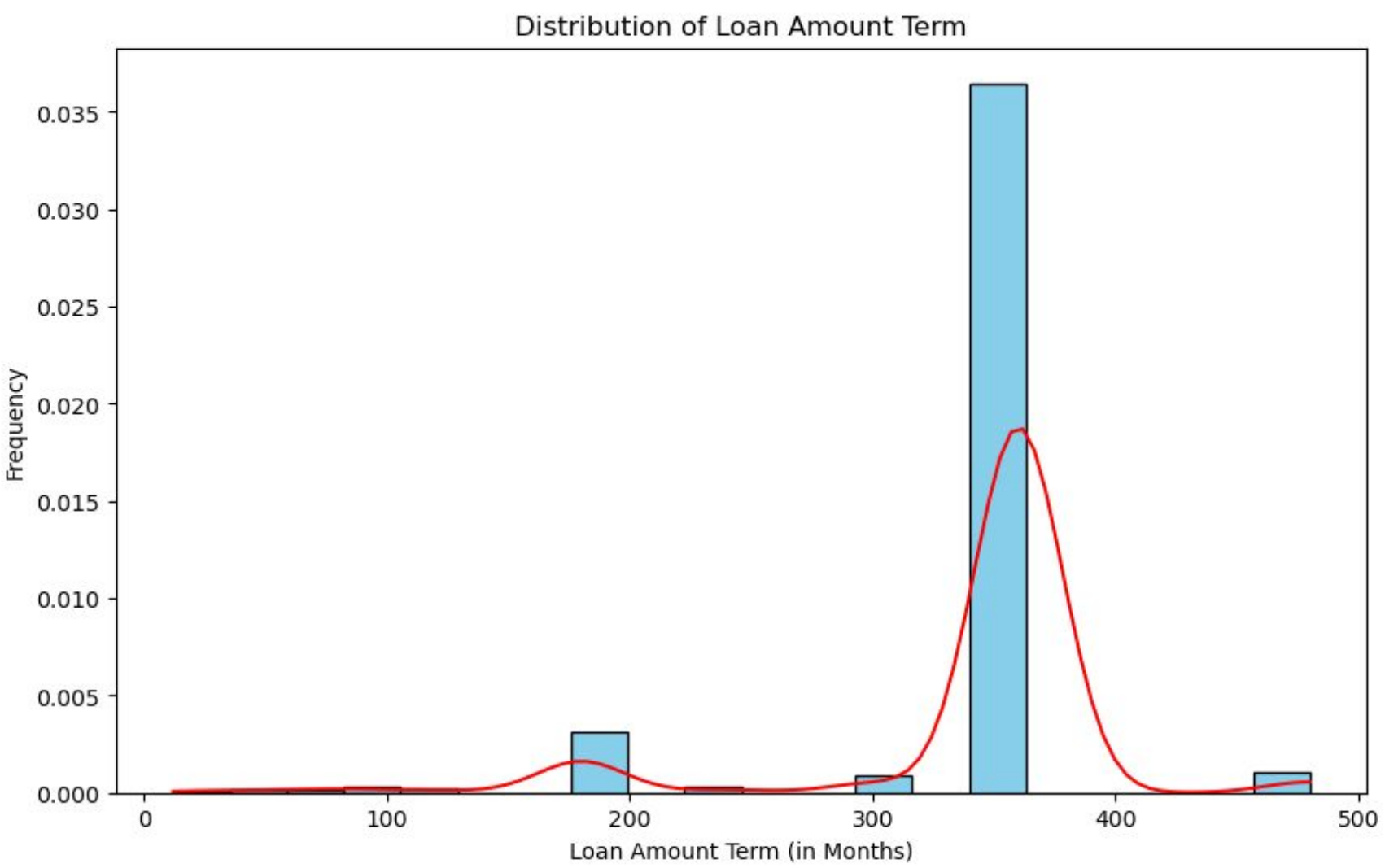

Property_Area vs Loan Status
Loan_Status
Approved
Denied
110 (61.45%)
69 (38.55%)
179 (76.82%)
54 (23.18%)
133 (65.84%)
69 (34.16%)
Count (Percentage)
Rural
Urban
Semi-Urban
Property_Area

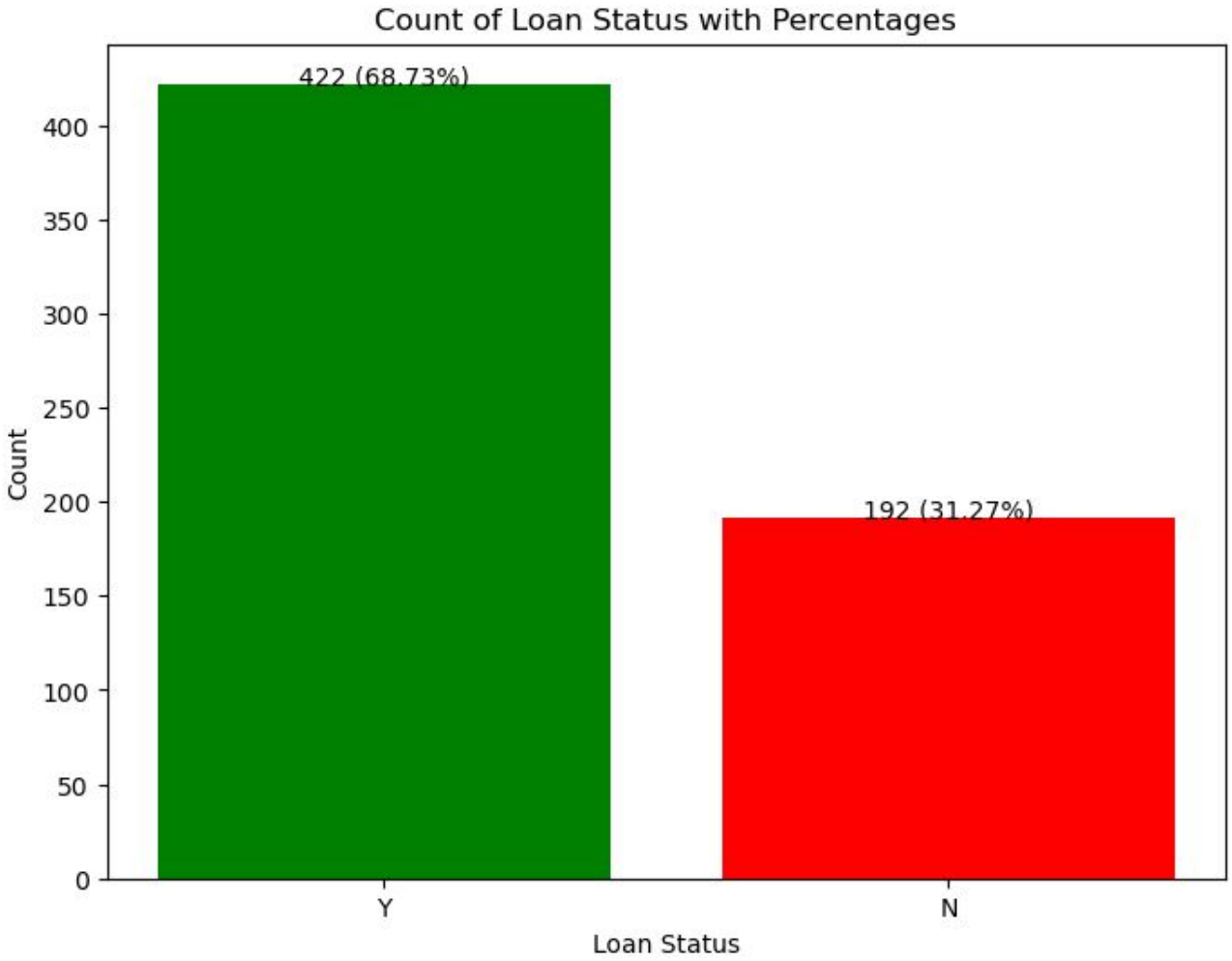

Scatter Plot of ApplicantIncome vs Loan Status
Loan Status
N
Y
0
10000
20000
30000
40000
50000
60000
70000
80000
ApplicantIncome


Scatter Plot of CoapplicantIncome vs Loan Status
Loan Status
N
Y
0
10000
20000
30000
40000
CoapplicantIncome


Scatter Plot of LoanAmount vs Loan Status
Loan Status
N
Y
0
100
200
300
400
500
600
700
LoanAmount (in thousands)


Scatter Plot of Loan_Amount_Term vs Loan Status
Loan Status
N
Y
0
100
200
300
400
500
Loan_Amount_Term (in months)


Distribution of Loan Amount Term
Frequency
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
100
200
300
400
500
Loan Amount Term (in Months)


Count of Loan Status with Percentages
422 (68.73%)
192 (31.27%)
Count
0
50
100
150
200
250
300
350
400
Y
N
Loan Status

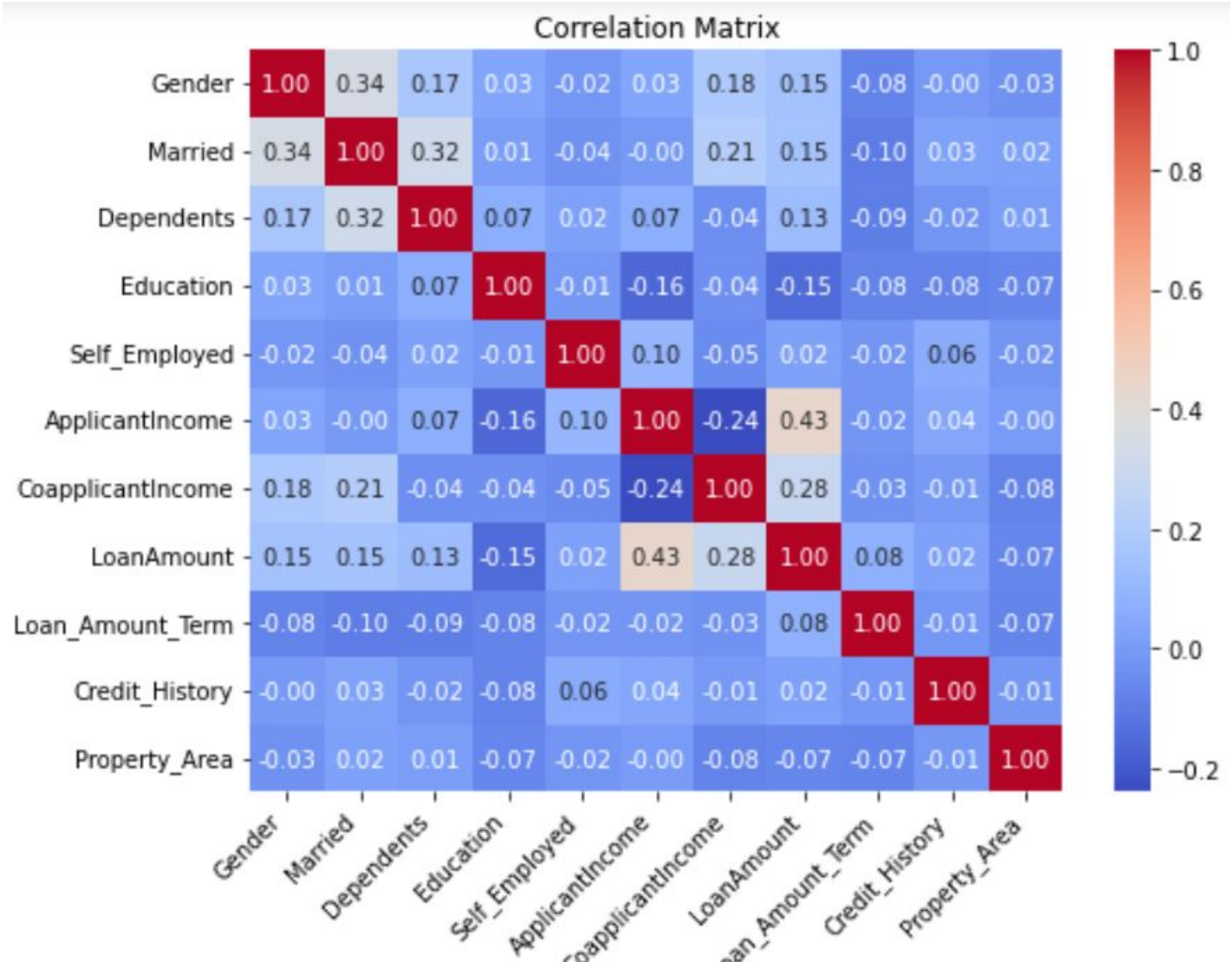
Preparation before modeling

Preprocessing Data

1. Convert categorical variables and non-standard numerical values to numeric values
   #For simplicity, we use Label Encoding for categorical variables
2. Impute missing values
   # For categorical columns, fill missing values with the mode
   # For numerical columns, fill missing values with the median
3. Remove outliers
   # Here, we use a simple method of removing values that are beyond 3 standard deviations from the mean

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	1	0	0.0	0	0	5849	0.0	128.0	360.0	1.0	2	1
1	1	1	1.0	0	0	4583	1508.0	128.0	360.0	1.0	0	0
2	1	1	0.0	0	1	3000	0.0	66.0	360.0	1.0	2	1
3	1	1	0.0	1	0	2583	2358.0	120.0	360.0	1.0	2	1
4	1	0	0.0	0	0	6000	0.0	141.0	360.0	1.0	2	1

Correlation Matrix



PCA

Explained variance ratio is set to 95%

Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
0	-1.048034	0.997088	0.197283	1.117371	-0.411208	-0.202591	0.131674	-0.636012	0.967531	-0.815652
1	0.712545	-0.113204	-0.507286	-0.475809	0.336084	-0.954566	0.339778	0.958273	0.291930	0.259092
2	-0.734810	-0.195042	0.841875	1.113072	0.963517	0.008335	-1.693967	0.074033	1.056185	0.736767
3	-0.018778	-1.580996	-0.121421	0.434776	0.042295	-0.132564	-0.145632	-2.056172	0.185487	0.616348
4	-0.925248	1.126658	0.155562	1.090603	-0.446349	-0.163390	0.170565	-0.708914	0.918788	-0.826128

PCA

Explained Variance Ratio

	PC	Explained Variance Ratio
0	PC1	0.166351
1	PC2	0.135786
2	PC3	0.113989
3	PC4	0.097313
4	PC5	0.094600
5	PC6	0.085422
6	PC7	0.081580
7	PC8	0.071628
8	PC9	0.070197
9	PC10	0.052805

Models

SVM

Classifier	Parameters
SVM	{'C': [0.01, 1, 100]}
KNN	{'n_neighbors': [1, 2, 3, 4]}
Logistic Regression	{'C': [0.01, 1, 100]}

Using Grid search to find the best model.

SVM

Model Accuracies

	Classifier	Parameters	Accuracy	Model
0	SVM	{'C': 0.01}	0.669	SVC(C=0.01, random_state=42)
1	KNN	{'n_neighbors': 3}	0.585	KNeighborsClassifier(n_neighbors=3)
2	Logistic Regression	{'C': 1}	0.763	LogisticRegression(C=1, max_iter=5000)

KNN

Classifier	Parameters
SVM	{'C': [0.01, 1, 100]}
KNN	{'n_neighbors': [1, 2, 3, 4]}
Logistic Regression	{'C': [0.01, 1, 100]}

Using Grid search to find the best model.

KNN

Model Accuracies

	Classifier	Parameters	Accuracy	Model
0	SVM	{'C': 0.01}	0.669	SVC(C=0.01, random_state=42)
1	KNN	{'n_neighbors': 3}	0.585	KNeighborsClassifier(n_neighbors=3)
2	Logistic Regression	{'C': 1}	0.763	LogisticRegression(C=1, max_iter=5000)

Logistic Regression

Classifier	Parameters
SVM	{'C': [0.01, 1, 100]}
KNN	{'n_neighbors': [1, 2, 3, 4]}
Logistic Regression	{'C': [0.01, 1, 100]}

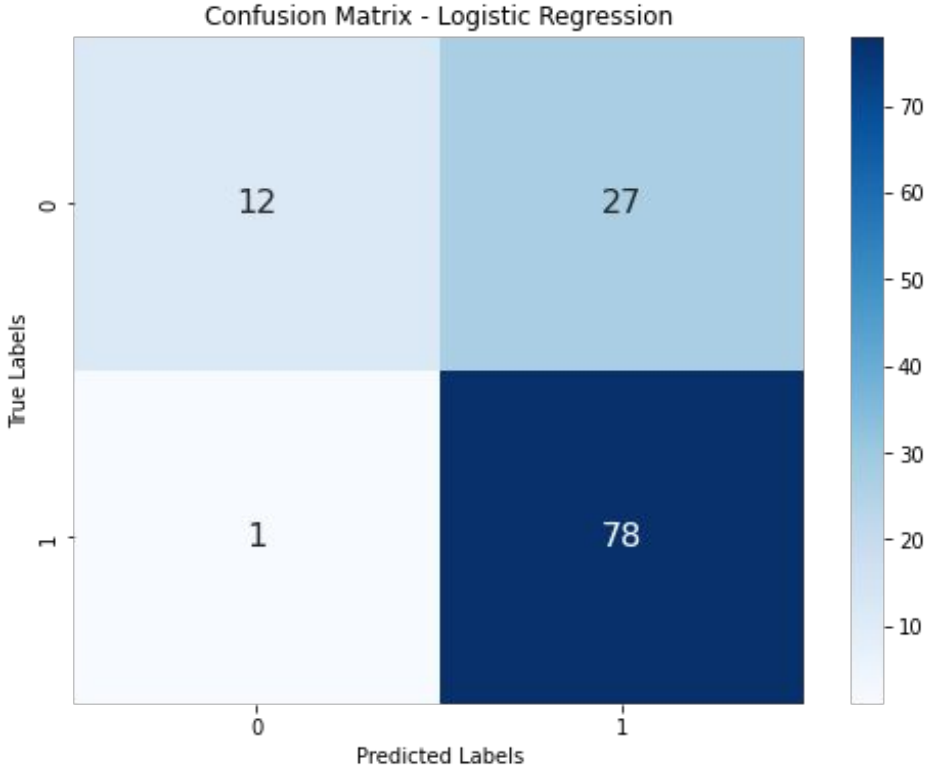
Using Grid search to find the best model.

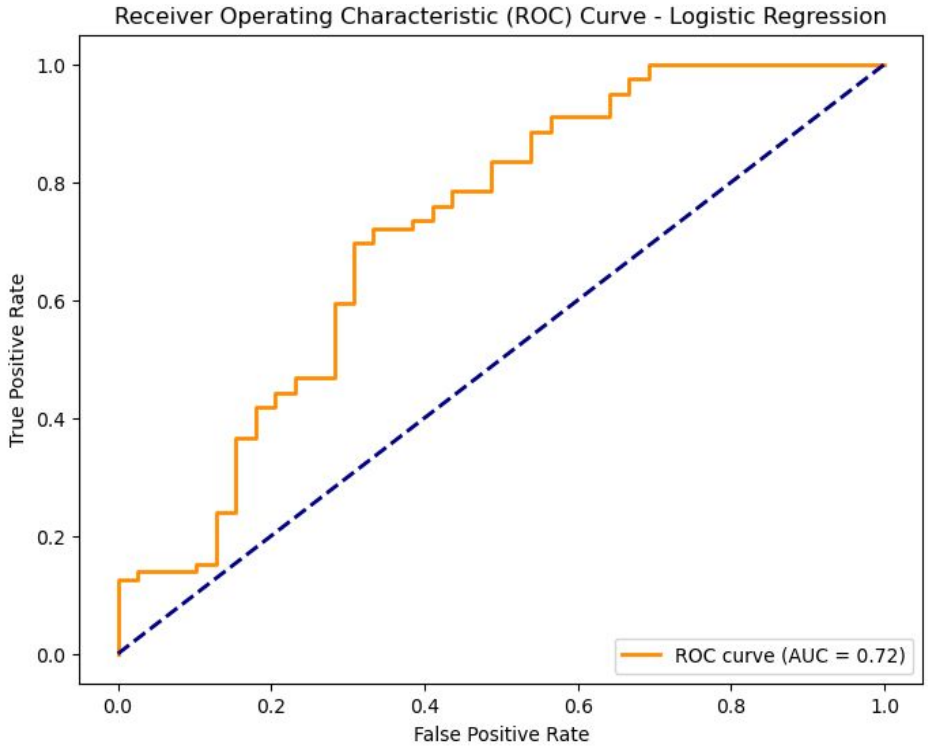
Logistic Regression

Model Accuracies

	Classifier	Parameters	Accuracy	Model
0	SVM	{'C': 0.01}	0.669	SVC(C=0.01, random_state=42)
1	KNN	{'n_neighbors': 3}	0.585	KNeighborsClassifier(n_neighbors=3)
2	Logistic Regression	{'C': 1}	0.763	LogisticRegression(C=1, max_iter=5000)

Logistic Regression





Logistic Regression

Logistic Regression Coefficients and Odds Ratios

	Feature	Coefficient	Odds Ratio
0	Gender	-0.5559	0.5735
1	Married	0.5258	1.6918
2	Dependents	0.1056	1.1113
3	Education	-0.6439	0.5253
4	Self_Employed	-0.1953	0.8226
5	ApplicantIncome	0.0000	1.0000
6	CoapplicantIncome	0.0001	1.0001
7	LoanAmount	-0.0058	0.9942
8	Loan_Amount_Term	-0.0020	0.9980
9	Credit_History	3.1788	24.0175
10	Property_Area	0.0850	1.0887

Compare Odds Ratio with 1

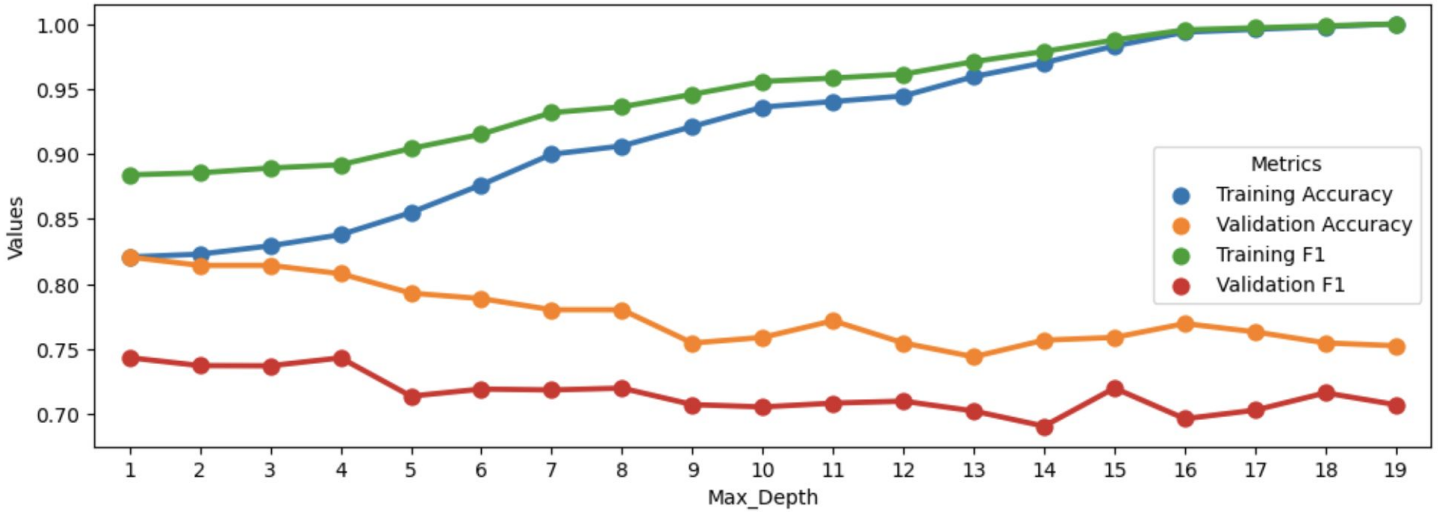
Education,Married, and Credit_history are most significant variables.

Decision Tree

```
Training Data Set Accuracy:  1.0
Training Data F1 Score  1.0
Validation Mean F1 Score:  0.704
Validation Mean Accuracy:  0.739
```

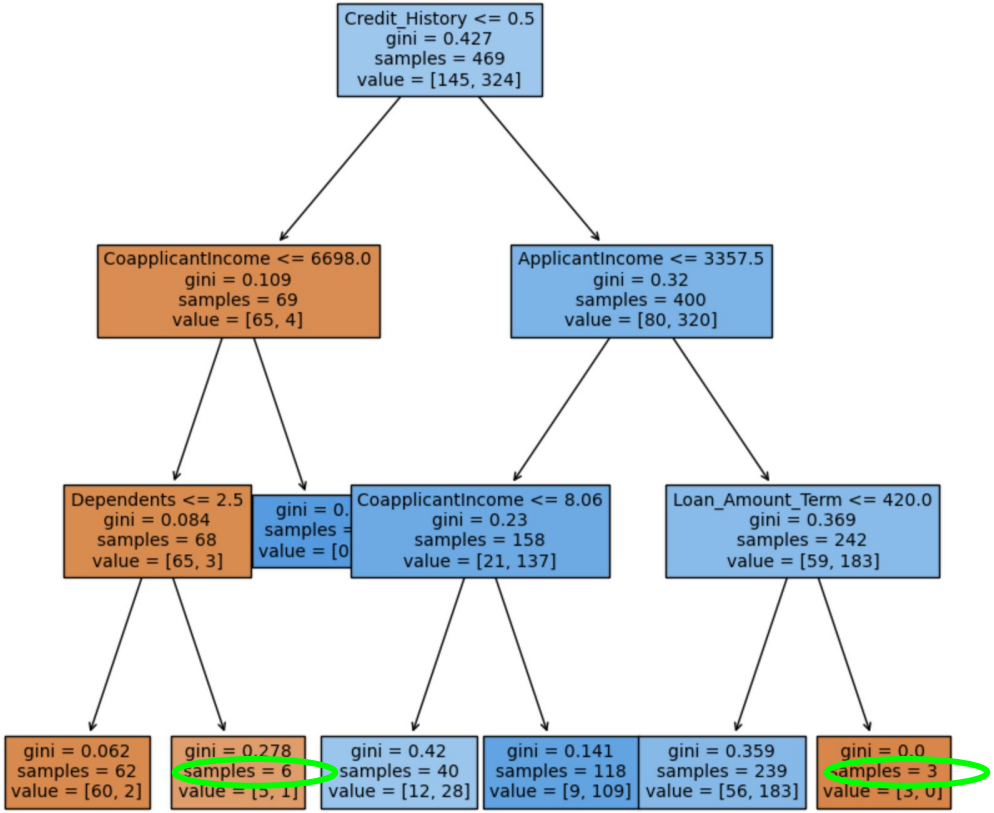
Decision Tree

Tuning Max_Depth




```
Test Accuracy:  0.779
Test F1 Score:  0.858
```

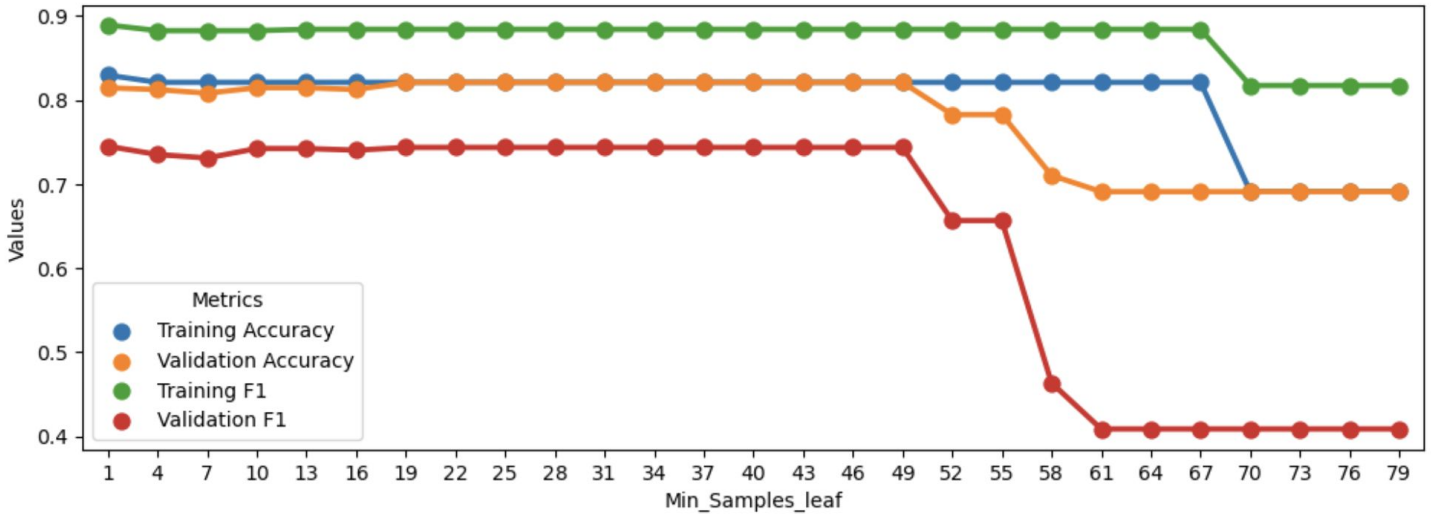
Decision Tree



Overfitting

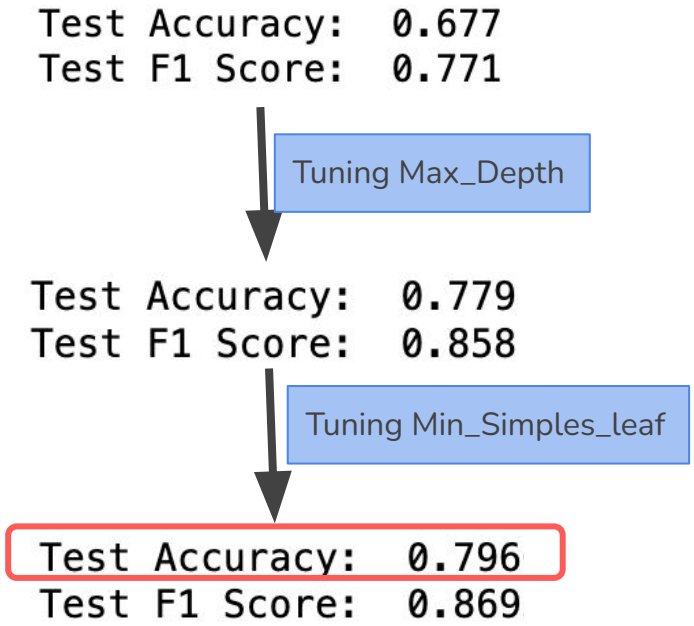
Decision Tree

Tuning Min_Simples_leaf




```
Test Accuracy:  0.796
Test F1 Score:  0.869
```

Decision Tree



Random Forest

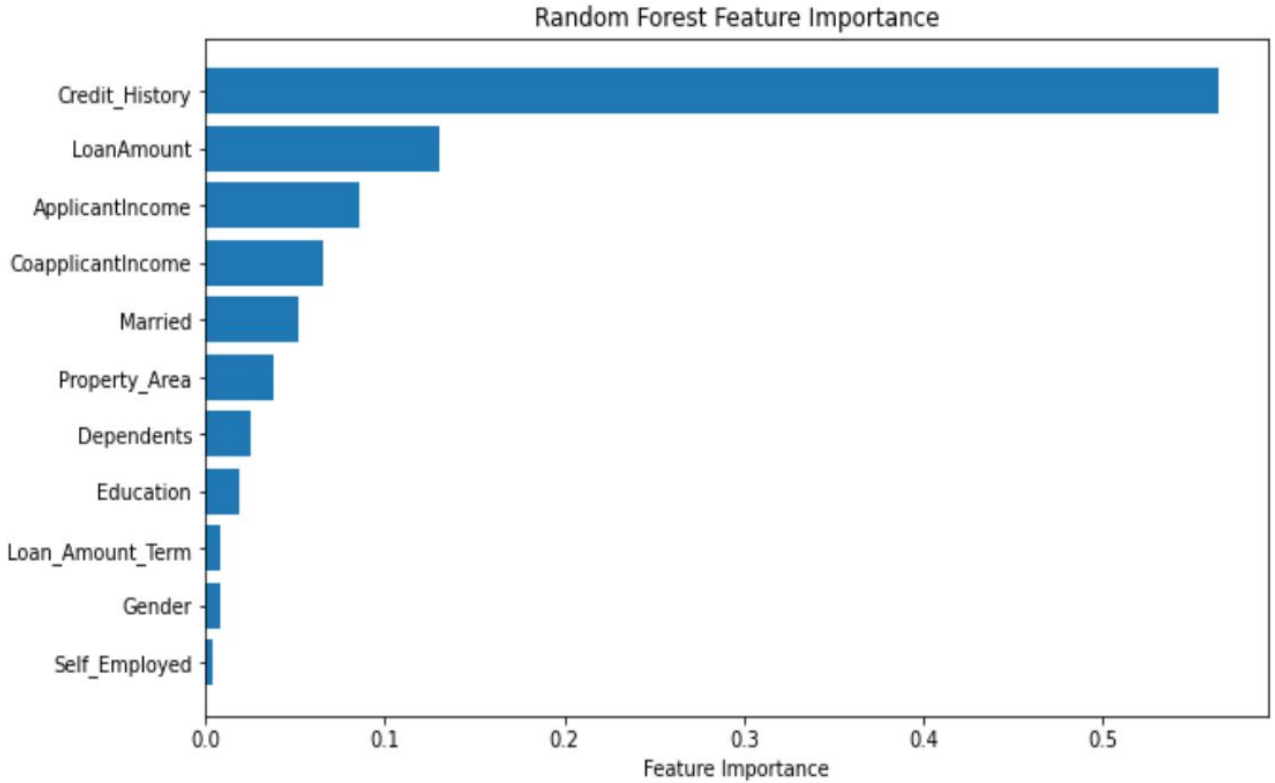
```
'n_estimators': [20, 30, 50],
'max_depth': [None, 10, 20],
'min_samples_split': [2, 5, 10],
'min_samples_leaf': [10, 20, 35]
```

Using Grid search to find the best model.

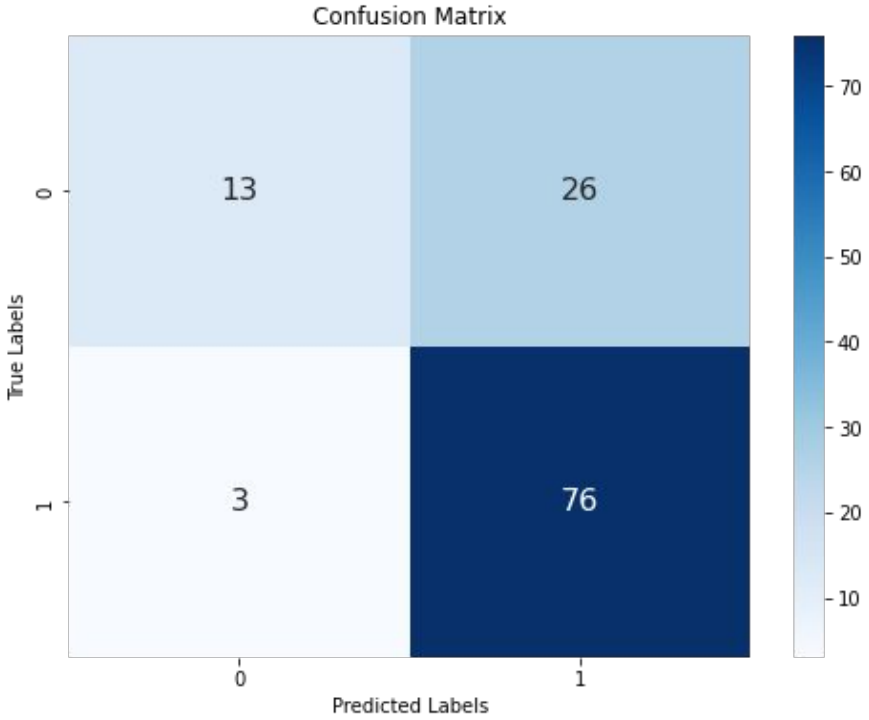
Random Forest

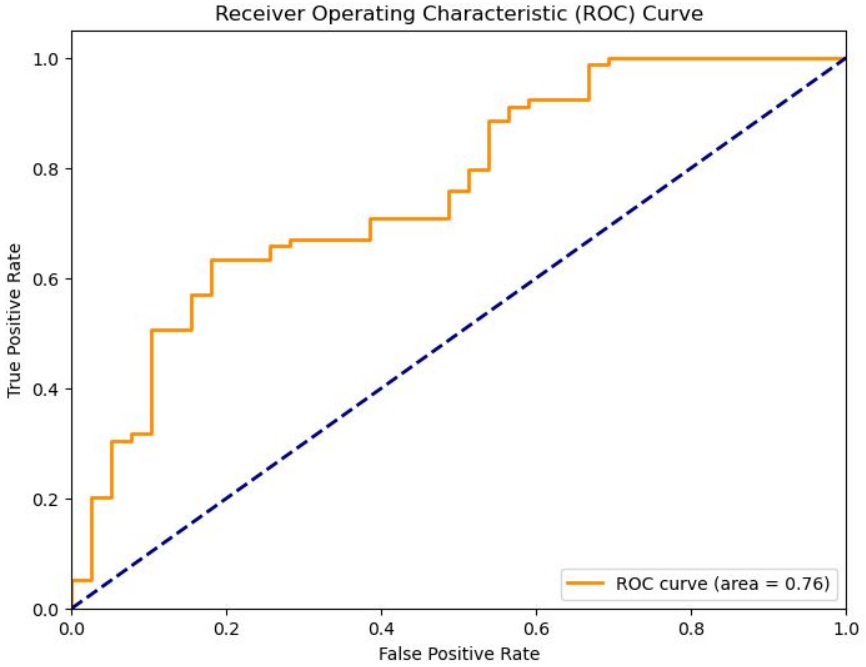
```
Best Parameters: {'max_depth': None, 'min_samples_leaf': 10, 'min_samples_split': 2, 'n_estimators': 20}
Accuracy on Test Set: 0.771
```

Random Forest

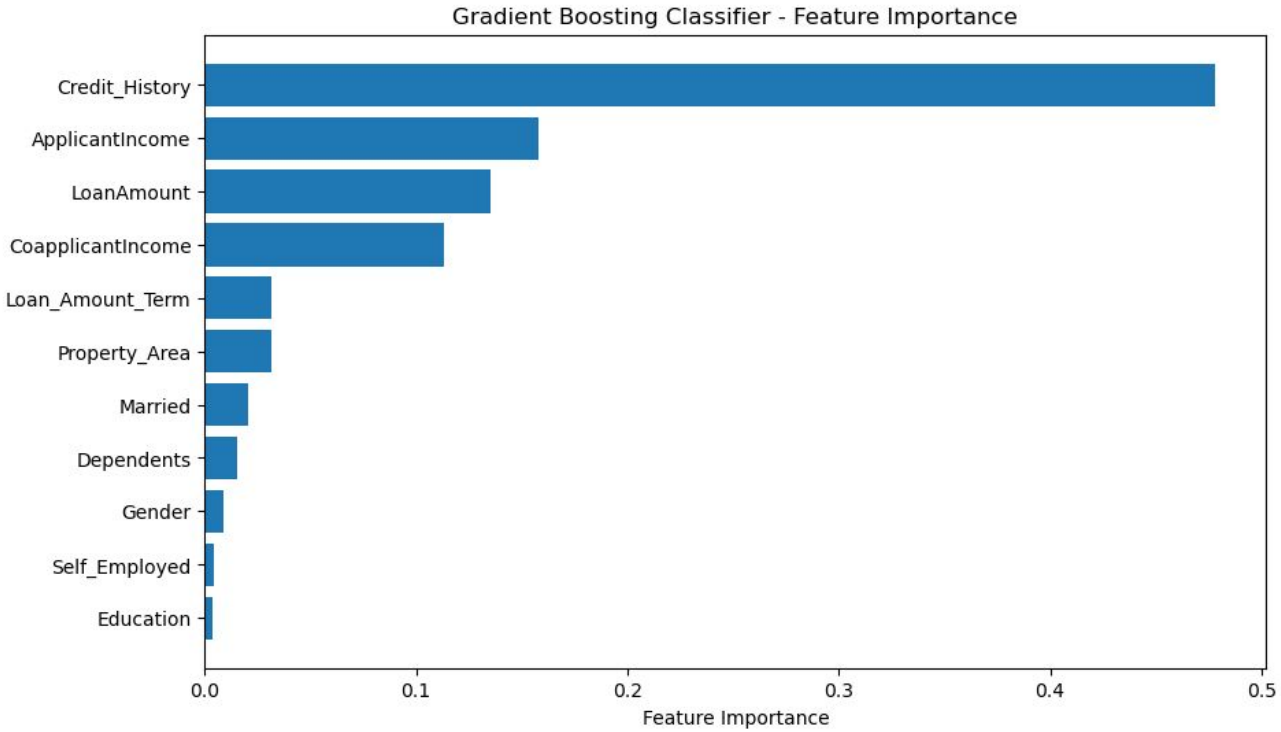


Random Forest

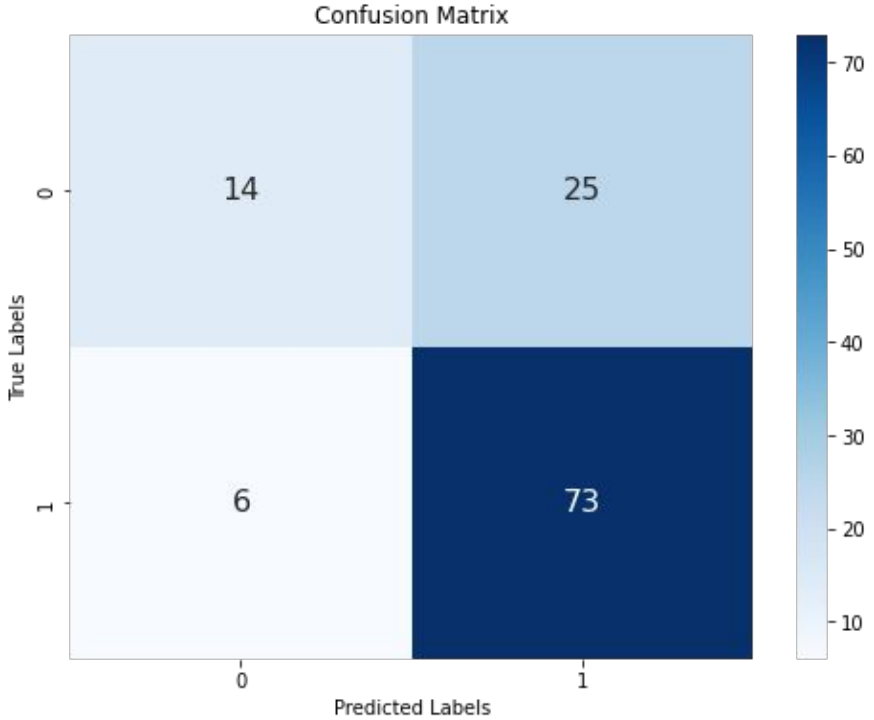




Gradient Boosting



Gradient Boosting



Gradient Boosting

```
Accuracy of Gradient Boosting on Test Set: 0.737
```

Best Model

Model	Accuracy on test dataset
SVM	0.669
KNN	0.585
Logistic Regression	0.763
Decision Tree	0.796
Random Forest	0.771
Gradient Boosting	0.737

Summary & Conclusion

- Decision tree performs the best when predicting the approval or denial of a loan application
- Important features differ between models:
  - Random Forest & Gradient Boosting: Credit History, Loan Amount, Applicant & Co Applicant Income
  - Logistic Regression: Credit History, Education, Marital Status
- Credit History matters

Future work based on weakness and limitations

Enhancing Data Quality:

Data Enrichment: Incorporate more diverse and comprehensive data sources to capture a wider range of variables that might affect loan approval decisions.

Handling Missing Data: Explore advanced imputation techniques or data augmentation methods to better handle missing data without introducing significant bias.

Future work based on weakness and limitations

Advanced Modeling Techniques:

- Ensemble Methods: Further explore ensemble methods that combine the predictions from multiple models to improve accuracy and robustness.

Future work based on weakness and limitations

Model Validation and Testing:

Cross-Validation with Diverse Datasets: Test the model on various datasets to ensure its generalizability and robustness.

Real-World Testing: Pilot the model in a real-world setting to observe its performance and gather feedback.

Thank you for listening.

Yifan Chen, Xupeng Tang, Chris Yang, Jiajie Yao, Hongyan Zhao